Morality is Non-Binary: Building a Pluralist Moral Sentence Embedding Space using Contrastive Learning

Jeongwoo Park*, **Enrico Liscio***, and **Pradeep K. Murukannaiah**

Delft University of Technology, the Netherlands

`E.Liscio@tudelft.nl`

Abstract

Recent advances in NLP show that language models retain a discernible level of knowledge in deontological ethics and moral norms. However, existing works often treat morality as binary, ranging from right to wrong. This simplistic view does not capture the nuances of moral judgment. Pluralist moral philosophers argue that human morality can be deconstructed into a finite number of elements, respecting individual differences in moral judgment. In line with this view, we build a pluralist moral sentence embedding space via a state-of-the-art contrastive learning approach. We systematically investigate the embedding space by studying the emergence of relationships among moral elements, both quantitatively and qualitatively. Our results show that a pluralist approach to morality can be captured in an embedding space. However, moral pluralism is challenging to deduce via self-supervision alone and requires a supervised approach with human labels.

1 Introduction

Morality helps humans distinguish right from wrong (Graham et al., 2013). As AI systems work with (or for) humans, it is crucial that they align with human morality (Gabriel, 2020; Liscio et al., 2023b). Several NLP methods have been proposed to recognize human morality in text (Forbes et al., 2020; Lourie et al., 2021; Jiang et al., 2022; Pyatkin et al., 2023). However, such methods typically treat morality as a score that ranges in a single dimension of right to wrong. This does not reflect the nuances in moral reasoning, differences among individuals, or the existence of moral value conflicts (Telkamp and Anderson, 2022).

Pluralist moral philosophers argue that morality should be represented through a finite number of basic elements, referred to as moral values (Graham

---

* Equal Contribution.

et al., 2013). Each situation triggers one or more moral values, and each of us assigns varying importance to each moral value. The combination of these two aspects determines the individual moral judgment in the situation. For instance, the debate on immigration touches on the moral values of *fairness* ("Everyone should be given equal opportunities") and in-group *loyalty* ("I worry about the preservation of our identity"). The way in which each of us prioritizes fairness vs. loyalty influences our moral judgment in this debate. Thus, morality cannot (and should not) be unidimensionally classified in text (Talat et al., 2022). Instead, the moral elements that are salient to a piece of text can be recognized, which can be used to reason about or assist the humans in the moral judgment.

The Moral Foundations Theory (MFT) is a popular pluralist approach to morality (Graham et al., 2013) which states that people have five innate moral foundations on which they base their moral judgments. There is a surge of interest in morality (Vida et al., 2023) and particularly in the MFT in the NLP community (Kobbe et al., 2020; Alshomary et al., 2022; Liscio et al., 2022a, 2023a), partly due to the Moral Foundation Twitter Corpus (MFTC) (Hoover et al., 2020), composed of 35k tweets annotated with the MFT foundations.

Prior research has focused on methods for classifying MFT elements in a textual discourse (Huang et al., 2022; Alshomary et al., 2022; Liscio et al., 2022a). However, such methods provide limited qualitative insight into the relations between text and MFT elements. We explore the mapping between text and MFT through sentence embeddings, which consist of a multi-dimensional representation that encapsulates knowledge from textual data. Instead of being limited to a specific task, a suitable sentence embedding space can be valuable across multiple NLP tasks, such as text classification, generation, and topic modelling (Henderson et al., 2020; Li et al., 2022; Zhang et al., 2022b).

Further, a sentence embedding space can be geometrically explored, allowing us to investigate the relationships among different moral elements.

Schramowski et al. (2022) show that pre-trained sentence embeddings contain a moral direction that maps actions from "do" to "don't", without the need for re-training on morally loaded data. In this work, we investigate whether the same holds for a pluralist approach to morality. That is: do pre-trained sentence embeddings contain discernible clusters corresponding to the different elements of a pluralist approach to morality, or is it necessary to re-train them with a supervised approach to disentangle the different moral elements?

Our contribution is twofold. First, we propose a novel approach for mapping the MFT elements to a sentence embedding space using the state-of-the-art SimCSE (Gao et al., 2021) method, which makes use of the Contrastive Learning paradigm (Le-Khac et al., 2020). Then, we evaluate the resulting embedding space in two ways. First, we perform an intrinsic evaluation to investigate the relationship between different moral elements and evaluate whether a supervised approach is necessary to disentangle the MFT elements in the embedding space. Second, to evaluate whether the relationships among the MFT elements have been adequately captured, we perform an extrinsic evaluation, generalizing the analyses to a novel test set and to the set of words from a moral dictionary.

Our experiments show that a pluralist approach to morality can be captured in a sentence embedding space, but also that human labels are necessary to successfully train the embeddings. Our work represents the starting point for incorporating a pluralist approach to morality in language models, with a warning that self-supervision alone is not sufficient to capture the complexity of human morality.

2   Background and Data

We introduce the method to train sentence embedding spaces (SimCSE) and the data we use.

SimCSE   Sentence embedding spaces represent sentences as points in a high-dimensional space, mapping semantically similar sentences to the same region of space. Contrastive Learning (CL) (Le-Khac et al., 2020) is an approach to training an embedding space based on a contrastive loss that aims to minimize the distance between positive (semantically similar) sentence pairs and maximize the distance between negative (semantically dis-

similar) sentence pairs. Formally, let x_i and x_i^+ be positively related and $\mathbf{h_i}$, $\mathbf{h_i^+}$ be their encoded representations. Then, the training loss for the two instances with a mini-batch of N pairs is:

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau}} \qquad (1)$$

where τ is a temperature hyperparameter and $\text{sim}(\mathbf{h_1}, \mathbf{h_2})$ the cosine similarity (Gao et al., 2021).

SimCSE (Gao et al., 2021) is a text-based CL framework built on BERT sentence embeddings (Reimers and Gurevych, 2019) that demonstrated better performance than other BERT variants (Gao et al., 2021). SimCSE supports *supervised* and *unsupervised* approaches. Supervised SimCSE seeks to minimize the distance between sentences with the same label and maximize the distance between sentences with different labels. Unsupervised SimCSE generates a positive instance by applying a slight variation of a reference sentence through dropout, and uses a random sentence as a negative instance. We detail the SimCSE supervised and unsupervised CL loss in Appendix A.1.

Moral Foundations Twitter Corpus   The MFT (Graham et al., 2013) is a popular pluralist theory of morality that postulates that human morality is composed of five innate moral foundations that combine to describe our moral stance over divisive issues. Each of the five foundations of the MFT is composed of a virtue-vice duality, resulting in the 10 moral elements shown in Table 1.

Element	Definition
Care/ Harm	Support for care for others/ Refrain from harming others
Fairness/ Cheating	Support for fairness and equality/ Refrain from cheating or exploiting others
Loyalty/ Betrayal	Support for prioritizing one's inner circle/ Refrain from betraying the inner circle
Authority/ Subversion	Support for respecting authority and tradition/ Refrain from subverting authority or tradition
Purity/ Degradation	Support for the purity of sacred entities/ Refrain from corrupting such entities

Table 1: The MFT moral foundations (virtue/vice).

The Moral Foundations Twitter Corpus (MFTC) (Hoover et al., 2020) is a collection of 35,108 tweets collected in seven domains: All Lives Matter, Baltimore Protest, Black Lives Matter, hate speech and offensive language (Davidson et al.,

[2017](), 2016 presidential election, MeToo movement, and hurricane Sandy. The tweets were annotated with one or more of the 10 MFT elements, or with a *non-moral* label. As each tweet was annotated by multiple annotators (ranging from 3 to 8), the authors of MFTC use a majority vote to choose the definitive label(s) of each tweet (thus resulting in one or more moral labels per tweet), and *non-moral* is assigned when no majority is present.

3 Training the Embedding Space

We train the moral embedding space by finetuning *unsupervised* and *supervised* SimCSE approaches. The unsupervised approach does not employ label information, thus the strategy described in Section 2 is used. In the supervised approach, SimCSE uses label information to construct the training triples for its supervised CL objective function. Each triple is composed of (1) a *reference* data point, (2) a data point whose distance from the reference should be minimized (*positive instance*), and (3) a data point whose distance from the reference should be maximized (*negative instance*).

Figure 1 shows an example of how the triples are constructed. In this example, the chosen reference instance is labeled with two moral elements—*harm* and *betrayal*. Then, the positive instance is chosen as a data point with the same labels as the reference instance. However, selecting negative instances is not trivial due to the structure of the MFT taxonomy, which is composed of five pairs of virtue-vice. Thus, we propose two policies, *opposite* and *outside*, to guide the choice of negative instances.
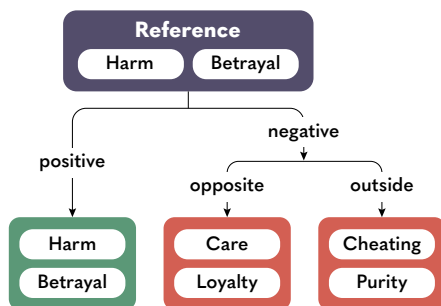

Figure 1: Example triple formation with the two policies for negative instance selection (*opposite* and *outside*).

The *opposite* policy selects the negative instance as a data point annotated with moral elements that are opposite virtue/vice of the reference labels (*care* and *loyalty* in the example). In contrast, the *outside* policy chooses the negative instance as a data point annotated with moral elements that belong to other moral foundations than the reference foundations (*cheating* and *purity* in the example).

In both policies, we prioritize data points with more negative labels when choosing the negative instance, when possible. For instance, in the example in Figure 1, with the *opposite* policy, we prioritize a data point with the labels *care* and *loyalty* over a data point with just the *care* label. We divide the MFTC training set into two halves and apply each policy to a half. We ensure that each data point appears in just one triple. When no suitable positive or negative instances are available, data points labeled as *non-moral* are used as positive or negative instances, until all morally-loaded data points have been used in a triple.

4 Evaluating the Embedding Space

We use 90% of the MFTC as the training set to train the moral embedding space (with the approaches described in Section 3) and the remaining 10% as the test set. To generate a balanced training (and test) set, we randomly selected 90% (and 10%) of data from each of the seven domains in MFTC, resulting in the label distribution in Table 2. Data pre-processing, hyperparameters, and training environment are detailed in Appendix A. The code is available on GitHub[1].

We first inspect the embedding space itself to evaluate whether a supervised approach is needed to disentangle the MFT elements in the MFTC training set (intrinsic evaluation). Then, to evaluate whether the relationships among MFT elements have been successfully captured, we test the embedding space on two downstream tasks (as suggested by Eger et al. (2019)) (extrinsic evaluation).

4.1 Intrinsic Evaluation

We investigate the embedding space by (1) showing a visualization of the training set data in the embedding space to gain an intuitive understanding of the relationships among MFT elements, and (2) computing a moral similarity table to inspect quantitative similarities among MFT elements. To show the effect of supervised labels during training, we compare (a) an off-the-shelf pre-trained supervised SimCSE embedding space, and the embeddings trained with (b) the unsupervised SimCSE and (c) the supervised SimCSE approaches.

---

[1] https://github.com/jeongwoopark0514/morality-is-non-binary

Dataset	Care	Harm	Fairness	Cheating	Loyalty	Betrayal	Authority	Subversion	Purity	Degradation	Non-moral
Train	2176	3269	1870	3068	1736	1736	1294	1816	698	1246	14428
Test	240	359	204	335	183	121	137	196	72	132	1611

Table 2: Distribution of MFT labels in the training and test sets used to train and evaluate SimCSE moral embeddings.

4.1.1 Visualization

We explore the relationships between the MFT elements in the embedding space through visual insight. Since the SimCSE embedding space is 1024-dimensional, we employ the Uniform Manifold Approximation and Projection (UMAP) method (McInnes et al., 2020), a nonlinear dimensionality reduction technique, to reduce the embedding space to two dimensions. We choose UMAP as it preserves both local and most of the global structure in the data, with a shorter run-time when compared to other dimensionality reduction techniques such as t-SNE and PCA (McInnes et al., 2020). We show all the data points in the MFTC training set in a two-dimensional plot and qualitatively discuss the relationships among MFT elements.

4.1.2 Moral Similarity

We perform a moral similarity task, inspired by the popular semantic similarity task (Agirre et al., 2013; Gao et al., 2021), to measure the similarity between moral elements using the MFTC training set. To calculate the moral similarity between two MFT elements m and n, we compute the cosine similarity between the moral embedding representations of each data point annotated with m and each data point annotated with n, and report the mean result. We apply the procedure for all combinations of the ten MFT elements plus the *non-moral* label, resulting in an 11x11 table of mean similarities.

4.2 Extrinsic Evaluation

To evaluate whether the relationships among MFT elements have been effectively captured in the embedding spaces, we evaluate (1) the generalizability on the held-out test set, and (2) the consistency between the embeddings and the Moral Foundation Dictionary 2.0 (MFD2.0) (Frimer, 2019), an independently collected MFT dictionary. As in Section 4.1, we compare (a) an off-the-shelf pre–trained SimCSE embedding space, and the embeddings trained with (b) the unsupervised SimCSE and (c) the supervised SimCSE approaches.

4.2.1 Generalizability on Test Set

We evaluate the moral embedding spaces on the MFTC test set to assess the generalizability to unseen data. As for the intrinsic evaluation described above, we evaluate the embedding spaces (1) via a visualization by plotting the MFTC test set on the embedding space and visualizing it via a UMAP plot, and (2) with a moral similarity table.

4.2.2 Comparison to MFD2.0

We measure the consistency of the generated moral embedding spaces with MFD2.0, a dictionary manually created by the authors of the MFT (Graham et al., 2013), containing sets of words representative of each MFT moral element.

Clustering We collect all words belonging to the MFD2.0 and use K-means clustering to test whether meaningful clusters can be discerned based on the words' embedding representations based on their Euclidean distance (we choose Euclidean since the K-means algorithm may not converge with other distances without data transformation).

First, we measure the coherence of the clusters via the silhouette coefficient (Rousseeuw, 1987):

$$s = \frac{\sum_{i=1}^{N} \frac{b(i) - a(i)}{max(a(i), b(i))}}{N} \qquad (2)$$

where N is the number of samples, $a(i)$ the mean intra-cluster distance and $b(i)$ the mean nearest-cluster distance for sample i. The coefficient ranges from -1 to 1. For each tested approach, we plot the silhouette coefficient for K ranging from 2 to 15 and choose $\hat{K}$ as the optimal number of clusters with the highest silhouette score.

Then, we measure the quality of the clusters via the purity score (Manning, 2009). To calculate the purity of a cluster, we first find the most frequent true label (L_f) of each cluster. Then, we sum the number of words labeled with L_f for each cluster and divide the sum by the total number of words in the dictionary. Thus, a high purity score indicates that the clusters primarily consist of words with the same label. However, the purity score tends to increase as K increases, since each cluster is

(a) Off-the-shelf SimCSE   **(b) Unsupervised SimCSE**   **(c) Supervised SimCSE**

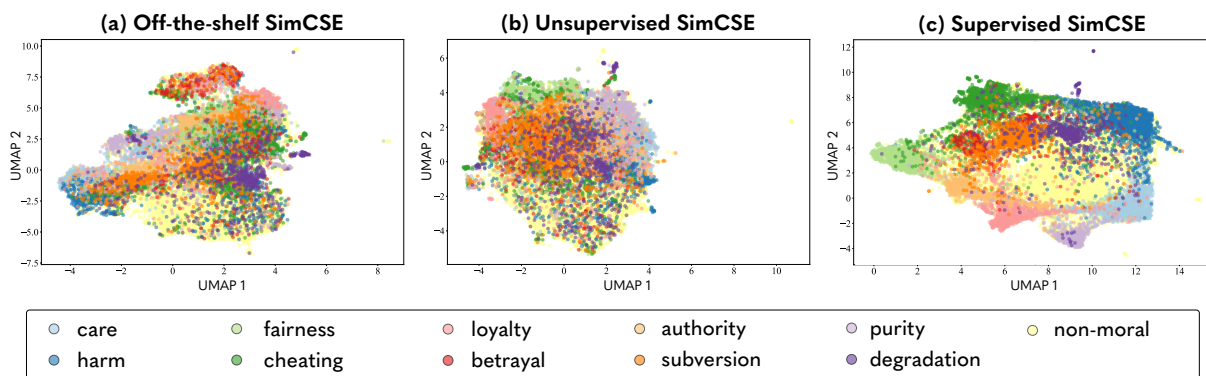| ○ care | ○ fairness | ○ loyalty | ○ authority | ○ purity | ○ non-moral |
| ○ harm | ○ cheating | ○ betrayal | ○ subversion | ○ degradation | |

Figure 2: UMAP plot of the **MFTC training set** data with off-the-shelf pre-trained SimCSE model (a, left), unsupervised SimCSE approach (b, middle), and supervised SimCSE approach (c, right).

at the purest state when there is only one item in the cluster. Due to this tradeoff between K and the clustering quality, we evaluate the clustering results via both the silhouette coefficient and the mean purity score over the clusters. We report the results for $K = \hat{K}$ and $K = 10$ (as the MFT taxonomy is composed of ten elements).

Moral Similarity (MFD2.0) We measure the similarity among the MFD2.0 words belonging to different MFT elements via moral similarity, as in Section 4.1.2. To calculate the moral similarity between two MFT elements m and n, we compute the cosine similarity between the moral embedding representations of each MFD2.0 word belonging to m and each MFD2.0 word belonging to n, and report the mean result. We apply the procedure for all combinations of the ten MFT elements, resulting in a 10x10 table of mean similarity.

5 Results and Discussion

We report the results of the intrinsic evaluations to judge the effect of supervised training, and the results of the extrinsic evaluation to assess the moral embeddings when used with external data.

5.1 Intrinsic Evaluation

We present the results of visualization and moral similarity evaluations on the MFTC training set.

5.1.1 Visualization

Figure 2 shows the dimension-reduced UMAP plot of the MFTC training set data mapped on the moral embedding spaces (a) resulting from the off-the-shelf pre-trained supervised SimCSE model, or trained with (b) the unsupervised SimCSE approach or (c) the supervised SimCSE approach.

We notice that the supervised approach (Figure 2c) shows distinguishable clusters for each vice and virtue element, exhibiting a visible improvement when compared to the off-the-shelf model (Figure 2a). However, the unsupervised approach (Figure 2b) displays no discernible clusters.

In Figure 2c, we observe a clear separation between virtues (located in the bottom half of the plot) and vices (located in the top half). Further, the values within the same foundation (e.g., *care-harm*) tend to be in symmetrical locations in the virtues and vices areas. Finally, tweets labeled as *non-moral* are spread throughout the plot, especially in the area between the vice and virtue clusters.

The noticeable difference between the off-the-shelf, unsupervised, and supervised approaches suggests that a CL-based moral embedding space can capture the relationships between virtues and vices and among moral foundations when employing label information. We investigate this further via a quantitative moral similarity evaluation.

5.1.2 Moral Similarity

To further analyze the insightful results observed with the supervised approach, we report in Table 3 the moral similarity across MFT elements calculated with the supervised SimCSE moral embedding representations of the MFTC training set. This table allows us to inspect in more detail the similarity across the different moral elements.

First, we notice a high similarity along the diagonal, indicating that the moral embedding space consistently clusters data points annotated with the same label. Further, the overall similarity between virtues and vices values (top-right and bottom-left quadrants) is visibly lower than the similarity between virtue-virtue (top-left quadrant) and vice-
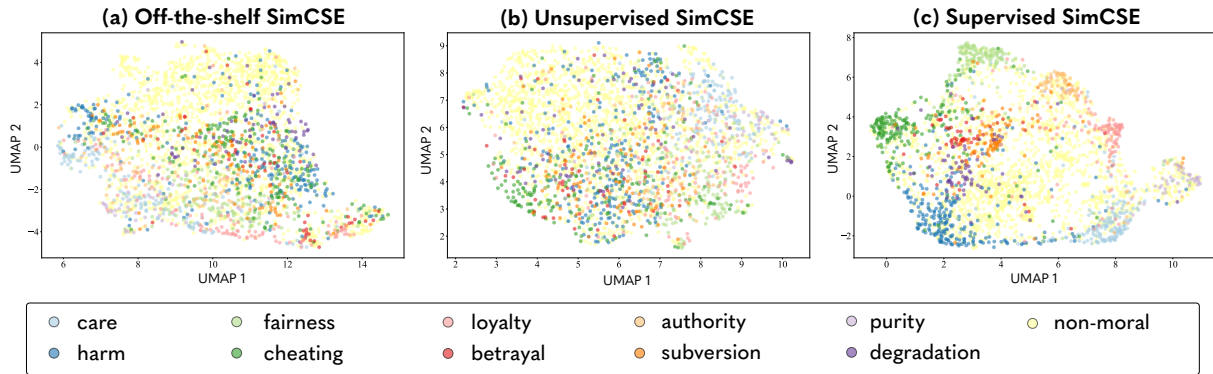
Figure 3: UMAP plot of the **MFTC test set** data with off-the-shelf pre-trained SimCSE model (a, left), unsupervised SimCSE approach (b, middle), and supervised SimCSE approach (c, right).

	Care	Fairness	Loyalty	Authority	Purity	Harm	Cheating	Betrayal	Subversion	Degradation	Non-moral
Care	81.2	25.4	41.0	35.2	49.5	27.6	4.7	21.0	15.2	11.6	28.8
Fairness	25.4	77.9	28.8	43.0	29.1	12.7	34.6	19.2	22.4	10.9	26.4
Loyalty	41.0	28.8	65.0	37.7	36.2	9.7	8.5	27.7	19.1	8.7	27.0
Authority	35.2	43.0	37.7	68.7	40.5	11.3	14.4	25.4	37.4	14.1	27.3
Purity	49.5	29.1	36.2	40.5	79.3	13.2	5.2	15.5	17.5	22.4	27.2
Harm	27.6	12.7	9.7	11.3	13.2	56.9	27.2	35.5	30.2	31.7	30.0
Cheating	4.7	34.6	8.5	14.4	5.2	27.2	58.9	40.8	35.8	32.7	26.8
Betrayal	21.0	19.2	27.7	25.4	15.5	35.5	40.8	58.3	50.6	35.7	32.5
Subversion	15.2	22.4	19.1	37.4	17.5	30.2	35.8	50.6	57.9	36.2	30.7
Degradation	11.6	10.9	8.7	14.1	22.4	31.7	32.7	35.7	36.2	46.5	28.5
Non-moral	28.8	26.4	27.0	27.3	27.2	30.0	26.9	32.5	30.7	28.5	30.8

Table 3: Moral similarity for MFTC training set with supervised SimCSE. Darker the cell higher the similarity.

vice values (bottom-right quadrant), which indicates that the model can clearly separate virtues and vices found in tweets. Moreover, a significant similarity between opposing virtues and vices (e.g., *fairness* and *cheating*) can be observed, showing that the embedding space has learned relationships among corresponding virtues and vices. Finally, the similarity between non-moral and moral values is modest, confirming that tweets labeled as *non-moral* are spread throughout the embedding space, without forming any significant cluster.

The results described above show the effectiveness of the training strategy described in Section 3. However, additional emergent results can be observed in Table 3. For instance, on the diagonal, virtue values (top-left quadrant) have a higher similarity than vice values (bottom-right quadrant), showing that tweets labeled with virtue values are more consistently clustered. Moreover, we observe that some elements have a high similarity despite not having been explicitly addressed by the training strategy, e.g., *care-purity* and *subversion-betrayal*.

To further investigate these similarities, we tokenize and lemmatize the tweets labeled with these elements and inspect whether they share commonly used lemmas. We provide some insightful examples to better understand such similarities. The word 'god' appears consistently in tweets labeled with *care* and *purity*, hinting that the correlation is driven by common concerns of religion and care, especially in the context of the Sandy hurricane relief tweets. The words 'Obama' and 'protest' are common for both *betrayal* and *subversion* tweets, showing how the correlation was driven by the political background behind tweets collected with the All Lives Matter and Black Lives Matter hashtags.

Lastly, similar to Figure 2, the moral similarity tables obtained with the off-the-shelf model and with the unsupervised SimCSE approach fail to produce meaningful similarities (see Appendix B.1.2).

5.2 Extrinsic Evaluation

We present the results of generalizability on the test set and comparison to MFD2.0 dictionary.

5.2.1 Generalizability on Test Set

Figure 3 shows the UMAP plot of the MFTC test set data mapped on the embedding spaces obtained with the three compared approaches. First, we remark that the lower density of the plotted data with respect to Figure 2 is due to the smaller size of the test set compared to the training set. Further, with the supervised SimCSE approach, we observe clear clusters corresponding to the MFT elements (similar to Figure 2c). Instead, the UMAP plots resulting from the off-the-shelf model and the unsupervised approach show no distinguishable clusters.

To quantitatively investigate the relationships among the MFT elements, we show in Table 4 the

659

	Care	Fairness	Loyalty	Authority	Purity	Harm	Cheating	Betrayal	Subversion	Degradation	Non-moral
Care	75.2	26.7	41.6	37.0	49.8	28.4	7.7	20.0	17.1	12.6	29.5
Fairness	26.7	72.0	28.1	41.3	30.8	15.6	35.1	22.1	24.1	15.2	26.5
Loyalty	41.6	28.1	60.8	37.8	37.0	12.6	10.3	26.9	19.9	11.6	27.6
Authority	37.0	41.3	37.8	62.9	42.4	14.7	16.2	23.9	34.3	19.1	27.7
Purity	49.8	30.8	37.0	42.4	75.5	15.1	6.3	13.9	17.6	18.7	27.6
Harm	28.4	15.6	12.6	14.7	15.1	52.1	26.4	35.0	32.2	32.5	30.2
Cheating	7.7	35.1	10.3	16.2	6.3	26.4	56.4	41.5	34.5	33.5	26.2
Betrayal	20.0	22.1	26.9	23.9	13.9	35.0	41.5	56.8	46.9	39.3	31.8
Subversion	17.1	24.1	19.9	34.3	17.6	32.2	34.5	46.9	51.8	40.4	30.4
Degradation	12.6	15.2	11.6	19.1	18.7	32.5	33.5	39.3	40.4	46.5	29.7
Non-Moral	29.5	26.5	27.6	27.7	27.6	30.2	26.2	31.8	30.4	29.7	30.9

Table 4: Moral similarity for MFTC test set with supervised SimCSE. Darker the cell higher the similarity.

moral similarity for the MFTC test set with the supervised SimCSE approach. These results are in line with Table 3, and show that the distribution of the MFT elements learned in the training set is consistent with the data in the test set.

5.2.2 Comparison to MFD2.0

We present the results of the clustering of the MFD2.0 words based on the three compared approaches (as described in Section 4.2). We further inspect the best-performing approach through the moral similarity evaluation of the MFD2.0 words.

Clustering Figure 4 shows the silhouette coefficient for K-means clustering with K ranging from 2 to 15 for the three compared approaches.
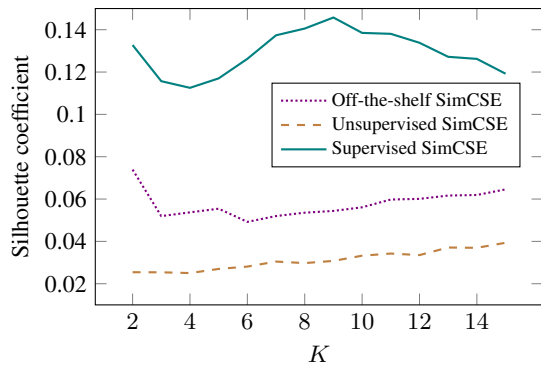

Figure 4: Silhouette coefficients for K ranging from 2 to 15 for the three compared approaches.

We observe that the supervised SimCSE approach performs best, with a silhouette coefficient that peaks at $K = 9$, close to the total number of MFT elements (10). Instead, the off-the-shelf model peaks at $K = 2$, aligning with previous research results that show that the pre-trained embedding spaces contain an intuitive distinction between

do's and don'ts (Schramowski et al., 2022). Further, we observe low silhouette coefficients due to the high dimensionality of the embedding space.

Table 5 shows purity and silhouette coefficients for $K = \hat{K}$ (the K that leads to the highest silhouette coefficient) and $K = 10$. The supervised SimCSE approach achieves the highest purity score for both $K = \hat{K}$ and $K = 10$, resulting in a purity of 0.71 in both cases. This result shows that the resulting embedding space allows for a coherent clustering of the MFD2.0 words, proving consistent with an independently generated MFT dictionary.

	Approach	K	Purity	Silhouette
$K = \hat{K}$	Off-the-shelf SimCSE	2	0.30	0.07
	Unsupervised SimCSE	15	0.51	0.04
	Supervised SimCSE	9	**0.71**	**0.15**
$K = 10$	Off-the-shelf SimCSE	10	0.56	0.06
	Unsupervised SimCSE	10	0.45	0.03
	Supervised SimCSE	10	**0.71**	**0.14**

Table 5: Purity and Silhouette coefficients for $K = \hat{K}$ and $K = 10$. The best scores are highlighted in bold.

Moral Similarity (MFD2.0) We further investigate the consistency between the supervised SimCSE embedding space approach and MFD2.0. Table 6 shows the moral similarity between the MFT elements, calculated with the supervised SimCSE embedding space representation of MFD2.0 words.

	Care	Fairness	Loyalty	Authority	Purity	Harm	Cheating	Betrayal	Subversion	Degradation
Care	57.8	30.7	36.7	32.4	39.4	30.1	18.9	23.2	19.4	22.4
Fairness	30.7	48.3	33.0	37.5	32.5	25.1	30.3	27.8	27.5	22.2
Loyalty	36.7	33.0	50.9	35.8	38.3	26.5	24.6	33.4	31.9	27.4
Authority	32.4	37.5	35.8	48.2	40.0	26.1	25.5	31.3	36.4	27.4
Purity	39.4	32.5	38.3	40.0	57.2	27.0	21.2	27.4	30.7	35.0
Harm	30.1	25.1	26.5	26.1	27.0	56.4	35.9	35.6	33.5	41.8
Cheating	18.9	30.3	24.6	25.5	21.2	35.9	52.4	45.9	40.9	39.3
Betrayal	23.2	27.8	33.4	31.3	27.4	35.6	45.9	54.9	51.0	39.3
Subversion	19.4	27.5	31.9	36.4	30.7	33.5	40.9	51.0	56.5	41.1
Degradation	22.4	22.2	27.4	27.4	35.0	41.8	39.3	39.3	41.1	53.9

Table 6: Moral similarity for MFD2.0 with supervised SimCSE. Darker the cell higher the similarity.

The high similarity along the diagonal indicates that MFD2.0 words that represent the same moral value are closer in embedding space with respect to words that represent different moral values. Further, we notice parallels with Table 3. That is, (1) the similarity between virtues and virtues (top-left quadrant) and vices and vices (bottom-right quadrant) is greater than the similarity between virtues

and vices (top-right and bottom-left quadrants), and (2) there is a noticeable similarity between corresponding virtues and vices (e.g., *authority* and *subversion*). These results confirm that the supervised SimCSE approach generates moral embeddings that align with an independently generated MFT dictionary, whereas the off-the-shelf and unsupervised approaches fail to do so (Appendix B.2.2).

6 Related Works

We review previous research on methods for detecting moral values and existing moral datasets.

6.1 Detecting Moral Values in Text

Traditionally, value lexicons—sets of words descriptive of each moral element—have been used to detect morality through text similarity (Bahgat et al., 2020; Pavan et al., 2020). Graham et al. (2009) developed the Moral Foundations Dictionary (MFD), which has been extended manually (Frimer, 2019) and via semi-automated methods (Rezapour et al., 2019; Araque et al., 2020; Kobbe et al., 2020; Hopp et al., 2020). However, word-level lexicons are limited by the ambiguity of natural language and the restricted range of lemmas, which can be solved by projecting the MFD lexicon on knowledge graphs that link moral entities and concepts (Hulpuș et al., 2020; Asprino et al., 2022). Other methods instead use the supervised classification paradigm (Lin et al., 2018; Johnson and Goldwasser, 2018; Hoover et al., 2020), exploiting an annotated dataset to train a classifier. In particular, BERT-based models have been successfully used on datasets annotated with the MFT taxonomy (Kobbe et al., 2020; Alshomary et al., 2022; Liscio et al., 2022a; Huang et al., 2022; Bulla et al., 2023).

Similar to our work, Priniski et al. (2021) map text onto a 10-dimensional space (corresponding to the MFT elements) where the position of a word in each dimension is determined by the moral valence that FrameAxis (an MFT-based lexicon (Kwak et al., 2021)) attributes to the word for the corresponding MFT element. Our work differs in that we use state-of-the-art pre-trained 1024-dimensional sentence embeddings that have been shown to be more effective at capturing semantic similarity compared to lexicon-based approaches.

6.2 Datasets with Moral Content

Besides the MFTC, other datasets based on different moral value taxonomies have been collected for NLP applications. The Schwartz value theory (Schwartz, 2012) is another commonly used taxonomy, composed of 20 values that form a continuum of meaning in a circumplex. Kiesel et al. (2022) presented a dataset of 5,270 arguments labeled with the Schwartz values and extended it to over 9K arguments for the SemEval-2023 Task 4 (Kiesel et al., 2023). Qiu et al. (2022) collected a dataset of dialogues in different social scenarios, also annotated with the Schwartz values. Jin et al. (2022) proposed MoralExceptQA, the novel challenge and dataset on moral exception question answering. Finally, Hendrycks et al. (2021) introduced a dataset with contextualized scenarios about commonsense moral intuitions. We opted for MFT and MFTC due to the strong psychological background and the availability of a large annotated dataset.

7 Conclusions and Future Work

AI agents ought to recognize the diversity and nuances of human moral perspectives. To this end, we propose a method to generate a pluralist moral sentence embedding space with a state-of-the-art contrastive learning approach and focus on its evaluation. First, we perform an intrinsic evaluation to evaluate the significance of label information for distinguishing among the different elements of pluralist morality. Our results show that a pluralist approach to morality cannot be simply learned through self-supervised learning, but human labels are essential. Then, we demonstrate that the embedding space trained through label supervision is aligned with externally sourced data such as an independently created lexicon of words that are descriptive of a pluralist approach to morality.

Our investigation opens avenues for incorporating a pluralist approach to morality in language models, overcoming a simplistic, binary interpretation, i.e., simply judging a situation as morally right or wrong. Pluralist moral embeddings can be used in a variety of applications, e.g., recognizing moral rhetoric from diverse social issues such as abortion and terrorism (Sagi and Dehghani, 2014), generating morally-aligned language (Ammanabrolu et al., 2022; Lorandi and Belz, 2023), measuring disagreement in online discussions (Shortall et al., 2022; van der Meer et al., 2023), and investigating the context specificity of moral judgment (Liscio et al., 2022b, 2023a) or the cultural influences on moral norms (Ramezani and Xu, 2023). Furthermore, the detection of pluralist morality could be extended

with Hybrid Intelligence approaches (Akata et al., 2020), aiming at devising AI systems that combine human and artificial intelligence by design (e.g., van der Meer et al. (2022); Siebert et al. (2022)).

Our experiments are limited to one dataset and one approach to moral pluralism. However, our experimental setup can be extended to other corpora to assess the generalizability to other approaches to pluralist morality. For instance, a comparative analysis would reveal differences between discrete and fuzzy approaches to moral pluralism, e.g., by comparing the MFT and the Schwartz value theory (Schwartz, 2012). Similarly, we chose SimCSE due to its proven efficacy, but additional CL approaches could extend our work, e.g., by incorporating label embeddings in the training procedure (Zhang et al., 2022a) or by exploiting adversarial examples to improve generalizability (Zhan et al., 2023). Finally, the MFTC was annotated by multiple annotators and we used the majority agreement to train the moral embedding space. To better reflect the subjective nature of morality, an avenue for future work is to employ all annotations, incorporating annotators' (dis)agreement through a perspectivist approach (Uma et al., 2022; Cabitza et al., 2023).

8   Ethical Considerations and Limitations

Morally-charged content poses a significant challenge for language models (Jin et al., 2022). This is particularly problematic when models trained to discern descriptive ethics (i.e., understand how humans reason about moral judgments) are used for normative ethics, (i.e., to make moral judgments such as religious prescriptions and medical advice) (Talat et al., 2022). For this reason, in this work, we limit ourselves to descriptive ethics. Further, the usage of our embedding space in highly sensitive domains, such as the legal field, requires additional cautious deliberation (Leins et al., 2020).

An additional challenge is introduced by the *dual-use* problem (Hovy and Spruit, 2016), that is when a system developed for a certain purpose leads to unintended negative consequences in another application. For instance, since liberals and conservatives rely on different moral foundations (Graham et al., 2009), the moral embedding space can be misused to identify and discriminate against people with certain political standpoints.

Next, we recognize the limitations regarding the dataset we use, the MFTC. First of all, the MFTC is composed of English tweets about US-centric topics, thus perpetuating Western biases (Mehrabi et al., 2021). Post-hoc debiasing techniques (Liang et al., 2020) can be applied to the current moral embedding space, preventing the need for re-training with large amounts of additional data. However, our method and evaluation procedure can be applied to larger and culturally diverse datasets as well. Then, the MFTC annotation procedure resulted in a low annotator agreement, which is to be expected in such a subjective annotation task (Hoover et al., 2020). Choosing the majority label as the true label reinforces the domination of the majority, suppressing the minority views. Employing a perspectivist approach, using all the annotations when training, can improve the representativity of the embedding space (Cabitza et al., 2023).

Finally, we recognize concerns on the evaluation procedure. First, the MFT dictionary (MFD2.0) is based on the WEIRD (Western, Educated, Industrialized, Rich, Democratic) sample. Dictionaries created from more diverse samples could reveal new strengths and weaknesses of the embedding space. Second, we used UMAP to easily visualize the embedding space and the effect of the training. Additional investigation is required for a detailed geometric analysis of the embedding space.

Acknowledgements

References

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Zeynep Akata, Dan Balliet, Maarten De Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, et al. 2020. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8):18–28.

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.

Prithviraj Ammanabrolu, Liwei Jiang, Maarten Sap, Hannaneh Hajishirzi, and Yejin Choi. 2022. Aligning to social norms and values in interactive narratives. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5994–6017, Seattle, United States. Association for Computational Linguistics.

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.

Luigi Asprino, Luana Bulla, Stefano De Giorgis, Aldo Gangemi, Ludovica Marinucci, and Misael Mongiovi. 2022. Uncovering values: Detecting latent moral content from natural language with explainable and non-trained methods. In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 33–41, Dublin, Ireland and Online. Association for Computational Linguistics.

Mohamed Bahgat, Steven R. Wilson, and Walid Magdy. 2020. Towards Using Word Embedding Vector Space for Better Cohort Analysis. In *Proceedings of the International AAAI Conference on Web and Social Media*, ICWSM '20, pages 919–923, Atlanta, Georgia. AAAI Press.

Luana Bulla, Stefano De Giorgis, Aldo Gangemi, Ludovica Marinucci, and Misael Mongiovì. 2023. Detection of Morality in Tweets Based on the Moral Foundation Theory. In *Machine Learning, Optimization, and Data Science: 8th International Conference*, LOD '22, pages 1–13. Springer Nature Switzerland.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI '23, pages 6860–6868.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, ICWSM '15, pages 512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages

4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Steffen Eger, Andreas Rücklé, and Iryna Gurevych. 2019. Pitfalls in the evaluation of sentence embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 55–60, Florence, Italy. 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. 2019. Moral foundations dictionary 2.0.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and Machines*, 30:411–437.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In *Advances in Experimental Social Psychology*, volume 47, pages 55–130. Elsevier, Amsterdam, the Netherlands.

Jesse Graham, Jonathan Haidt, and Brian Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96:1029–46.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2161–2174, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI With Shared Human Values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

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*, 11:1057–1071.

Frederic R. Hopp, Jacob T. Fisher, Devin Cornell, Richard Huskey, and René Weber. 2020. 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.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

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, pages 121–131. 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.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, Yulia Tsvektov, Oren Etzioni, Maarten Sap, Regina Rini, and Yejin Choi. 2022. Can machines learn morality? the delphi experiment. *arXiv preprint arXiv:2110.07574*.

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 *Advances in Neural Information Processing Systems*, volume 35, pages 28458–28473. Curran Associates, Inc.

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.

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 *17th International Workshop on Semantic Evaluation*, SemEval '23, pages 2290–2306, 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.

Haewoon Kwak, Jisun An, Elise Jing, and Yong-Yeol Ahn. 2021. FrameAxis: characterizing microframe bias and intensity with word embedding. *PeerJ Computer Science*, 7:e644.

Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. 2020. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934.

Kobi Leins, Jey Han Lau, and Timothy Baldwin. 2020. Give me convenience and give her death: Who should decide what uses of NLP are appropriate, and on what basis? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2908–2913, Online. Association for Computational Linguistics.

Ruiqi Li, Xiang Zhao, and Marie-Francine Moens. 2022. A brief overview of universal sentence representation methods: A linguistic view. *ACM Comput. Surv.*, 55(3).

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online. Association for Computational Linguistics.

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

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

Enrico Liscio, Alin Dondera, Andrei Geadau, Catholijn Jonker, and Pradeep Murukannaiah. 2022a. Cross-domain classification of moral values. In *Findings of the Association for Computational Linguistics:*

NAACL 2022, pages 2727–2745, Seattle, United States. Association for Computational Linguistics.

Enrico Liscio, Roger Lera-Leri, Filippo Bistaffa, Roel I.J. Dobbe, Catholijn M. Jonker, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, and Pradeep K. Murukannaiah. 2023b. Value Inference in Sociotechnical Systems. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, pages 1774–1780, London, United Kingdom. IFAAMAS.

Enrico Liscio, Michiel van der Meer, Luciano C Siebert, Catholijn M Jonker, and Pradeep K Murukannaiah. 2022b. What values should an agent align with? An empirical comparison of general and context-specific values. *Autonomous Agents and Multi-Agent Systems*, 36(1):23.

Michela Lorandi and Anya Belz. 2023. How to Control Sentiment in Text Generation: A Survey of the State-of-the-Art in Sentiment-Control Techniques. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 341–353, Toronto, Canada. Association for Computational Linguistics.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI '21, pages 13470–13479.

Christopher D Manning. 2009. *An introduction to information retrieval*. Cambridge university press.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6).

Jeongwoo Park, Enrico Liscio, and Pradeep K. Murukannaiah. 2024. Morality is Non-Binary: Building a Pluralist Moral Sentence Embedding Space using Contrastive Learning - models. 4TU.ResearchData.

Matheus C. Pavan, Vitor G. Santos, Alex G. J. Lan, Joao Martins, Wesley Ramos Santos, Caio Deutsch, Pablo B. Costa, Fernando C. Hsieh, and Ivandre Paraboni. 2020. Morality Classification in Natural Language Text. *IEEE Transactions on Affective Computing*, 3045(c):1–8.

J. Hunter Priniski, Negar Mokhberian, Bahareh Harandizadeh, Fred Morstatter, Kristina Lerman, Hongjing Lu, and P. Jeffrey Brantingham. 2021. Mapping Moral Valence of Tweets Following the Killing of George Floyd. In *Poceedings of the ICWSM International Workshop on Social Sensing*, SocialSens '21. Association for the Advancement of Artificial Intelligence.

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)*, ACL '23, pages 11253–11271.

Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. 2022. ValueNet: A New Dataset for Human Value Driven Dialogue System. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11183–11191.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '23, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

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 3982–3992, Hong Kong, China. 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.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Eyal Sagi and Morteza Dehghani. 2014. Measuring moral rhetoric in text. *Soc. Sci. Comput. Rev.*, 32(2):132–144.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin Rothkopf, and Kristian Kersting. 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. 2012. An Overview of the Schwartz Theory of Basic Values. *Online readings in Psychology and Culture*, 2(11):1–20.

Ruth Shortall, Anatol Itten, Michiel van der Meer, Pradeep Murukannaiah, and Catholijn Jonker. 2022. Reason against the machine? Future directions for mass online deliberation. *Frontiers in Political Science*, 4:946589.

Luciano C Siebert, Enrico Liscio, Pradeep K Murukannaiah, Lionel Kaptein, Shannon Spruit, Jeroen Van Den Hoven, and Catholijn Jonker. 2022. Estimating Value Preferences in a Hybrid Participatory System. In *HHAI2022: Augmenting Human Intellect*, pages 114–127. IOS Press.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. On the machine learning of ethical judgments from natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 769–779, Seattle, United States. Association for Computational Linguistics.

Jake Telkamp and Marc Anderson. 2022. The Implications of Diverse Human Moral Foundations for Assessing the Ethicality of Artificial Intelligence. *Journal of Business Ethics*, 178:961–976.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from disagreement: A survey. *J. Artif. Int. Res.*, 72:1385–1470.

Michiel van der Meer, Enrico Liscio, Catholijn M. Jonker, Aske Plaat, Piek Vossen, and Pradeep K. Murukannaiah. 2022. Hyena: A hybrid method for extracting arguments from opinions. In *HHAI2022: Augmenting Human Intellect*, pages 17–31. IOS Press.

Michiel van der Meer, Piek Vossen, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2023. Do Differences in Values Influence Disagreements in Online Discussions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, EMNLP '23, pages 15986–16008, Singapore. Association for Computational Linguistics.

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.

Pengwei Zhan, Jing Yang, Xiao Huang, Chunlei Jing, Jingying Li, and Liming Wang. 2023. Contrastive Learning with Adversarial Examples for Alleviating Pathology of Language Model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '23, pages 6493–6508, Toronto, Canada. Association for Computational Linguistics.

Zhenyu Zhang, Yuming Zhao, Meng Chen, and Xiaodong He. 2022a. Label anchored contrastive learning for language understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1437–1449, Seattle, United States. Association for Computational Linguistics.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad Reza Namazi Rad. 2022b. Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3886–3893, Seattle, United States. Association for Computational Linguistics.

A Experimental Details

For the sake of reproducibility, we share further details on our experimental procedure. The trained models are available online (Park et al., 2024).

A.1 SimCSE Contrastive Losses

We present the SimCSE contrastive losses as introduced by Gao et al. (2021). For unsupervised SimCSE, we take a collection of sentences $\{x_i\}_{i=1}^{m}$, and uses $x_i^+ = x_i$. It constructs a positive pair for each input x_i by encoding the input twice using different dropout masks, z and z'. We denote $\mathbf{h}_i^z = f_\theta(x_i, z)$, where z is a random mask for dropout. Note that in the standard transformer models, there are dropout masks placed on fully-connected layers. The training objective for the unsupervised SimCSE approach is the following:

$$\ell_i = -\log \frac{e^{\operatorname{sim}\left(\mathbf{h}_i^{z_i}, \mathbf{h}_i^{z_i'}\right)/\tau}}{\sum_{j=1}^{N} e^{\operatorname{sim}\left(\mathbf{h}_i^{z_i}, \mathbf{h}_j^{z_j'}\right)/\tau}},$$

For supervised SimCSE, instead of using dropout, it takes predefined positive and negative instances, x_i^+ and x_i^- respectively. The training objective for the supervised SimCSE approach is the following:

$$\ell_i = -\log \frac{e^{\operatorname{sim}\left(\mathbf{h}_i, \mathbf{h}_i^+\right)/\tau}}{\sum_{j=1}^{N} \left(e^{\operatorname{sim}\left(\mathbf{h}_i, \mathbf{h}_j^+\right)/\tau} + e^{\operatorname{sim}\left(\mathbf{h}_i, \mathbf{h}_j^-\right)/\tau}\right)}$$

A.2 Data Processing

We preprocess the tweets by removing URLs, emails, usernames, and mentions. Next, we employ the Ekphrasis package[2] to correct common spelling mistakes and unpack contractions. Finally, emojis are transformed into their respective words using the Python Emoji package[3]. Moreover, there are some independent tweets with duplicated content, in some cases with different labels. We reduced repeated instances of distinct tweet annotations to one instance by applying a majority vote. The final unsupervised SimCSE training set consists of 29,147 triples (i.e., the size of the training set). The final supervised SimCSE training set consists of 5,304 triples, due to the large number of *non-moral* labels (Table 2) that did not appear in any triple.

---

[2] https://github.com/cbaziotis/ekphrasis
[3] https://pypi.org/project/emoji/

A.3 Hyperparameters

To select the most optimal combination of hyperparameters for SimCSE, we perform a grid search based on the F_1-scores of the classification result, which is further discussed in Appendix B.2.3. Table A1 and Table A2 show the hyperparameters that were compared, highlighting in bold the best-performing option. We used these hyperparameters for every experiment in this paper for consistency. If a parameter is not present in the table, the default value supplied by the framework[4] was used.

Hyperparameters	Options
Model name	sup-simcse-bert-large-uncased
Max Sequence Length	**64**, 128
Epochs	**2**, 3, 5
Batch Size	16, **32**
Learning Rate	5×10^{-5}
Temperature	0.01, 0.05, **0.1**
Pooler	**cls**

Table A1: Hyperparameters tested for training SimCSE with the supervised approach.

Hyperparameters	Options
Model name	unsup-simcse-bert-large-uncased
Max Sequence Length	**64**, 128
Epochs	**1**, 2, 3
Batch Size	16, **32**
Learning Rate	3×10^{-5}
Temperature	0.01, **0.05**, 0.1
Pooler	**cls**

Table A2: Hyperparameters tested for training SimCSE with the unsupervised approach.

The time taken for the supervised SimCSE hyperparameter search is roughly 6-7 hours, and the time taken for the unsupervised SimCSE hyperparameter search is approximately 15-16 hours.

A.4 Computing Infrastructure

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

- PyTorch: 1.13.0

- Huggingface's Transformers: 4.2.1

- SimCSE: 0.4

- NVIDIA A40 GPU

- CUDA 11.6

---

[4] https://github.com/princeton-nlp/SimCSE

A.5 Random Seeds

In our experiments, we ensure that the same train-test splits are used across different runs of each experiment. Further, to control for any randomness throughout code execution, we fixed the random seeds (to 42) in the following libraries:

- Python (`random.seed`);

- NumPy (`numpy.random.seed`);

- PyTorch (`torch.manual_seed`);

- Tensorflow
  (`tensorflow.random.set_seed`).

A.6 Artifacts Used

We primarily use two different types of artifacts, data and models.

MFTC is a collection of 35,108 tweets annotated based on MFT (Hoover et al., 2020). MFTC can be accessed[5] and used under Creative Commons Attribution 4.0 license. MFD2.0 (Frimer, 2019) can be freely accessed[6].

SimCSE (Gao et al., 2021) can be used under MIT license[7]. BERT (Devlin et al., 2019) is used as a baseline model to compare with SimCSE. The license of BERT is Apache License 2.0[8].

B Extended Results

We extend the results shown in the main paper for intrinsic and extrinsic evaluation.

B.1 Intrinsic Evaluation

We provide additional visualizations and quality metrics of the trained embedding spaces.

B.1.1 Visualization

Figures B1 and B2 show the UMAP plot of the MFTC training set mapped on the off-the-shelf SimcSE model the supervised SimCSE approach, respectively. The figures are similar to Figure 2, however grouping the 10 moral elements as vices or virtues.

Figure B1 does not show any distinguishable cluster. Instead, Figure B2 shows a clearer separation between vice and virtue elements—vice and virtue clusters are less mixed together, and a bigger gap can be found between them.

Figure B1: UMAP plot of MFTC training set with the off-the-shelf SimCSE model (only vices and virtues).

Figure B2: UMAP plot of MFTC training set with the supervised SimCSE approach (only vices and virtues).

B.1.2 Moral Similarity

In the main paper we show the moral similarity table for the supervised SimCSE approach, here we show for the off-the-shelf model (Table B1) and for the unsupervised SimCSE approach (Table B2). Both tables show relatively low similarity along the diagonal when compared to Table 3. The diagonal similarity of the virtue elements is higher than the vice elements for both tables, suggesting that a limited level of knowledge is already present in the off-the-shelf SimCSE. Moreover, the poor result of the unsupervised SimCSE approach aligns with the findings in the main paper, indicating that labels are necessary to grasp a pluralist approach to morality.

B.1.3 Alignment and Uniformity

Alignment and *unifomity* are metrics commonly used to assess the quality of an embedding space, measuring *alignment* between positive pairs and *uniformity* of the embedding space (Gao et al., 2021). They can be calculated as follows:

$$\mathcal{L}_{\text{align}}(f; \alpha) \triangleq \mathbb{E}_{(x,y) \sim p_{\text{pos}}} \left[\|f(x) - f(y)\|_2^\alpha \right]$$

$$\mathcal{L}_{\text{uniform}}(f; t) \triangleq \log \mathbb{E}_{x,y} \overset{\text{i.i.d}}{\sim} p_{\text{data}} \left[e^{-t\|f(x)-f(y)\|_2^2} \right]$$

	Care	Fairness	Loyalty	Authority	Purity	Harm	Cheating	Betrayal	Subversion	Degradation	Non-moral
Care	27.8	19.3	21.8	17.6	20.5	14.6	10.4	11.6	11.9	8.6	10.3
Fairness	19.3	29.7	23.7	20.5	18.2	16.9	17.5	17.2	17.1	12.6	11.6
Loyalty	21.8	23.7	28.5	18.4	17.5	14.7	13.8	16.8	16.0	9.9	11.4
Authority	17.6	20.5	18.4	22.5	16.4	13.0	13.7	14.6	15.7	10.4	10.2
Purity	20.5	18.2	17.5	16.4	25.5	10.9	9.8	10.2	10.3	9.8	8.7
Harm	14.6	16.9	14.7	13.0	10.9	22.0	18.9	19.5	18.5	18.3	12.2
Cheating	10.4	17.5	13.8	13.7	9.8	18.9	22.4	20.5	19.6	19.5	11.9
Betrayal	11.6	17.2	16.8	14.6	10.2	19.5	20.5	23.0	20.9	18.4	12.3
Subversion	11.9	17.1	16.0	15.7	10.3	18.5	19.6	20.9	22.0	17.7	12.0
Degradation	8.6	12.6	9.9	10.4	9.8	18.3	19.5	18.4	17.7	23.7	11.9
Non-Moral	10.3	11.6	11.4	10.2	8.7	12.2	11.9	12.3	12.0	11.9	9.8

Table B1: Moral similarity on MFTC train set using the off-the-shelf SimCSE model.

	Care	Fairness	Loyalty	Authority	Purity	Harm	Cheating	Betrayal	Subversion	Degradation	Non-moral
Care	26.1	19.4	21.7	21.5	23.2	19.7	18.3	18.9	19.3	19.0	19.6
Fairness	19.4	25.1	20.8	21.8	20.2	18.7	20.4	19.6	20.3	18.6	19.0
Loyalty	21.7	20.8	25.5	21.8	21.1	18.8	18.8	20.9	21.0	18.7	20.0
Authority	21.5	21.8	21.8	26.6	22.3	19.6	20.8	21.7	23.1	19.9	20.7
Purity	23.2	20.2	21.1	22.3	27.5	18.4	18.7	19.1	19.7	20.1	19.4
Harm	19.7	18.7	18.8	19.6	18.4	22.3	20.4	20.8	21.0	20.3	19.3
Cheating	18.3	20.4	18.8	20.8	18.7	20.4	23.1	21.4	21.9	20.8	19.7
Betrayal	18.9	19.6	20.9	21.7	19.1	20.8	21.4	23.1	22.9	20.6	20.0
Subversion	19.3	20.3	21.0	23.1	19.7	21.0	21.9	22.9	24.4	21.0	20.5
Degradation	19.0	18.6	18.7	19.9	20.1	20.3	20.8	20.6	21.0	22.8	19.6
Non-Moral	19.6	19.0	20.0	20.7	19.4	19.3	19.7	20.0	20.5	19.6	20.4

Table B2: Moral similarity on the MFTC train set using the unsupervised SimCSE approach.

Our goal is to generate the best possible embedding space mapping for this corpus—however, we only train on a relatively small and limited corpus, and thus we do not strive for a state-of-the-art *alignment* and *uniformity*. Nevertheless, for completeness, we report the *alignment* and *uniformity* using the test dataset. Table B3 displays the result of *alignment* and *uniformity* metrics. The supervised SimCSE outperforms in *alignment*, but gets a worse score in *uniformity* when compared to the other two approaches. This is consistent with the findings in the SimCSE paper (Gao et al., 2021) where the supervised SimCSE amends the *alignment* and the unsupervised SimCSE effectively improves *uniformity*.

Approach	Alignment	Uniformity
Off-the-shelf SimCSE	1.49	-3.13
Unsupervised SimCSE	1.50	-3.12
Supervised SimCSE	0.77	-2.27

Table B3: *Alignment* and *uniformity* on MFTC test dataset. For both, lower numbers are better.

B.2 Extrinsic Evaluation

We provide additional details on generalizability and comparison to MFD2.0 evaluation results, and offer further insight through a classification task.

B.2.1 Generalizability on Test Set

Figures B3 and B4 show the UMAP plot of the MFTC test set mapped on the moral embedding space with the off-the-shelf model and with the supervised SimCSE approach, respectively. The figures are similar to Figure 3, however grouping the 10 moral elements as vices or virtues. Figure B3 does not show clearly distinguishable cluster. Instead, Figure B4 shows a clearer separation between vice and virtue values—vice and virtue clusters are less mixed together, and a bigger gap can be found between them.
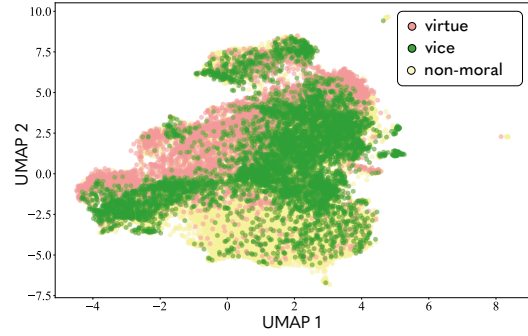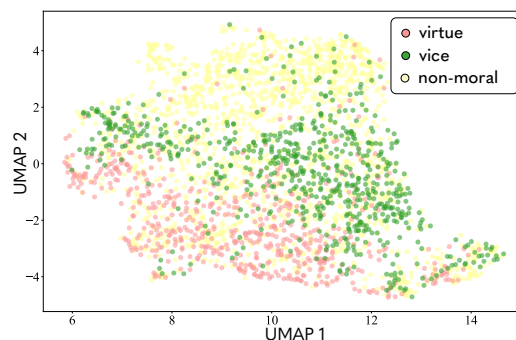

Figure B3: UMAP plot of MFTC test set with the off-the-shelf SimCSE model (only vices and virtues).

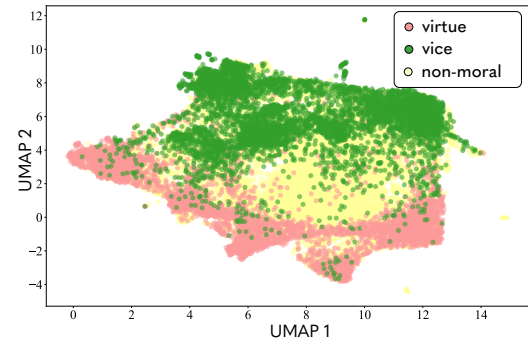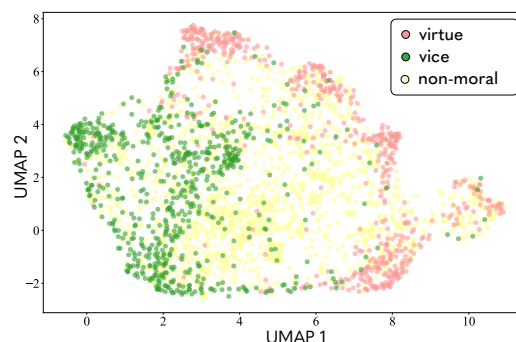Figure B4: UMAP plot of MFTC test set with the supervies SimCSE approach (only vices and virtues).

Table B4 and Table B5 show the moral similarity obtained with off-the-shelf SimCSE model and unsupervised SimCSE approach (similar to Table 4). These tables confirm the visual intuition found in Figure 3, with a low similarity along the diagonal. Further, these tables are consistent with the corresponding training set tables from the intrinsic evaluation (Tables B1 and B2)).

	Care	Fairness	Loyalty	Authority	Purity	Harm	Cheating	Betrayal	Subversion	Degradation	Non-moral
Care	27.8	19.7	22.4	17.5	21.8	13.9	10.6	10.8	10.8	7.8	10.4
Fairness	19.7	30.6	24.3	20.3	20.3	17.4	18.2	18.1	16.9	12.2	11.7
Loyalty	22.4	24.3	29.4	18.2	18.8	15.4	14.1	17.6	15.3	9.5	11.6
Authority	17.5	20.3	18.2	22.2	17.9	12.9	13.4	13.3	14.8	10.2	10.1
Purity	21.8	20.3	18.8	17.9	28.5	10.6	10.1	10.0	9.9	8.3	9.0
Harm	13.9	17.4	15.4	12.9	10.6	21.5	18.4	20.1	17.9	17.0	11.8
Cheating	10.6	18.2	14.1	13.4	10.1	18.4	22.8	21.5	18.8	18.5	11.5
Betrayal	10.8	18.1	17.6	13.3	10.0	20.1	21.5	26.3	21.1	19.6	12.6
Subversion	10.8	16.9	15.3	14.8	9.9	17.9	18.8	21.1	21.8	17.6	11.3
Degradation	7.8	12.2	9.5	10.2	8.3	17.0	18.5	19.6	17.6	23.8	11.8
Non-Moral	10.4	11.7	11.6	10.1	9.0	11.8	11.5	12.6	11.3	11.8	9.6

Table B4: Moral similarity on MFTC test set using the off-the-shelf SimCSE model.

	Care	Fairness	Loyalty	Authority	Purity	Harm	Cheating	Betrayal	Subversion	Degradation	Non-moral
Care	27.1	19.5	22.1	21.6	23.6	19.4	18.3	18.3	19.1	18.6	19.7
Fairness	19.5	25.2	21.0	22.0	21.2	18.8	20.3	19.2	20.4	19.1	19.0
Loyalty	22.1	21.0	26.0	21.8	21.3	19.2	18.6	20.6	21.2	19.6	20.1
Authority	21.6	22.0	21.8	27.0	22.8	19.7	20.7	21.0	23.0	20.7	20.8
Purity	23.6	21.2	21.3	22.8	29.2	18.4	19.2	19.5	20.1	19.9	19.6
Harm	19.4	18.8	19.2	19.7	18.4	22.5	20.4	20.7	21.3	20.3	19.5
Cheating	18.3	20.3	18.6	20.7	19.2	20.4	23.7	21.3	21.6	20.9	19.6
Betrayal	18.3	19.2	20.6	21.0	19.5	20.7	21.3	24.2	22.8	21.4	19.9
Subversion	19.1	20.4	21.2	23.0	20.1	21.3	21.6	22.8	24.9	21.9	20.7
Degradation	18.6	19.1	19.6	20.7	19.9	20.3	20.9	21.4	21.9	23.6	20.2
Non-Moral	19.7	19.0	20.1	20.8	19.6	19.5	19.6	19.9	20.7	20.2	20.6

Table B5: Moral similarity on MFTC test set using the unsupervised SimCSE approach.

B.2.2 Comparison to MFD2.0

Clustering In Figure B5 we report the purity score for K ranging from 2 to 15 (similar to the Silhouette coefficient in Section 5.2.2).
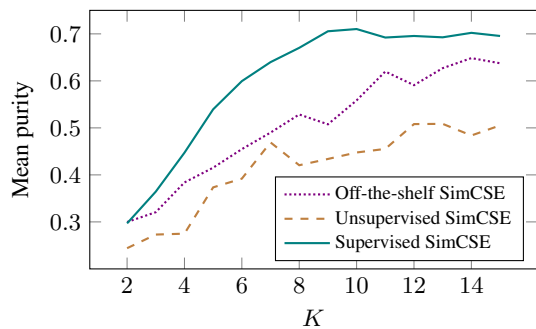

Figure B5: Mean purity for K ranging from 2 to 15 for the three compared embedding spaces.

We observe an overall increase in the mean purity score for all approaches as K increases, which is to be expected due to the calculation of the purity score (Section 4.2.2). We notice that the supervised SimCSE results in higher mean purity compared to other approaches, reaching its peak at $K = 9$ and

$K = 10$. These values are similar to the number of moral values, indicating that corresponding embedding spaces are consistent with the MFT taxonomy and the MFD2.0 lexicon. Further, we observe that the supervised SimCSE approach and the off-the-shelf SimCSE model lead to a higher mean purity compared to the unsupervised SimCSE approach.

Moral Similarity In Table 6 we report the moral similarity for MFD2.0 with the supervised SimCSE approach, whereas in Tables B6 and B7 we report the analogous results with the off-the-shelf model and the unsupervised SimCSE approach. We notice how the unsupervised approach only slightly captures the similarity among words belonging to the same MFT element, in strong contrast with the supervised approach. We observe the same pattern with off-the-shelf SimCSE approach in Table B6. The strong similarity of Tables B6 and B7 corresponds with the clustering findings described in Figure 4 and Figure B5, with the off-the-shelf SimCSE model leading to slightly better results to the unsupervised SimCSE approach.

	Care	Fairness	Loyalty	Authority	Purity	Harm	Cheating	Betrayal	Subversion	Degradation
Care	32.0	18.0	20.0	17.1	19.2	16.6	13.3	13.3	12.3	13.6
Fairness	18.0	28.0	17.4	17.8	16.0	13.1	16.1	16.1	16.2	11.4
Loyalty	20.0	17.4	30.0	20.2	18.2	15.0	16.2	20.3	19.9	14.3
Authority	17.1	17.8	20.2	25.4	18.2	15.0	14.5	17.4	19.0	13.2
Purity	19.2	16.0	18.2	18.2	26.2	12.7	11.0	14.0	14.8	14.5
Harm	16.6	13.1	15.0	15.0	12.7	35.6	23.7	26.5	25.5	27.8
Cheating	13.3	16.1	16.2	14.5	11.0	23.7	31.4	31.4	25.7	24.1
Betrayal	13.3	16.1	20.3	17.4	14.0	26.5	31.4	42.6	32.8	25.9
Subversion	12.3	16.2	19.9	19.0	14.8	25.5	25.7	32.8	36.5	24.6
Degradation	13.6	11.4	14.3	13.2	14.5	27.8	24.1	25.9	24.6	33.7

Table B6: Moral similarity for MFD2.0 with the off-the-shelf SimCSE approach.

	Care	Fairness	Loyalty	Authority	Purity	Harm	Cheating	Betrayal	Subversion	Degradation
Care	36.4	26.8	30.0	28.2	29.4	30.0	26.7	28.3	26.5	28.1
Fairness	26.8	32.1	27.8	27.7	27.0	26.5	28.2	28.2	27.9	26.8
Loyalty	30.0	27.8	38.3	31.3	29.9	28.7	29.6	34.1	32.0	29.0
Authority	28.2	27.7	31.3	33.9	29.7	28.2	28.6	30.4	31.4	28.0
Purity	29.4	27.0	29.9	29.7	33.5	28.0	27.2	29.2	28.8	29.2
Harm	30.0	26.5	28.7	28.2	28.0	34.7	28.7	30.8	30.4	30.3
Cheating	26.7	28.2	29.6	28.6	27.2	28.7	33.5	33.7	32.0	29.8
Betrayal	28.3	28.2	34.1	30.4	29.2	30.8	33.7	41.7	36.2	31.3
Subversion	26.5	27.9	32.0	31.4	28.8	30.4	32.0	36.2	38.7	31.0
Degradation	28.1	26.8	29.0	28.0	29.2	30.3	29.8	31.3	31.0	33.0

Table B7: Moral similarity for MFD2.0 with the unsupervised SimCSE approach.

B.2.3 Classification

As suggested in the literature (Eger et al., 2019), we test the resulting embedding spaces by adding a linear layer (i.e., a fully connected layer) with 11 output features as a classification head on top of the trained moral embedding spaces, to predict the 11 labels described in Table 2. We compare the off-the-shelf SimCSE model and the embeddings trained with unsupervised and supervised approaches to judge the effectiveness of the (un)supervised training of the moral embeddings for the classification task. The three compared embedding spaces are not retrained—we only train the linear layer on the test set with 5-fold cross-validation and report mean and standard deviation. The hyperparameters used for the linear classifier are reported in Table B8. Default and commonly used values were chosen.

Hyperparameters	Options
Max Sequence Length	64
Epochs	10
Batch Size	16
Learning Rate	0.01
Dropout	0.1
Loss function	Binary Cross Entropy

Table B8: Hyperparameters used for the linear classifier.

Results We report the mean and standard deviation of the micro and macro F_1-scores in Table B9.

Approach	Micro F_1	Macro F_1
Supervised SimCSE	**68.4 ± 3.1**	**56.7 ± 2.6**
Unsupervised SimCSE	58.0 ± 2.9	36.2 ± 3.4
Off-the-shelf SimCSE	59.4 ± 3.1	39.4 ± 3.9

Table B9: Classification results for the three compared approaches.

First, we notice that the supervised SimCSE approach clearly outperforms the off-the-shelf model and the unsupervised approach, confirming that label information is crucial to recognize a pluralist approach to morality. Further, the reported F_1-scores are in line with previous experiments on the same dataset (Liscio et al., 2022a), which we reproduce in the next section. Second, the unsupervised approach does not improve over the off-the-shelf model despite having been exposed to the training set, showing that the necessity of labels overshadows the need for large amounts of training data for the task of pluralist moral classification.

BERT Baseline We also add two baselines by performing multi-label classification with BERT (Devlin et al., 2019), which is considered state-of-the-art in the classification of the MFT taxonomy (Alshomary et al., 2022; Liscio et al., 2022a; Huang et al., 2022; Bulla et al., 2023). In the first variant (referred to as 'BERT'), we first train BERT on the MFTC training set and then we continue to train it on the test set with a 5-fold cross-validation. In the second variant (referred to as 'BERT (base)'), we only train BERT on the test set with a 5-fold cross-validation. We base the hyperparameters on the ones used by Liscio et al. (2022a), who performed experiments with the same corpus and model. We set the number of epochs to 10, similar to the linear classifier used in the previous experiments. The hyperparameters are listed in Table B10 and the results are shown in Table B11.

Hyperparameters	Options
Model name	bert-large-uncased
Max Sequence Length	64
Epochs	10
Batch Size	16
Optimizer	AdamW
Learning Rate	**2e-5**, 5e-5
Loss function	Binary Cross Entropy

Table B10: Hyperparameters for the BERT baseline. In bold, the chosen hyperparameters.

Approach	Micro F_1	Macro F_1
BERT	71.0 ± 1.5	62.2 ± 1.1
BERT (base)	66.2 ± 2.4	55.8 ± 1.2

Table B11: Classification results for the BERT baseline.

The end-to-end training of BERT offers an advantage with respect to the split training (sentence embeddings + linear classifier) of the SimCSE approaches. Further, we only choose a simple linear layer as classifier head on top of the SimCSE embeddings, yet being aware that a more complex classifier could lead to better performance. As a result, the results of the supervised SimCSE approach (Table B9) are comparable to the BERT baseline in micro F_1-score and worse in macro F_1-score, showing BERT's better capacity at handling imbalanced datasets. However, the goal of the SimCSE classification evaluation is not to improve the classification performance over the BERT baselines but rather to compare the effectiveness of the different training approaches.

Misclassification Error Analysis   To further analyze the results of the five classification approaches, we inspect (1) the confusion between moral and non-moral texts and (2) the confusion between and within foundations. In Table B12 we show the following four types of misclassification errors (which add up to 100%), as previously performed for a similar classification task (Liscio et al., 2022a).

Error I A tweet labeled with one or more moral values is classified as non-moral or no prediction.
Error II A tweet labeled as non-moral is classified with one or more moral values.
Error III A tweet labeled with a moral value is classified with values from other foundations.
Error IV A tweet labeled as a vice/virtue is classified as the opposite virtue/vice within that foundation.

Approach	I	II	III	IV
Supervised SimCSE	50.5	30.6	17.3	1.60
Unsupervised SimCSE	62.9	24.6	11.3	1.15
Off-the-shelf SimCSE	62.2	24.8	11.6	1.40
BERT	28.5	36.9	30.7	3.86
BERT (base)	29.3	38.0	29.8	2.89

Table B12: Misclassification errors (reported as percentages over the total number of errors).

The SimCSE approaches mostly incur in Error I and Error II (i.e., distinguishing between moral and non-moral texts). Instead, the BERT models show an approximately equal distribution of Error I, Error II, and Error III. This means that, compared to SimCSE, BERT is better at distinguishing moral vs. non-moral, but worse at predicting the correct foundation. This difference can be explained by the training procedure of BERT (which uses all labeled data points, which are mostly composed of non-moral labels) vs. supervised SimCSE (which focuses on distinguishing among the moral elements). Finally, BERT makes more mistakes between virtue and vice within a foundation (Error IV) compared to the SimCSE approaches.

Training Time   Table B13 displays the time needed for training the models. Off-the-shelf SimCSE and BERT (base) are not trained on the MFTC training set, thus the first values are 0. The supervised SimCSE takes significantly less total time for the training process than BERT and than the unsupervised SimCSE (which takes longer due to the larger number of triples used during training, as

described in Section 3 and A.2). Considering the small difference in the final F_1-scores (Tables B9 and B11), there is a trade-off in using the supervised SimCSE approach. Further, the embedding space can be re-used in different applications (e.g., language classification and generation).

Approach	Training Time (s)
Supervised SimCSE	249 + 10
Unsupervised SimCSE	493 + 11
Off-the-shelf SimCSE	0 + 10
BERT	3521 + 327
BERT (base)	0 + 313

Table B13: Training time comparison. The first value shows the training time on the MFTC training set and the second value is the cross-validation on the test set.

Per-label Classification Results   Table B14 and B15 show the mean and standard deviation of F_1-scores for each label. Overall, a common pattern can be observed. *Cheating* and *harm* are the easiest vice values to classify, while *fairness* and *care* are the easiest virtues value to classify. On the other hand, the *purity* element is always difficult to identify for all approaches, likely due to the presence of fewer examples with this label in the dataset.

	Sup. SimCSE	Unsup. SimCSE
Care	67.9 ± 5.2	56.7 ± 3.7
Harm	57.5 ± 4.8	48.1 ± 6.7
Fairness	71.4 ± 6.3	50.3 ±8.8
Cheating	66.0 ± 3.6	40.1 ± 7.7
Loyalty	61.1 ± 6.0	36.7 ± 15.0
Betrayal	51.0 ± 9.4	16.8 ± 3.3
Authority	54.9 ± 10.4	30.2 ± 14.1
Subversion	37.1 ± 13.1	16.3 ± 3.9
Purity	46.3 ± 21.8	14.3 ± 10.1
Degradation	32.2 ± 12.4	14.6 ± 13.6
Non-moral	78.0 ± 3.7	73.9 ± 3.1

Table B14: Per-label classification mean and standard deviation for the compared SimCSE approaches.

Foundations-only Results   We additionally experimented with 6 labels, i.e., the 5 foundations (combining vices and virtues) plus the *non-moral* label. The supervised approach dataset construction slightly differs as vice and virtue from the same foundation are in this case assigned the same label. Thus, the positive instance is chosen as a data point annotated with the same foundation, and the negative instance as a data point annotated with a different foundation.

	BERT	BERT (base)
Care	70.5 ± 4.1	67.0 ± 3.3
Harm	64.7 ± 4.5	57.9 ± 4.3
Fairness	70.8 ± 7.8	68.7 ± 6.1
Cheating	71.2 ± 4.5	64.8 ± 4.9
Loyalty	65.4 ± 4.5	59.9 ± 5.2
Betrayal	55.5 ± 13.2	48.2 ± 9.7
Authority	59.6 ± 7.8	51.5 ± 12.9
Subversion	44.8 ± 10.2	39.1 ± 13.5
Purity	50.1 ± 8.1	41.7 ± 10.7
Degradation	52.5 ± 14.0	38.4 ± 14.5
Non-moral	80.3 ± 2.3	77.2 ± 3.5

Table B15: Per-label classification mean and standard deviation for the BERT models.

We show the results with 6 and 11 labels (as in Table B9) in Table B16. The used hyperparameters are in Tables B17 and B18. We observe that the results are comparable. Since distinguishing between vice and virtue allows for a more fine-grained interpretation of morality with respect to only distinguishing among foundations, we opted for the 11-label approach.

Approach	Micro F_1	Macro F_1
Supervised SimCSE (6 labels)	68.0	56.7
Unsupervised SimCSE (6 labels)	57.5	39.4
Supervised SimCSE (11 labels)	68.4	56.7
Unsupervised SimCSE (11 labels)	58.0	36.2

Table B16: Classification result with 6 and 11 labels.

Hyperparameters	Options
Model name	sup-simcse-bert-large-uncased
Max Sequence Length	64
Epochs	3
Batch Size	16
Learning Rate	5×10^{-5}
Temperature	0.05
Pooler	cls

Table B17: Hyperparameters chosen for the 6-label supervised SimCSE approach.

Hyperparameters	Options
Model name	unsup-simcse-bert-large-uncased
Max Sequence Length	64
Epochs	1
Batch Size	16
Learning Rate	3×10^{-5}
Temperature	0.05
Pooler	cls

Table B18: Hyperparameters chosen for the 6-label unsupervised SimCSE approach.