# Learning Moral Diversity: Modelling Individual Perspectives in Moral Classification of Texts

**Yi Ren, Lewis Mitchell, Matthew Roughan**

School of Computer and Mathematical Sciences, University of Adelaide
{yi.ren, lewis.mitchell, matthew.roughan}@adelaide.edu.au

## Abstract

Understanding moral values expressed in social media texts provides us valuable insights into how people form moral judgements and motivate a range of social science studies. Many NLP techniques have been applied and supervised models trained on crowdsourced datasets have shown great classification performance. However, most approaches simplify the problem by aggregating multiple annotators' labels into a single "ground truth", overlooking the inherent subjectivity of the task that caused disagreements in annotations. In this work, we introduce the Annotator Layer, an extension to finetuning pretrained large language model BERT that learns annotator-specific features, allowing training classifiers directly on raw annotations. Our model improves predictions of individual-level annotations and yields interpretable representations that reveal meaningful differences in annotators' moral perspectives. We further show that models trained on aggregated labels may hide this variation in moral judgement and give a misleading impression of performance. Finally, we discuss that the greater performance gains on certain moral foundations can indicate various levels of ambiguity across foundations. Overall, we demonstrate that disagreement reflects the inherent subjectivity of the task and that modelling individual perspectives grants benefits for moral classification of texts.

## Introduction

Morality plays a vital role in shaping people's opinions and forming judgement towards social events. Accordingly, analysing morality allows for better understanding of what people believe and how people interact and form communities, inspiring diverse research directions — from analysing political ideology and polarisation (Haidt and Graham 2007) to understanding how people engage in public-health discourse (Zhou et al. 2024; Jiang, Luceri, and Ferrara 2025). Thus, it is extremely valuable to be able to extract moral values from human-created content. In particular, people express opinions and stances via language, revealing moral beliefs that they hold. Consequently, studies have focused on the task of moral value classification in texts.

Many Natural Language Processing (NLP) techniques and approaches have been applied and integrated with an empirically validated psychological framework called the **Moral Foundations Theory** (MFT) (Graham, Haidt, and Nosek 2009; Haidt 2012; Graham et al. 2013). The MFT decomposes human moral beliefs into 5 distinct moral foundations each with its own virtue and vice dimensions: *Authority/Subversion*, *Care/Harm*, *Fairness/Cheating*, *Loyalty/Betrayal* and *Purity/Degradation*. Approaches including lexicons (Graham and Haidt 2012; Frimer et al. 2019; Araque, Gatti, and Kalimeri 2020; Hopp et al. 2021) and supervised machine learning models (Lin et al. 2018; Beiró et al. 2023; Huang, Wormley, and Cohen 2022) have been employed to classify text.

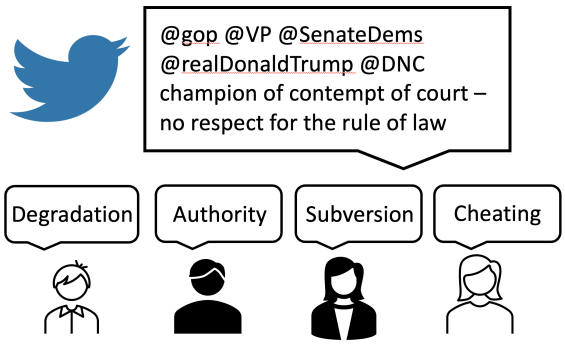


Figure 1: A tweet example with disagreeing labels given by four annotators in the Moral Foundations Twitter Corpus (Hoover et al. 2020), demonstrating how individuals interpret moral expressions differently.

Recently, transformer-based large language models have shown promising results in classification tasks. BERT in particular have been widely applied due to its ability to generate rich contextual and semantic embeddings (Devlin et al. 2018). Fine-tuning BERT with crowdsourced data has become a standard and effective approach for achieving state-of-the-art performance in moral value classifications (Preniqi et al. 2024; Nguyen et al. 2024; Guo, Mokhberian, and Lerman 2023).

Human moral judgement is inherently subjective — different people hold different beliefs, interpret social contents differently and prioritise different foundations. Additionally, classifying texts from social discourse further extends the subjectivity due to the ambiguity of text data. However, many approaches train a universal classifier that predicts a moral label, treating the task as if there exists a "ground

truth". In reality, such a model may fail to represent moral judgements of a diverse group of people. Moreover, crowdsourced datasets in this field often exhibit substantial disagreement among annotators (Figure 1), yet existing work typically disregards this information, treating disagreement as noise. A more sophisticated approach is to build models that learn how individuals or groups of people who share similar moral beliefs differ in their judgement from others.

In this work, we take a step towards this goal by training classifiers that model annotators in crowdsourced data specifically. We do this by adding a neural network layer on top of finetuning BERT. This additional layer learns how a particular annotator interprets moral content differently from the shared text embeddings. Furthermore, the added layer can be seen as interpretable representations that capture meaningful differences between individuals, revealing biases and tendencies towards different moral values. Our results show substantial improvement in predicting individual annotations and highlight concerns that training classifiers on aggregated labels may appear highly accurate but mask inconsistencies across annotators.

In summary, this work makes three key contributions: (1) we present one of the first modelling approaches that captures individual perspectives in moral value classification, accounting for task subjectivity; (2) we demonstrate that annotator biases and tendencies in crowdsourced datasets are learnable features rather than noise; and (3) we raise concerns about aggregating annotations into a single ground truth label in such task, urging future modelling approaches to incorporate individual-level variation.

## Related Work

### MFT NLP

**Dictionary-based Approaches** Traditionally, moral value classifications of texts are performed using dictionary approaches. These approaches utilise human-crafted lexicons, where certain words are associated with a set of moral values. When classifying a document, words contained in the document will be compared with tokens in the lexicons and a score for each foundation is accumulated based on either occurrence of matching words or embedding similarities of words (Araque, Gatti, and Kalimeri 2020; Hopp et al. 2021). However, despite their widespread usage there are limitations to dictionary approaches. Nguyen et al. analyses three existing moral foundations dictionaries (MFDs): MFD (Graham and Haidt 2012), MFD 2.0 (Frimer et al. 2019) and the extended MFD (Hopp et al. 2021), and reports issues of reliability of the MFDs. The key limitations include a lack of consensus across three MFDs; unclear instructions on word variation handling; correlation between document length and score; and personal and social biases in lexicon creation processes (Nguyen et al. 2024). Moreover, Kraft and Klemmensen argue that dictionary approaches may be insufficient to measure moral rhetoric due to the inability to capture contextual and semantic meanings (Kraft and Klemmensen 2024).

**Supervised Learning Approaches** Many studies have also focused on supervised learning approaches where classifiers are trained to map texts to moral values. Classical machine learning models such as logistic regression and support vector machines (SVMs), as well as deep learning models such as long short-term memory networks (LSTMs), have been widely adopted to perform the classification task (Lin et al. 2018; Araque, Gatti, and Kalimeri 2020; Hoover et al. 2020; Trager et al. 2022; Beiró et al. 2022). More recent work uses transformer-based pretrained language models (PLMs), particularly Bidirectional Encoder Representations from Transformers (BERT). Trager et al. report baseline performance from finetuning BERT models on a label dataset of Reddit posts (Trager et al. 2022). Ngyuen et al. and Preqini et al. provide in-depth analyses of BERT finetuning procedures and evaluations of performance in practice (Nguyen et al. 2024; Preniqi et al. 2024). Supervised learning approaches have shown better results as they are capable of capturing varying moral expressions of words under different context in comparison to dictionary approaches. Despite strong performance, construction and training of supervised models typically rely on large-scale crowdsourced datasets that often exhibit annotation disagreement as a result of the subjective nature of the task. Existing work often handle this disagreement through label aggregation, implicitly assuming the existence of a ground-truth label. Yet studies in moral psychology demonstrate that moral judgement is inherently subjective and varies across individuals, indicating that aggregation may hide meaningful differences in how people interpret the same content.

### Subjectivity in Moral Judgement

Human moral judgement is widely recognised as subjective and shaped by individual differences. Haidt's social intuitionist model highlights how moral judgements arise from intuitive, socially and culturally shaped processes (Haidt 2001). This model later informed MFT, which underpins most NLP research on moral value classification and posits that moral judgement varies across individuals (Haidt 2012). Cultural differences were among the most influential factors shaping variability in moral judgement. Early work shows that people from different countries make different moral evaluations, even when presented with the same scenarios (Haidt, Koller, and Dias 1993). Subsequent studies using the Moral Foundations Questionnaires (Graham et al. 2011; Atari et al. 2023) further validated substantial demographic and cultural differences in moral intuitions. Furthermore, studies also show that interpretation of morality shifts depends on social identities (Koleva et al. 2012; Ellemers and Van der Toorn 2015). One must then consider individual perspectives when forming moral judgement.

NLP work has focused on capturing the variations in the expression of moral values in language, but lacked attentions on the subjectivity in how individuals interpret morality in texts. During the construction of crowdsourced datasets, researchers acknowledged the presence of annotator disagreement due to inherent subjectivity and ambiguity introduced by limited contextual information (Hoover et al. 2020). Possible resolution such as discussion or further training can be applied to improve consensus but this fabricated agreement may hide important insights into the subjectivity of the task

(Hoover et al. 2017). The dataset authors explicitly encouraged researchers to investigate the impact of annotators on labelling (Trager et al. 2022).

Following previous work, this research aims to enrich the limited studies on how to address individual perspectives in MFT NLP and reveal insights into the impact of annotators and the subjective nature of moral foundations.

## Data

In this work we use the **Moral Foundations Twitter Corpus** (MFTC) to run our experiments. The MFTC contains 35,108 tweets collected from Twitter (now X), across seven socially and politically relevant topics identified by hashtags. Twenty three annotators were trained to manually assign labels when they recognise moral foundations and their polarity expressed in the tweets. An additional "non-moral" label is included for annotators to flag tweets with no presence of any moral foundations. Every tweet receives between 3 to 8 annotation, in which the majority are annotated by 3 or 4 annotators. As shown in Table 1, on average, each annotator has labelled only about one-seventh of the entire tweet collection, reflecting the highly sparse annotation structure. Large standard deviation highlights the variability of numbers of tweets each annotator labelled, indicating that a small number of annotators covered a majority of the dataset.

| # Annotator | Mean | Median | Min | Max | s.d |
|---|---|---|---|---|---|
| 23 | 5585 | 4588 | 560 | 19556 | 4747 |

Table 1: Summary statistics of the number of tweets each annotator has annotated, showing the sparsity of annotation structure.

Table 2 shows the label distribution for all five foundations where each foundation is considered for the entire dataset. There exists substantial class imbalance between the moral classes and the absent class.

| Foundation | Virtue | Vice | Absent | Proportion |
|---|---|---|---|---|
| Authority | 8618 | 9855 | 109945 | 14.39 |
| Care | 10289 | 15852 | 102277 | 20.36 |
| Fairness | 10248 | 14387 | 103783 | 19.18 |
| Loyalty | 10053 | 7384 | 110981 | 13.58 |
| Purity | 4420 | 7562 | 116436 | 9.33 |

Table 2: Annotation distributions for all five foundations, with the proportion of annotations for moral classes (*virtue* and *vice*).

Figure 2 shows the co-annotation network between annotators where the edges have weights that represent the number of tweets two annotators both labelled. Edges with weight less than 500 are removed to identify clusters where annotators within one cluster have been presented similar twitter content, allowing us to examine and validate various annotator features that we can learn via our model later on.

Some prior work has disregarded the polarity of a foundation by merging virtue and vice labels or treated virtue and vice dimensions as two individual foundations. These approaches allow researchers to simplify the problem and avoid additional "noise". However, we preserve the virtue and vice labels, as we aim to study the subjectivity of moral judgement not only identifying whether a foundation is expressed, but also in how annotators differ in assigning opposite moral valences.

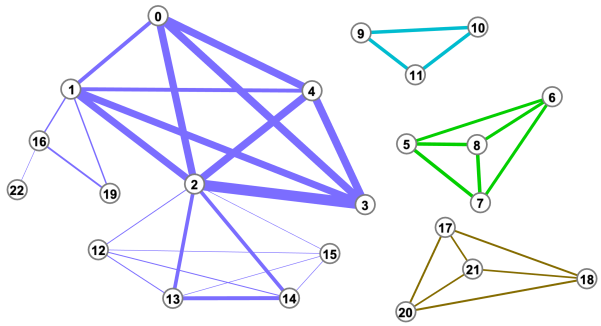


Figure 2: Co-annotation network between annotators. Edges represent the number of tweets co-annotated, and edges with weight less than 500 are filtered out.

## Methods

### Problem Setup

Our goal is to predict the moral foundation expressed in a piece of social media text while accounting for labelling biases and subjectivity among human annotators. Each text instance $x_i$ is annotated by annotator $a_j$ in a group of $N$ annotators, producing a set of labels for all present moral values. For each foundation $k \in \{1, ..., 5\}$, we extract the label $y_{ij}^{(k)}$ and consider the tuple $(x_i, a_j, y_{ij}^{(k)})$ as a single observation. This way we keep annotators' individual labels rather than aggregating annotations into a single "ground truth" label. We assume independence between foundations to simplify the multi-label classification problem into multiple single-label, multi-class classification tasks separately. We train a separate classifier $f_k(x_i, a_j; \theta_k)$ for each of the five moral foundations (*Authority, Care, Fairness, Loyalty, Purity*) to predict $y_{ij}^{(k)} \in \{1, 2, 3\}$, representing labels *virtue, vice* and *absent*.

### Model Overview

Our model extends a standard finetuned BERT classifier by incorporating neural network structures that learn annotator-specific features. This design is inspired by the *Crowd Layer* framework proposed by Rodrigues and Pereira, which applies neural network designs that directly learn from crowd-sourced labels from multiple annotators (Rodrigues and Pereira 2018). Adapting this idea to our setting, we introduce the **Annotator Layer** that adjusts the shared text features according to the learned representations of annotators, allowing the model to capture systematic difference in annotation patterns and annotators' individual perspectives. The model has 3 parts (Figure 3). Firstly, the pretrained language model

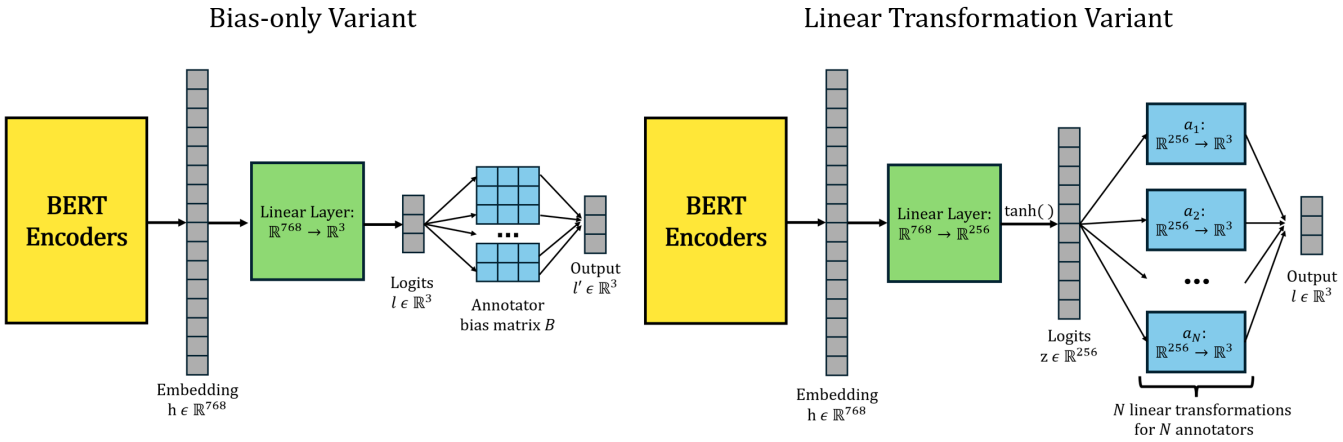


Figure 3: BERT finetuning with **Annotator Layer** that models annotators' individual labelling pattern and featrues. Left: Bias-only variant; Right: Linear transformation variant.

BERT takes texts as inputs and outputs contextual embeddings. We use the pretrained encoder-only language model **BERT-base-uncased** to encode text $x_i$, producing an embedding $h_i \in \mathbb{R}^{768}$ from the final hidden layer [CLS] token. The BERT model and the embeddings are finetuned with the following layers during the training process. The second part is a linear transformation layer normally denoted as the fine-tuning layer. This linear transformation layer has an input dimension 768 and an output dimension that depends on the following Annotator Layer variants. Here we present two variants of the Annotator Layer:

- **Bias-only**: this variant provides more interpretability as the bias terms can be directly interpreted as the annotators' biases towards certain classes.
- **Linear transformation**: this variant models each annotator with a linear transformation to better represent the annotators and provide greater predicting power.

**Bias-only** The objective of this variant is to provide interpretable annotator features learned from data. The fine-tuning layer has an output dimension of 3 that corresponds to the final output class probabilities. It applies a linear transformation on the embedding $h_i$ to get the base logits $l_i$:

$$l_i = Wh_i + b$$

where $W$ is the $3 \times 768$ weight matrix and $b$ is the $c \times 1$ bias vector of the linear layer. For the Annotator Layer, we use an $N \times c$ matrix for $N$ annotators where each row of the matrix corresponds to the biases of an annotator towards each class. The base logits $l_i$ is adjusted according to the annotator. For the base logits $l_i$ with annotator id $a_j$, we compute the adjusted logits:

$$l'_{ij} = l_i + B_j^T$$

where $B \in \mathbb{R}^{N \times c}$ is the annotator bias matrix and $B_j^T$ denotes the transpose of the $j$-th row, representing annotator $a_j$'s bias across all classes.

**Linear Transformation** The objective of this variant is to provide greater predicting power as we use neural network components with much more parameters to represent the annotators. The fine-tuning layer has a tunable output dimension which we choose to use 256 as an intermediate value between 768 and 3. We apply a *tanh* activation function to the output:

$$z_i = \tanh(Wh_i + b)$$

where $W$ is the $256 \times 768$ weight matrix and $b$ is the $256 \times 1$ bias vector. For the Annotator Layer, each annotator $a_j$ is a linear transformation with a $3 \times 256$ weight matrix and a $3 \times 1$ bias vector. We compute the adjusted logits:

$$l'_{ij} = W_{a_j} z_i + b_{a_j}$$

Both variants produce a 3 dimensional vector $l'_{ij}$, we then apply a *softmax* function to yield the predicted probability distribution over classes:

$$p_{ij} = \text{softmax}(l'_{ij}), \quad \hat{y}_{ij}^{(k)} = \arg\max_c p_{ij}^{(c)}$$

and the predicted class $\hat{y}_{ij}^{(k)}$ for text $x_i$ and annotator $a_j$ is the class with the greatest probability.

## Training Process

We train the model by minimising the cross-entropy loss function

$$\mathcal{L}_{CE} = -\sum_{c=1}^{C} y_c \log p_c,$$

where $C$ is the total number of classes.

Moreover, two regularisation terms are used to improve generalisation and pose constraints:

1. **L2 Norm (Weight Decay):** standard L2 regularisation is applied to the model parameters to prevent overfitting.

$$\mathcal{R}_{\text{L2}} = \|\theta_k\|_2^2.$$

2. **Centred Bias Penalty:** For the bias-only variant of the Annotator Layer, we add a centred penalty on the bias

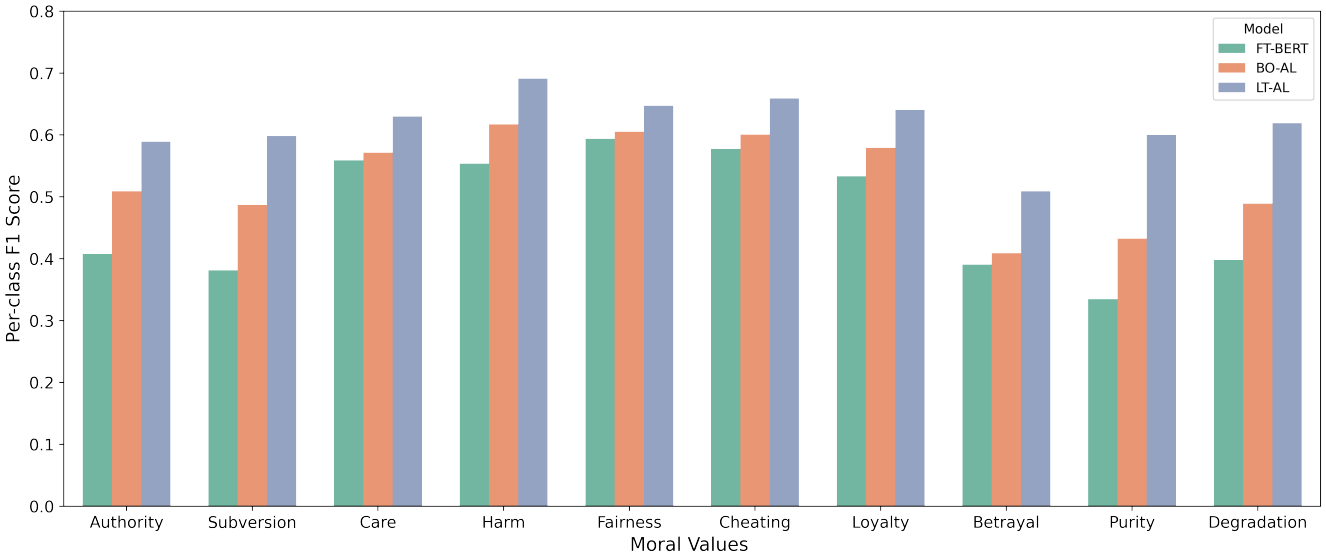


Figure 4: Average per-class F1 scores (five runs on different splits) for 10 moral values across five foundations (without class *absent*), comparing three models: FT-BERT, BO-AL and LT-AL, evaluated on raw annotations.

matrix $B$:

$$\mathcal{R}_{\text{centre}} = \left\| \frac{1}{N} \sum_{j=1}^{N} B_j \right\|_2^2 .$$

This is to ensure that the average annotator has zero bias towards each of the classes.

In summary, the model combines a shared representation of each text with annotator-specific adjustments. BERT encodes the input into a contextual embedding, which is then mapped to base logits via a linear classification layer. The Annotator Layer modifies these logits according to the learned parameters of the annotator who supplied the label, modelling their individual perspectives. Via backpropagation, updates to the BERT parameters incorporates annotator information, enabling the encoders to refine text representations according to annotators' various moral perspectives.

## Experiments

Through our experiments, we aim to demonstrate two main aspects of our approach: (1) its effectiveness in predicting individual annotations, and (2) the interpretability of the learned annotator features.

We begin by preprocessing the texts, removing URLs, non-alphanumeric characters, punctuations and retweet markers. All text is lowercased, and user mentions are replaced with the token "@user". Stopwords may optionally be removed, though we found this to have negligible effect on model performance. We then separate the data into subsets that correspond to each of the five foundations, where we keep the cleaned texts, annotator ids and annotations. For each foundation, we create five folds using a **StratifiedGroupKFold** from scikit-learn (Pedregosa et al. 2011), which maintains the overall label distribution across folds and prevents data leakage by ensuring that all annotations belonging to the same tweet are grouped together in either the train or test set. Models are built and evaluated using the 5 non-overlapping splits and any metrics reported is an average score calculated across the five splits.

We compare the two variants of Annotator Layer: Bias-only Annotator Layer (**BO-AL**) and Linear Transformation Annotator Layer (**LT-AL**); with a baseline of finetuned BERT without Annotator Layer (**FT-BERT**). For finetuning, we add a linear classifier that maps embeddings from BERT into a 3-class probability distribution and finetune parameters in the linear classifier and pretrained BERT. Finetuned BERT is currently regarded as the state-of-the-art approach in moral value classification. We deploy the same pretrained BERT-base-uncased and apply an identical training process wherever possible, allowing our experiments to also function as an ablation study.

We implement and train all models using **Pytorch** (Paszke et al. 2019) **v2.7** and optimise the parameters using the **AdamW** optimiser (Loshchilov and Hutter 2017). The initial learning rate is 2e-5 for the BERT parameters and is 1e-4 for the parameters in the linear layer and Annotator Layer, with linear decay and no warm-up. The lower learning rate is used to update the parameters of BERT moderately, avoiding deterioration of BERT's ability of capturing semantic and contextual meaning with the embeddings. In training, we use a batch size of 8, and the maximum input text length is set to be 64 tokens as all texts in the dataset are short in length. We set the L2 regularisation coefficient to 0.01 and the centred bias penalty to 0.05. We train the models for 5 epochs and freeze the BERT parameters during the first epoch to allow the newly added layers to stabilise before full finetuning. All experiments are run using one Nvidia A100-SXM4-40GB GPU.

We report classification performance using F1 score, which is calculated using the recall and precision for each of $c$ classes :

$$\text{F1}_c = \frac{2 \cdot \text{precision}_c \cdot \text{recall}_c}{\text{precision}_c + \text{recall}_c}.$$

And the macro-averaged F1 score is calculated by treating each class equally importantly:

$$\text{macro-averaged F1} = \frac{1}{C} \sum_{c=1}^{C} \text{F1}_c$$

Due to the imbalance label distributions between the moral classes (*virtue* and *vice*) and the *absent* class, F1 score better reflects the model performance than classification accuracy.

## Results

We first present a visualisation of the extracted features from the bias-only Annotator Layer to provide interpretable neural network weights that represent annotator-specific features. This is followed by quantitative results of the classification performance of our model to show its effectiveness in predicting individual annotations.

### Interpretable Annotator Features

The bias-only Annotator Layer provides interpretability and insights into annotators' tendencies of labelling a class for a foundation. The bias matrix weights are visualised using heatmaps to show annotators' biases towards or against a class. Figure 5 presents an example of the annotators' biases to all classes in one foundation. A detailed analysis of the learned features is presented in the discussion section.

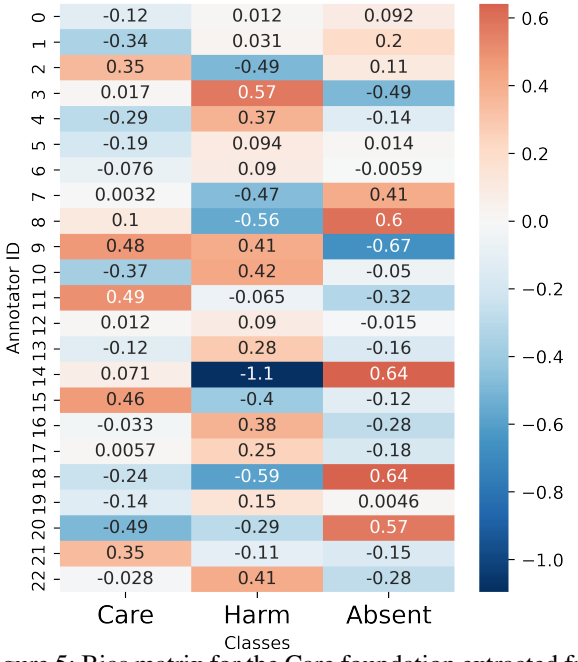


Figure 5: Bias matrix for the Care foundation extracted from the bias-only Annotator Layer.

### Classification Performance

Figure 4 shows that as we increase the complexity of neural network structures that model the annotators (from none for FT-BERT, to linear transformation for LT-AL), the classification performance improves across all moral foundations and their polarities. With more model parameters, the LT-AL model has greater representational power for modelling individual annotators, which leads to a largely improved classification performance over the baseline when predicting individual annotations. Even the BO-AL model yields a clear performance gain, despite adding only a small bias matrix (69 parameters). In particular, foundations *Authority*, *Loyalty* and *Purity* have gained significant improvement in F1 scores over the baseline. Table 3 further validate the results, showing improvements in all foundations when adding Annotator Layers.

| Foundation | Macro F1 (%) | | |
|---|---|---|---|
| | FT-BERT | BO-AL | LT-AL |
| Authority | 57.3 | 64.6 | 71.3 |
| Care | 67.4 | 70.2 | 75.1 |
| Fairness | 69.7 | 71.1 | 74.9 |
| Loyalty | 62.1 | 64.3 | 70.0 |
| Purity | 56.3 | 62.7 | 72.9 |
| Overall | 62.6 | 66.8 | 72.8 |

Table 3: Macro F1 scores for each foundation across three models: FT-BERT, BO-AL, and LT-AL.

Table 4 indicates that the greatest gains in classification performance occur in the moral classes (*virtue* and *vice*). For instance, the greatest improvement over the baseline model is the classification of *purity* (virtue aspect of foundation *Purity*) with the LT-AL model, yielding an increase of 0.265 in F1 score. All three models perform comparably when it comes to the effectiveness of classifying *absent* class (a tweet that is not expressing a certain moral value). However, the addition of Annotator Layer continues to provide improvement for the *absent* class, even when the classification performance of the baseline is sufficiently great.

## Discussion

### Interpretability of Annotator Layer

To illustrate the interpretability of the bias-only Annotator Layer, we can extract the bias matrix from the trained models and learn annotators' bias towards each class. Figure 5 gives us a visualisation from a per foundation level. Positive cell values indicate biases towards a class whereas negative cell values indicate biases against a class. Several bias patterns are observed in the bias matrix and we will use the *Care* foundation as an example. Annotator 2 shows a moderate bias toward the virtue aspect *Care* and, correspondingly, a bias against the vice aspect *Harm*. However, this complementary behaviour between the two polarities does not always hold. When an annotator possesses a tendency towards or against one moral class, the complementary class may instead be *absent*. We observe this pattern in several annotators (for example Annotator 3 and 7). In some cases, annotator

| Foundation | Per-class F1 (%) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Virtue | | | Vice | | | Absent | | |
| | FT-BERT | BO-AL | LT-AL | FT-BERT | BO-AL | LT-AL | FT-BERT | BO-AL | LT-AL |
| Authority | 40.8 | 50.9 | 58.9 | 38.1 | 48.7 | 59.8 | 93.0 | 94.1 | 95.0 |
| Care | 55.9 | 57.1 | 63.0 | 55.4 | 61.7 | 69.1 | 90.8 | 91.8 | 93.1 |
| Fairness | 59.4 | 60.6 | 64.7 | 57.7 | 60.1 | 65.9 | 92.1 | 92.7 | 93.9 |
| Loyalty | 53.3 | 57.9 | 64.1 | 39.0 | 40.9 | 50.9 | 93.8 | 94.2 | 94.9 |
| Purity | 33.5 | 43.2 | 60.0 | 39.8 | 48.9 | 61.9 | 95.5 | 96.0 | 96.8 |
| Overall | 48.5 | 53.9 | 62.1 | 46.0 | 52.1 | 61.5 | 93.0 | 93.8 | 94.7 |

Table 4: Per-class F1 scores (Virtue, Vice, Absent) for each foundation across three models: FT-BERT, BO-AL, and LT-AL.

exhibit biases towards or against both moral classes, with the *absent* class acting as the complement.

To better demonstrate, we show bias weights for a subset of annotators for the *Care* and *Fairness* Foundations. Annotators 5, 6, 7 and 8 are within one co-annotation cluster (Figure 2) and have been presented similar tweet contents when labelling, hence we pick this group to show how the labelling patterns share similarity and difference between foundations. Figure 6 shows the biases over two foundations for the group of selected annotators. We observe that Annotators 7 and 8 show a consistent tendency to favour the *absent* class, with only minor differences in their biases toward the *virtue* classes across the two foundations. In contrast, Annotators 5 and 6 both have different tendencies between the two foundations. Both of them have a similar bias pattern for foundation *Care*, where they favour the vice aspect moderately and label against the *virtue* class as a compliment. However, when considering foundation *Fairness*, Annotators 5 and 6 exhibit opposite bias patterns. Annotator 5 tends to label *virtue* less and favouring the *absent* label, whereas Annotator 6 is more likely to give a *virtue* label and less likely for the *absent* class. Although there exists consistent bias patterns across different foundations, one should examine this consistency carefully and consider these biases separately for all foundations rather than attempting to group annotators by a universal bias pattern.

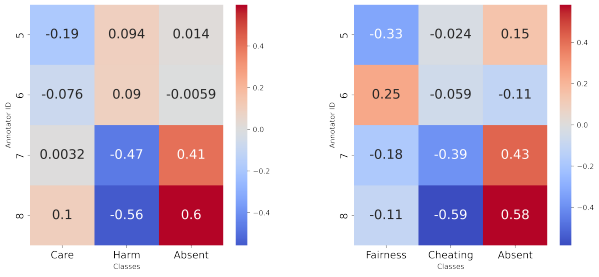


Figure 6: Bias Heatmaps of Foundations Care (Left) and Fairness (Right), for Annotators 5, 6, 7 and 8.

By summarising these observations, we may identify groups of annotators who share similar perspectives when interpreting moral values for each foundation. These groups exhibit consistent patterns in moral judgement, such as virtue/vice-oriented annotators, annotators who frequently give moral labels, and those who give moral labels more cautiously, leading to dominating non-moral (absent) annotations. Such patterns suggest meaningful "annotator types" that reveals insights into the diversity of moral judgement and provide potential categorisation for all individuals, not just annotators. This provides possible modelling approaches that learns group behaviours instead of modelling individual annotators, such as mixture-of-experts models where experts represents groups of people with similar perspectives. One can also study the correlations between individual political ideology, cultural background and other demographic factors with the "types" that we may identify, to gain insights into how this diversity of perspectives has developed.

## Raw Annotations and Aggregated Labels

As described in the **Experiments** section, the baseline model, FT-BERT, is trained using the same pretrained encoder and, where possible, the same training procedure as the models with Annotator Layers, making the evaluation effectively an ablation study. The improvements observed for both BO-AL and LT-AL demonstrate the effectiveness of learning annotator representations. These gains indicate that annotators provide consistent and learnable features in how they interpret morality expressed in text, and that including these features into the model yields more accurate annotation-level predictions.

Using the trained models with Annotator Layers, we apply a simple rule to obtain an aggregated label for each tweet. For each tweet, we activate all "annotators" in the Annotator Layer, regardless of their IDs, and obtain 23 predictions, yielding 23 probability distributions over the 3 classes. We then average these distributions and select the class with the highest mean probability as the aggregated prediction. For this analysis, we train the baseline model FT-BERT directly on aggregated labels, mimicking the common practice used when fine-tuning BERT for moral value classification.

We show the changes in macro F1 scores in Figure 7 as we move from standard finetuned BERT, to the two variants of the Annotator Layers. The horizontal axis can be interpreted as an increasing capacity to model annotator-specific features from left to right. We've already shown in the **Results** section that as we increase the capacity, the prediction accuracy of raw annotations also increases (solid lines in Figure 7). When evaluating on aggregated labels,

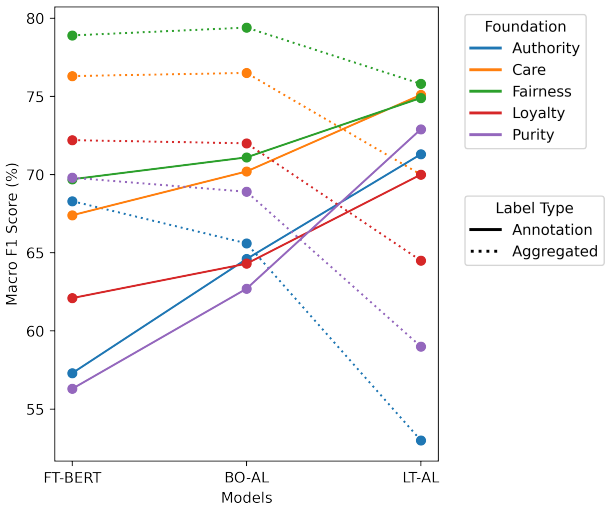


Figure 7: Trend of macro F1 scores as the modelling capacity to represent the annotators increases, compared for predicting raw annotations and aggregated labels.

the FT-BERT and BO-AL models achieve comparable performance, whereas the LT-AL model shows a substantial decrease. This decrease when evaluating on aggregated labels is more obvious where we see the largest gain when evaluating on raw annotations. Interestingly, the transition from BO-AL to LT-AL yields the strongest improvement on raw annotations, but it also produces substantial decrease under aggregated-label evaluation. This is expected, the LT-AL models are designed to capture annotator-specific features and better learn how individuals interpret texts and assign labels. When activating all annotators and collapsing their outputs into a "consensual" label, additional noise is introduced due to the sparse annotation structure. In essence, we are asking the learned annotator representations to make predictions on tweets that the corresponding annotators never see, with potentially great domain differences. These observations demonstrate that a model that's capable of representing annotators' individual perspectives does not necessarily agree with the aggregated labels. We want to highlight an important limitation of training models on aggregated labels. A strong performance of such model may hide substantial underlying variations in individual perspectives and may not reflect the true effectiveness of the model.

### Ambiguity of Foundations

While the addition of Annotator Layers improves overall performance, the gains vary across foundations, suggesting that some moral foundations exhibit greater ambiguity and therefore benefit more from annotator-specific modelling. Greater improvements occur in *Authority*, *Loyalty*, and *Purity*, compared to *Care* and *Fairness* (Table 5). This pattern mirrors the distinction between individualising foundations (*Care, Fairness*) and binding foundations (*Authority, Loyalty, Purity*). Individualising foundations are generally considered more morally relevant and are endorsed across the political spectrum, whereas binding foundations tend to receive endorsement from a smaller portion of the population (Graham, Haidt, and Nosek 2009). When annotator-specific features benefits the binding foundations more, it suggests that human judgement on these foundations are less consistent and more sensitive to individual interpretation, reflecting the ambiguity of foundations. Through our experiments, we observe that the binding foundations appear to be more ambiguous to human annotators than the individualising foundations.

| Foundation | Virtue | Vice | Average |
|---|---|---|---|
| Authority | 18.1 | 21.7 | **19.9** |
| Care | 7.1 | 13.7 | 10.4 |
| Fairness | 5.3 | 8.2 | 6.8 |
| Loyalty | 10.8 | 11.9 | **11.4** |
| Purity | 26.5 | 12.1 | **19.3** |

Table 5: Performance improvement in F1 scores (%) between LT-AL and FT-BERT for the moral classes for five foundations.

## Limitations and Future Work

Our work has two primary limitations.

First, although the dataset publication notes that annotator metadata (e.g., demographic information, political ideology and moral values measured by MFQ) exists, this information was not available to us and is therefore not incorporated into the analysis. As a result, while the Annotator Layer learns annotator-specific features and identified potential differences of annotation patterns between groups of annotators, we cannot directly examine how these patterns relate to known characteristics. Studies in moral psychology have validated that these characteristics have a direct impact to human moral judgement. Hence, access to such metadata can help validate the learned representations and provide explanations to some of the observed patterns.

Second, our approach does not provide a strong mechanism for producing high-quality aggregated predictions to moral values. We've demonstrated the bias-only variant's comparable classification performance to finetuned BERT on aggregated labels, but the linear transformation variant has shown a substantial decrease in performance. Many downstream applications ultimately requires a single, aggregated label for a text observation, yet our annotator-specific models requires annotator (human) information to provide accurate predictions which typically lacks in these tasks. The model learns fine-grained human-specific behaviours and does not generalise well for aggregated labels. Our naive approach to obtain consensual labels by activating all annotator corresponding neural network structures and averaging the prediction distributions introduces noise, especially given the sparse and uneven co-annotation structure. Developing principled aggregation methods that leverage annotator features is a vital future direction.

Moreover, as discussed before, the learned annotator representations such as the bias patterns, can be used to inform the construction of more expressive models. Ensem-

ble learning models such as mixture-of-experts may be used to approximate different groups of people who share similar moral perspectives and hence learn homogeneous patterns within groups. This points toward future work that continues modelling moral judgement at the level of individuals or groups, further exploring the subjectivity and ambiguity of the task and informing studies on human moral judgements.

## Conclusion

In this work we introduced the Annotator Layer for moral classification of texts that captures annotator-specific moral perspectives and annotation patterns, extending on fine-tuning BERT models. Our experiments demonstrate improved classification performance of individual annotations in crowdsourced dataset, along with interpretable representations of annotators' bias patterns. It is shown that disagreement between annotators in such subjective tasks is a learnable feature instead of annotation noise. The results reveal that modelling individual perspectives is particularly beneficial for ambiguous moral foundations and suggest that relying solely on aggregated labels can hide important information. We hope this work encourages future research to move beyond training a universal classifier that predicts a "ground truth" and develop models that better reflects diversity of moral judgement and understand the subjectivity of moral classification of texts.

**Acknowledgements** This work was supported with supercomputing resources provided by the Phoenix HPC service at the University of Adelaide.

## References

Araque, O.; Gatti, L.; and Kalimeri, K. 2020. MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowl. Based Syst.*, 191(105184): 105184.

Atari, M.; Haidt, J.; Graham, J.; Koleva, S.; Stevens, S. T.; and Dehghani, M. 2023. Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *J. Pers. Soc. Psychol.*, 125(5): 1157–1188.

Beiró, M. G.; D'Ignazi, J.; Perez Bustos, V.; Prado, M. F.; and Kalimeri, K. 2023. Moral Narratives Around the Vaccination Debate on Facebook. In *Proceedings of the ACM Web Conference 2023*, WWW '23, 4134–4141. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394161.

Beiró, M. G.; D'Ignazi, J.; Prado, M. F.; Bustos, V. P.; and Kalimeri, K. 2022. Moral narratives around the vaccination debate on Facebook.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of deep bidirectional Transformers for language understanding.

Ellemers, N.; and Van der Toorn, J. 2015. Groups as moral anchors. *Curr. Opin. Psychol.*, 6: 189–194.

Frimer, J. A.; Boghrati, R.; Haidt, J.; Graham, J.; and Dehgani, M. 2019. The Moral Foundations Dictionary for Linguistic Analyses 2.0. https://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/moral-foundations-dictionary/. Unpublished manuscript.

Graham, J.; and Haidt, J. 2012. The Moral Foundations Dictionary. https://moralfoundations.org/other-materials/. Unpublished manuscript.

Graham, J.; Haidt, J.; Koleva, S.; Motyl, M.; Iyer, R.; Wojcik, S. P.; and Ditto, P. H. 2013. Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In *Advances in Experimental Social Psychology*, Advances in experimental social psychology, 55–130. Elsevier.

Graham, J.; Haidt, J.; and Nosek, B. A. 2009. Liberals and conservatives rely on different sets of moral foundations. *J. Pers. Soc. Psychol.*, 96(5): 1029–1046.

Graham, J.; Nosek, B. A.; Haidt, J.; Iyer, R.; Koleva, S.; and Ditto, P. H. 2011. Mapping the moral domain. *J. Pers. Soc. Psychol.*, 101(2): 366–385.

Guo, S.; Mokhberian, N.; and Lerman, K. 2023. A data fusion framework for multi-domain morality learning. *Proceedings of the International AAAI Conference on Web and Social Media*, 17: 281–291.

Haidt, J. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychol. Rev.*, 108(4): 814–834.

Haidt, J. 2012. *The Righteous Mind: Why Good People are Divided by Politics and Religion*. Penguin UK.

Haidt, J.; and Graham, J. 2007. When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize. *Soc. Justice Res.*, 20(1): 98–116.

Haidt, J.; Koller, S. H.; and Dias, M. G. 1993. Affect, culture, and morality, or is it wrong to eat your dog? *J. Pers. Soc. Psychol.*, 65(4): 613–628.

Hoover, J.; Johnson-Grey, K. M.; Dehghani, M.; and Graham, J. 2017. Moral Values Coding Guide.

Hoover, J.; Portillo-Wightman, G.; Yeh, L.; Havaldar, S.; Davani, A. M.; Lin, Y.; Kennedy, B.; Atari, M.; Kamel, Z.; Mendlen, M.; Moreno, G.; Park, C.; Chang, T. E.; Chin, J.; Leong, C.; Leung, J. Y.; Mirinjian, A.; and Dehghani, M. 2020. Moral Foundations Twitter Corpus: A collection of 35k tweets annotated for moral sentiment. *Soc. Psychol. Personal. Sci.*, 11(8): 1057–1071.

Hopp, F. R.; Fisher, J. T.; Cornell, D.; Huskey, R.; and Weber, R. 2021. The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behav. Res. Methods*, 53(1): 232–246.

Huang, X.; Wormley, A.; and Cohen, A. 2022. Learning to adapt domain shifts of moral values via instance weighting. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*. New York, NY, USA: ACM.

Jiang, J.; Luceri, L.; and Ferrara, E. 2025. Moral Values Underpinning COVID-19 Online Communication Patterns. In *Companion Proceedings of the ACM on Web Conference 2025*, 2642–2650. New York, NY, USA: ACM.

Koleva, S. P.; Graham, J.; Iyer, R.; Ditto, P. H.; and Haidt, J. 2012. Tracing the threads: How five moral concerns (especially Purity) help explain culture war attitudes. *J. Res. Pers.*, 46(2): 184–194.

Kraft, P. W.; and Klemmensen, R. 2024. Lexical Ambiguity in Political Rhetoric: Why Morality Doesn't Fit in a Bag of Words. *British Journal of Political Science*, 54(1): 201–219.

Lin, Y.; Hoover, J.; Portillo-Wightman, G.; Park, C.; Dehghani, M.; and Ji, H. 2018. Acquiring Background Knowledge to Improve Moral Value Prediction. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 552–559.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization.

Nguyen, T. D.; Chen, Z.; Carroll, N. G.; Tran, A.; Klein, C.; and Xie, L. 2024. Measuring Moral Dimensions in Social Media with Mformer. *Proceedings of the International AAAI Conference on Web and Social Media*, 18: 1134–1147.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E. Z.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *CoRR*, abs/1912.01703.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12(null): 2825–2830.

Preniqi, V.; Ghinassi, I.; Ive, J.; Saitis, C.; and Kalimeri, K. 2024. MoralBERT: A Fine-Tuned Language Model for Capturing Moral Values in Social Discussions. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, GoodIT '24, 433–442. New York, NY, USA: Association for Computing Machinery. ISBN 9798400710940.

Rodrigues, F.; and Pereira, F. C. 2018. Deep learning from crowds. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press. ISBN 978-1-57735-800-8.

Trager, J.; Ziabari, A. S.; Davani, A. M.; Golazizian, P.; Karimi-Malekabadi, F.; Omrani, A.; Li, Z.; Kennedy, B.; Reimer, N. K.; Reyes, M.; Cheng, K.; Wei, M.; Merrifield, C.; Khosravi, A.; Alvarez, E.; and Dehghani, M. 2022. The Moral Foundations Reddit Corpus.

Zhou, A.; Liu, W.; Kim, H. M.; Lee, E.; Shin, J.; Zhang, Y.; Huang-Isherwood, K. M.; Dong, C.; and Yang, A. 2024. Moral Foundations, Ideological Divide, and Public Engagement with U.S. Government Agencies' COVID-19 Vaccine Communication on Social Media. *Mass Communication and Society*, 27(4): 739–764.