Contextualizing Language Models for Norms Diverging from Social Majority

Niklas Kiehne, Hermann Kroll, Wolf-Tilo Balke
Institute for Information Systems
TU Braunschweig
Braunschweig, Lower Saxony, Germany
{kiehne, kroll, balke}@ifis.cs.tu-bs.de

Abstract

To comprehensibly contextualize decisions, artificial systems in social situations need a high degree of awareness of the rules of conduct of human behavior. Especially transformer-based language models have recently been shown to exhibit some such awareness. But what if norms in some social setting do not adhere to or even blatantly deviate from the mainstream? In this paper, we introduce a novel mechanism based on deontic logic to allow for a flexible adaptation of individual norms by de-biasing training data sets and a task-reduction to textual entailment. Building on the popular 'Moral Stories' dataset we on the one hand highlight the intrinsic bias of current language models, on the other hand characterize the adaptability of pre-trained models to deviating norms in fine-tuning settings.[1]

1 Introduction

Social norms - whether explicitly codified or just widely agreed upon - to a large degree govern the interaction of humans. In that sense they allow for assessing and making sense of everyday situations. Thus, also the successful deployment of AI systems in social settings, e.g., conversational agents or decision making systems, will depend on the ability of such systems to adequately reflect existing social norms (Bicchieri, 2005). Recently, studies on transformer-based language models (LMs) have shown that indeed there seems to be a *'moral dimension' to LMs*, as they show high accuracy in related downstream tasks such as moral reasoning and action classification (Forbes et al., 2020; Emelin et al., 2021; Schramowski et al., 2022). Arguably, this notion of morality can be attributed to the LMs' pre-training corpora containing social majority biases also exhibited by the later used benchmarks, which are often gained by general crowd-sourcing tasks. Thus, throughout this paper we will understand the acquisition of the social norms by LMs in a descriptive, rather than an explicit prescriptive fashion.

While it is notoriously difficult to effectively remove *all* bias from AI systems, using known biases to fulfill some specific and clearly defined goals has been investigated, see e.g, (Hendrycks et al., 2021b; Ammanabrolu et al., 2022). *But what if desirable norms in some social setting do not adhere to or even blatantly deviate from norms agreed upon by the social majority, e.g., specific norms of social subgroups?* Is a completely new pre-training needed to override generally accepted norms in a seemingly consistent normative system in an LM's moral dimension? Since the costs of building huge and necessarily well-curated pre-training corpora for each social subgroup are clearly prohibitive, this option is of a rather theoretical nature. Could simpler techniques like fine-tuning then effectively create sufficient awareness in language models to allow for successful downstream tasks?

In this paper we will investigate the question of *how well general purpose language models can 'tune in' to norms deviating from majority society*. Building on deontic logic for *norm inversion* we perform extensive experiments allowing to remove arbitrary norms with respect to benchmark corpora or even impose contrasting norms during fine tuning. On the technical level, we show how to *construct the necessary datasets for fine tuning* such that models can achieve a high degree of accuracy. Thus, the actual norm acquisition always stays of a strongly descriptive nature, since we impose no explicit mechanisms to guarantee that the LM will always adhere to explicitly altered norms.

Due to the problems of deriving adequate document sets for individual social subgroups in the real world, however, within the scope of this paper we perform only *synthetic* experiments on often used real world datasets (in particular Social Chemistry,

[1] Data and code on GitHub: https://github.com/nikrruun/contrastive_moral_stories

Forbes et al. 2020 and Moral Stories, Emelin et al. 2021). Although this is a clear limitation of the work presented here, the paper's basic techniques and insights promise to allow for generalization. We will point out and critically assess limitations and possible problems for generalization in all parts of this work.

This paper is organized as follows: In Section 2 we will revisit related work and especially take a closer look at typical datasets and downstream tasks in the field of Moral AI. As the goal of this paper is to investigate the stability of pre-trained language models in the face of conflicting norms, we take a closer look at the task of moral action classification in Section 3. We provide a detailed description of our dataset design and creation in Section 4 as a basis for later experiments. This also includes the inversion of arbitrary norms following the rules provided by deontic logic. Section 5 will then present the actual experimental investigation of our hypothesis that fine-tuning may be a suitable remedy for the task of reflecting norms differing from social majority in downstream tasks. After a discussion in Section 6, we close with our conclusions in Section 7.

2 Related Work

There is a growing body of work concerning the development of AI/machines that behave ethically and/or are aligned with human values. For example, Prabhumoye et al. (2021) investigate potential applications of deontological ethics in the context of NLP and, similarly to Hooker and Kim (2018), study the first-principles of *generalization* and *autonomy*. Other works have prioritized aligning artifical agents with *shared human values* (Soares, 2018). Value alignment has been approached from numerous angles, including preference learning (Gabriel, 2020; Christiano et al., 2017), imitation (Ho and Ermon, 2016) and inverse reinforcement learning (Nahian et al., 2020; Hadfield-Menell et al., 2016). Additionally, several approaches concerning controllable text generation have been proposed to steer model generation towards specific attributes (Dathathri et al., 2020; Keskar et al., 2019). However, our experimental setup focuses on classification tasks instead of generation. Similarly, Kulkarni et al. (2021) incorporate speaker context into language model pre-training objectives. Here, we do not consider pre-training, but rather only investigate fine-tuning.

Several datasets of normative knowledge have been published to assess to which extent current models are able to represent specific *morality* or ethical rules. One important aspect of the benchmarks is their degree of implicitness of normativity. In this regard, implicit datasets usually contain examples of right and wrong behavior with according labels (Hendrycks et al., 2021a,b; Nahian et al., 2020; Lourie et al., 2021), whereas others rely on explicitly stating the social rules at play (Emelin et al., 2021; Forbes et al., 2020; Jiang et al., 2021b).

Forbes et al. (2020) introduce Social Chemistry 101, a large collection of so-called rules-of-thumb (RoT) associated with a rich structure of human annotations. According to the authors, these RoTs were designed to represent social norms and moral judgment as experienced by crowd-workers.

With Moral Stories, Emelin et al. (2021) propose self-contained branching narratives consisting of norms, context, moral and immoral actions and their expected consequences as written by crowd-workers. The authors suggest that the RoBERTa (Liu et al., 2019) model exhibits a normativity bias due to pre-training, but they do not follow up with investigations.

Other datasets, e.g., ETHICS (Hendrycks et al., 2021a) or Scruples (Lourie et al., 2021) also present resources containing normative information, but only through examples, whereas explicit mentions of the appropriate norms and rules are required for our purposes. Although conceivable, we leave adaptation of full paragraphs or stories to reflect contrary values or norms for future work, since current language models have been shown to lack the capabilities of dealing with the various nuances of negation and contradiction (Jiang et al., 2021a). In this paper, we focus on the action-classification task introduced by Emelin et al. and propose a controlled approach for negation grounded in deontic logic.

With COMMONSENSE NORM BANK, Jiang et al. (2021b) compile several benchmarks into a large collective of moral judgment Q&A tasks. One aspect of their work is similar to ours, as they also derive augmented norms from the Moral Stories dataset. However, their focus is on deriving equivalent norms through morality-preserving transformations, whereas we explicitly opt for the derivation of opposite norms.

Finally, perhaps most similar is the work of Arora et al. (2022), who investigate to which degree

pre-trained language models reflect cross-cultural values according to external value surveys. Their experiments employ probing techniques and provide evidence of normativity biases, but only weak alignment to the surveys. In contrast, we aim to analyze to which extent the models are able to reflect norms explicitly deviating from majority-imposed bias.

3 Moral action classification

Several works have provided evidence of pre-trained language models achieving notable results in approximating human decision-making in social context. To assess to which extent the models are able to generalize, researchers frequently turn to analyses of previously unseen situations. However, in many cases, these "test" sets stem from the same population that curated the data used for priming the models in the first place (e.g., gathered from crowd-workers with high agreement levels). This connection becomes even more apparent in the case of language models utilizing pre-training, which have been suspected to contain a *normativity* bias (Jiang et al., 2021b; Emelin et al., 2021). Recently, Arora et al. (2022) found such bias in their study of cross-cultural value alignment of PLMs. In light of these findings, we aim to test model generalizability from a broader perspective. If PLMs do contain a bias towards a specific set of norms, then to which extent are they adaptable to new norms? In this paper, we focus on the case of norms explicitly contrary to what has been argued to be picked up during pre-training. The motivation is to reflect the inherent nature of social subgroups, which usually oppose certain norms imposed by majority society. However, it is not yet clear, what *opposites*, or *inversions* of norms are. We adopt deontic logic, which formalizes relations between norms, such as *contrary* or *contradictory*. Deontic logic requires norms to be directly expressed, which rules out many of the published benchmarks as potential bases. To the best of our knowledge, Moral Stories is the only benchmark so far incorporating norms, actions and corresponding labels of adherence or violation. Therefore, of the many proposed tasks to assess normative knowledge in language models, we consider action classification for its explicitness and clear-cut semantics.

Thus, we build on and extend definitions and data from Moral Stories (Emelin et al., 2021) and Social Chemistry 101 (Forbes et al., 2020). More specifically, we define the action classification task following Emelin et al. (2021): We focus on the setting of actions grounded by corresponding norms. Although Moral Stories also provides context as well as consequences for grounding, we omit them in this paper due to the increased complexity. For the remainder of the paper we understand moral action classification as the (norm, action)-scenario, where the task is to decide for any such pair whether the action is deemed moral or immoral with respect to the norm. Models are evaluated on the accuracy metric. For further details, see Emelin et al. (2021).

4 Dataset design and creation

In the following subsections we briefly introduce deontic logic as the theoretic foundationn of our norm inversion procedure. Then we show how Moral Stories and Social Chemistry 101 datasets relate to the theoretical considerations, and lastly we present and evaluate two automatically derived sets of opposing norms, namely *anti-ms* and *optional-ms*. Note that we use deontic logic only as means to derive new norms by inversion. We explicitly do not use it for logical inference, as this would require the underlying datasets to be consistent. But, as already pointed out by their authors, neither Moral Stories nor Social Chemistry 101 are designed to be free of contradictions. For example, "You shouldn't let animals suffer." and "You should not kill animals.", both from Moral Stories, could be mutually exclusive under certain circumstances.

4.1 Deontic Logic

Deontic logic is a field in philosophical logic that is most concerned with inferring what follows from what in terms of *obligation*, *permission* and their related concepts (McNamara and Van De Putte, 2022). It is of special interest for our work, as it provides a logical framework for normative statuses and their connections. Here, we adopt the Standard Deontic Logic (SDL) (von Wright, 1951b; Prior and Prior, 1955). Several different, though equivalent, options exist to define operators such as $\mathbf{OB}p$ (it is *obligatory* that p is the case) or $\mathbf{IM}p$ (it is *impermissible* that p is the case). In the so-called *Traditional Definitional Scheme*, **OB** is chosen as a primitive and the remaining are defined as shown in Equation 1. For example, stating that p is impermissible can be expressed as $\neg p$ *ought to be the case* or, more formally, as $\mathbf{OB}\neg p$. It has to

norm	*action-moral-judgment*	proposed SDL operator
It's cruel to kill an animal.	very bad	**I**M, impermissible
It's rude to laugh at others.	bad	**I**M, impermissible
It is okay to not find someone attractive.	expected/ok	**O**P, optional
You should follow through on your promises.	good	**O**B, obligatory
It's good to fight for the rights of others.	very good	**O**B, obligatory

Table 1: Examples of norms from Moral Stories with their associated moral judgment as provided by the Social Chemistry 101 corpus and our proposed matching to SDL operators. Note, that the neutral class "expected/ok" is not represented in Moral Stories. The example is taken from Social Chemistry 101 instead.

be noted that the theoretical framework of SDL has been heavily discussed. Several shortcomings have been brought forward, perhaps most notably Chisholm's puzzle (see, e.g., McNamara and Van De Putte 2022).

$$\begin{aligned}
\mathbf{P}\mathrm{E}p &\stackrel{def}{=} \neg\mathbf{O}\mathrm{B}\neg p \textit{ (permissible)} \\
\mathbf{I}\mathrm{M}p &\stackrel{def}{=} \mathbf{O}\mathrm{B}\neg p \textit{ (impermissible)} \\
\mathbf{O}\mathrm{M}p &\stackrel{def}{=} \neg\mathbf{O}\mathrm{B}p \textit{ (omissible)} \\
\mathbf{O}\mathrm{P}p &\stackrel{def}{=} (\neg\mathbf{O}\mathrm{B}\neg p \wedge \neg\mathbf{O}\mathrm{B}p) \textit{ (optional)} \\
\mathbf{N}\mathrm{O}p &\stackrel{def}{=} (\mathbf{O}\mathrm{B}p \vee \mathbf{O}\mathrm{B}\neg p) \textit{ (non-optional)}
\end{aligned} \tag{1}$$

4.2 Moral Stories with a twist

How can we reason by Moral Stories in terms of deontic logic? First, we aim to map the provided norms to the six SDL operators. Ideally, due to the subjective nature of the topic, such a classification should be based on human judgments. The Moral Stories dataset itself does not provide any such means; however, Social Chemistry 101 (Forbes et al., 2020), the benchmark it was derived from, does. According to Forbes et al., the crowdworkers were instructed to classify the *moral judgment* of the rules-of-thumb into "very bad", "bad", "expected/OK", "good", and "very good".[2] See Table 1 for examples. We interpret all negatively judged statements as elements of the *impermissible* and their positive counterparts as elements of the *obligatory* category. Further, although not present in the original Moral Stories, we map the neutral statements ("expected/ok") to the optional SDL operator.

Next, we turn to applying operator equivalences to derive new statements. For practical purposes, implementable counterparts in the natural language domain are needed for the logical transformations of SDL. We only consider the operators needed, and, since only **O**B and **I**M occur in the dataset, we thus focus on these. Furthermore, of the many possible transformations in SDL, we restrict ourselves to only those reflecting negation. As per definitions, we then receive *omissible* and *permissible* operators as opposites, as shown below:

$$\begin{aligned}
\neg\mathbf{O}\mathrm{B}p &\stackrel{def}{=} \mathbf{O}\mathrm{M}p \textit{ (omissible)} \\
\neg\mathbf{I}\mathrm{M}p = \neg\mathbf{O}\mathrm{B}\neg p &\stackrel{def}{=} \mathbf{P}\mathrm{E}p \textit{ (permissible)}
\end{aligned} \tag{2}$$

Negation in natural language The ability to negate or contradict statements is a unique property of human language (Horn and Wansing, 2020). In contrast to logical negation, humans often deal with varying shades of semantic opposition (Jiang et al., 2021a). For example, consider the statement "You should not eat meat", which might be a valid norm in the subgroup of vegetarians. At least two opposites to this statement are conceivable in terms of semantic negation, ranging from the rather obligatory "You *should* eat meat" to a more optional perspective reflecting that eating meat is permissible but not obligatory. Both statements are *contrary* to the original, but not *contradictory*, since neither fully exhaust their domain. Deontic logic behaves similarly in this regard and provides well-defined domains for contrary and contradictory statements (von Wright, 1951a). Here, as shown in Equation 2, the contradictory opposite to "You should not eat meat" (impermissible) is *eating meat* as a permissible action, which, by definition, entails both the obligatory and optional cases. In the context of deontic logic, the correct approach to deriving opposites of norms is thus to generate permissible statements from impermissible ones and to use the obligatory norms to derive omissible statements. Natural language, however, does not necessarily provide equally clear-cut definitions for permissible and omissible notions as SDL does. Here, we decide to decompose both operators into disjunctions of less ambiguous operators. For ex-

[2] We refer to the *action-moral-judgment* column here.

norm	dataset	rating
It's wrong to ignore someone after a one-night stand.	Moral Stories	
It is ambitious to ignore someone after a one-night stand.	*anti-ms*	0/3
It is acceptable to ignore someone after a one-night stand.	*optional-ms*	3/3
Not tipping your server is rude.	Moral Stories	
You should not tip your server.	*anti-ms*	3/3
It is okay to tip your server.	*optional-ms*	3/3
You shouldn't stop communicating with a friend.	Moral Stories	
It's bad to communicate with a friend.	*anti-ms*	1/3
It's ok to stop communicating with a friend.	*optional-ms*	3/3

Table 2: Examples of model-generated norms and correctness ratings.

ample, Equation 3 shows the equivalence of the omissible operator to a disjunction of impermissible and optional.

$$\begin{aligned} \mathbf{OM}p &= \neg\mathbf{OB}p \wedge (\neg\mathbf{OB}\neg p \vee \mathbf{OB}\neg p) \\ &= \begin{array}{l}(\neg\mathbf{OB}p \wedge \neg\mathbf{OB}\neg p) \vee \\ (\neg\mathbf{OB}p \wedge \mathbf{OB}\neg p)\end{array} \\ &= \mathbf{OP}p \vee \mathbf{IM}p \end{aligned} \tag{3}$$

The permissible case can be approached analogously and we arrive at:

$$\mathbf{PE}p = \mathbf{OB}p \vee \mathbf{OP}p \tag{4}$$

We derive *two* datasets of opposing norms from Moral Stories based on the appropriate decomposition. Since both disjunctions consist of a non-optional (**OB** or **IM**) and an optional operator, we refer to the two datasets as *non-optional-ms* or *anti-ms* and *optional-ms*, respectively.

But how do the logical transformations relate to the domain of natural language? In the following, we carefully retrace the steps presented above on the textual representation of norms and show how to assign appropriate labels for corresponding actions.

optional-ms The case of optional statements is quite different from the impermissible or obligatory norms, mainly due to the non-existence of norm-divergent behavior. Continuing the earlier example, the norm "It is okay to eat meat" would not be violated both by *eating meat* and by *eating salad*. Note that "It is okay not to eat meat" is an equivalent statement, which, in SDL, immediately follows from $\mathbf{OP}p \Leftrightarrow \mathbf{OP}\neg p$. However, recent works have shown that language models, especially PLMs, are performing much worse on negated concepts as compared to the affirmative versions (Kassner and Schütze, 2020). To minimize the effect on our dataset, we represent norms in *optional-ms* without the added negation. Finally, given a norm from Moral Stories with its respective normative and norm-divergent actions, we consider *both* actions to be normative to the norm's optional counterpart.

norm	moral action	immoral action
Fat shaming is mean	moral	immoral
It's okay to fat shame	moral	moral
It's good to fat shame	immoral	moral

Table 3: Labels of moral and immoral actions on an original norm from Moral Stories and two variants from *optional-ms* and *anti-ms*. Note that the terms "moral"- and "immoral action" are always interpreted from the Moral Stories perspective. Thus, in the last example, we consider a formerly *moral* action to be immoral.

anti-ms The non-optional cases cover negation in a symmetrical fashion, since obligatory and impermissible are mutually contrary here. Still, there are multiple options of carrying out the negation in the text domain. For example, the corresponding obligatory statement to the impermissible "It's rude to laugh at others" could be expressed as either "It's not rude to laugh at others" or "It's rude not to laugh at others". Here, we opt for the first version in order not to complicate the task unnecessarily. We simplify negated judgments ("It's not rude") whenever possible (e.g. "It's nice") to specifically rule out any optional characteristics. Lastly, the labels for non-optionally negated assessments are derived as opposites to the originals. That is, formerly normative actions are considered non-normative and

vice versa. Table 3 shows an example of the label derivation.

Generating coherent norms For either dataset, the general idea is to adapt a norm from Moral Stories in a way that reflects the semantics of the corresponding operator. To this end, we utilize the plethora of examples in the Social Chemistry 101 corpus. Note that we filter out entries of low agreement and only consider the categories *social-norms* and *morality-ethics*. We extract ∼110k triples of moral judgment ("It is rude"), associated action ("laughing at others") and the resulting rule-of-thumb ("It is rude to laugh at others"). Next, we finetune a text-to-text language model to predict norms from judgment and action parts.[3] The goal is to later replace judgments according to a specific operator and to apply the model to create grammatically sound sentences.

For training, we split into 80%/10%/10% train, validation and test data and perform hyperparameter grid-search[4] on two encoder-decoder models T5 (Raffel et al., 2020) and BART (Lewis et al., 2020).[5] Refer to Appendix A.1 for details. Finally, the best performing model as shown in Table 4 is used to generate our two contrary datasets.

Since the opposing norms should not always show the same linguistic representation (rude-nice, good-bad, etc.), we sample the expressions to be used for a concrete norm from a pool of a-priori collected, human-written positive/negative samples from the Social Chemistry-101 dataset. In particular, we sample from about 500 unique linguistic expressions for obligatory norms and from about 1000 expressions for impermissible statements. This results in a wide variety of linguistically different expressions present in the data. Effectively, we allow for 500k different conversions from obligatory to impermissible norms and vice versa, although random sampling does of course not select all of them. Moreover, the norms in the parent corpora are not necessarily represented in a unique form themselves. For instance, we observed multiple statements regarding the action of "stealing something" phrased in different ways (theft, robbery, etc.), which taken with the random sampling accounts for even more variety.

[3]For the generation task the input was formatted as `<CLS>judgment<SEP>action<SEP>`.

[4]We explore ranges of batch size $\{16, 32, 64, 128\}$ and learning rates $\{1e\text{-}5, 3e\text{-}5, 5e\text{-}5\}$

[5]We use implementations of the popular *transformers* library (Wolf et al., 2020).

Model	Loss	BLEU-4	ROUGE-L
t5-small	0.019	88.33	94.96
t5-base	0.034	89.10	95.34
bart-base	0.018	89.63	95.46
bart-large	0.022	**90.00**	**95.62**
baseline		56.48	89.40

Table 4: Best achieved generation metrics for the two architectures T5, BART and baseline on test data.

We ran an ablation experiment investigating whether the sampled judgement expressions might introduce any cues for models to exploit on the later classification task. Consider two settings: first, models are given only the action to decide for moral/immoral classes and second, models have access to judgement+action (omitting the behavior description part of the norm). Neither BERT nor RoBERTa showed statistically significant differences in accuracy between both settings. Hence, we can safely argue that the judgements in anti-ms indeed cannot be readily exploited.

Quality In our evaluation we first apply automatic metrics to find best working settings and then perform a quantitative analysis of the generated samples on the leading approach. We report BLEU-4 (Papineni et al., 2002) and ROUGE-L (Lin, 2004) metrics in 4. While the metrics might seem unusually high, it has to be stressed that the task difficulty is considerably lower than for usual text generation problems with a more open-ended task. In our case, much of the needed output is already contained in the input data and only minor morphological transformations need to be carried out, e.g., verb inflection ("laughing", "to laugh"), which pre-trained language models have been shown to perform well on (Cotterell et al., 2018). As a baseline we include simple concatenation of the two input parts.

For the qualitative evaluation we asked annotators to judge the correctness of a random sample of 200 generated norms. They had to assess whether the generated norms do express the opposite judgement of the original norm and whether the generated sentences were grammatically correct. We trained graduate students how to annotate arbitrary norms by providing categories for positive, neutral, and negative normative judgements and showed how this reflects on the possible counterparts. A generated counterpart would only be annotated as correct, if the respective action is still the same, the judgement has been inverted and the sentence was

grammatically correct. In any other case, a generated norm was to be annotated as incorrect. We set up three crowdsourcing tasks for each norm of the random sample and for each norm recorded the majority decision and the annotator agreement. In summary, about 95% of our generations were rated as correct, with all three raters positively agreeing in almost 90% of assessments. Multiple examples of correct and incorrect generations are shown in Table 2. Appendix A.1 provides further details into rater agreement.

5 Experiments

We conduct several experiments based on the original Moral Stories (*original-ms*), *anti-ms* and *optional-ms* datasets over variations of the moral action classification task. We include seven models in our studies: DistilBERT (66M) (Sanh et al., 2019) as a rather small model, BERT (110M & 336M) (Devlin et al., 2019), since they are among the most used, RoBERTa (355M) (Liu et al., 2019) to ensure comparability with Moral Stories, ALBERT (223M) (Lan et al., 2020) for its exceptional performance on the ETHICS benchmark and lastly, GPT-Neo (1.3B & 2.7B) (Black et al., 2021) as representatives of larger transformer models.

5.1 Transfer learning

In our first setting we investigate whether pre-trained language models transfer well from one dataset to the others. To this end, we adopt the following procedure: Each model is fine-tuned separately on the three datasets.[6] After fine-tuning, the best model configuration per dataset is loaded and tested against all others, see Table 5 for results. Note that *optional-ms* does not contain samples of norm-diverging behavior and therefore only contains a single label, serving as an extreme case.

The achieved accuracy of RoBERTa on Moral Stories effectively reproduces the original paper (Emelin et al., 2021) and ALBERT sets a new state-of-the-art accuracy of 94.3%. Second, larger models do not automatically perform better, which is in contrast to findings of other studies (Kaplan et al., 2020). Even the largest model (2.7B) is outperformed by models a tenth its size. The amount of pre-training data also does not seem to majorly influence the scores for moral reasoning. For example, both best (ALBERT) and worst performing models (DistilBERT) rely on the same corpora. However, it is unclear whether this is due to insufficient fine-tuning, the models' architectures, or other differences. More work is needed to explore these discrepancies.

Concerning *optional-ms* as fine-tuning corpus, we found all hyper-parameter configurations across all models to produce the same outcome. Although, the perfect accuracy is expected for actual optional norms since only a single class needs to be considered. Hence, optional norms alone are inadequate. On the other hand, neither original nor *anti-ms* datasets allow models to correctly infer optional norms. It seems that fine-tuning does not transfer more general reasoning capabilities from optional to non-optional or vice versa. Interestingly, when comparing original to *anti-ms*, the picture is quite different. Here, models do not collapse to random guessing (50%), but perform even worse. It appears that fine-tuning causes models to adapt to the presented norms beyond the task and that some aspects are internalized.

Emelin et al. suggest a *normativity bias* attributed to pre-training. Arora et al. (2022) further corroborate these findings via probing tasks. However, as soon as fine-tuning is involved, such bias does not seem to significantly favor datasets of similarly biased norms.[7] For example, models fine-tuned on *anti-ms* were found to perform comparably to those trained on Moral Stories, with the largest difference of 1.2% (ALBERT & GPT-Neo).

To further analyze the effect of pre-training, we conducted additional experiments where models have access to more than one of the datasets. Due to their unsatisfactory performance, GPT-Neo variants are not included hereafter.

5.2 Conflicting Moral Stories

So far, models were only given access to single datasets at a time. This restriction appears reasonable from the perspective of dataset consistency, since all subsets are in some sense opposing the others. In this setting we explicitly study the ability of LMs to pick up the various notions of contrary norms during fine-tuning. Consequently, models were trained on the union of Moral Stories, *anti-ms* and *optional-ms*. We refer to this set as *conflicting-ms*.

The results in Table 6 show generally decreased

[6] Again, we explore ranges of batch size $\{16, 32, 64, 128\}$ and learning rates $\{1e\text{-}5, 3e\text{-}5, 5e\text{-}5\}$ on four epochs.

[7] Emelin et al. argue that Moral Stories likely contains norms also present in pre-training corpora.

Fine-tuned on:	Moral Stories			*anti-ms*		
Model	Tested on:					
	ms	***anti-ms***	***o.-ms***	***ms***	***anti-ms***	***o.-ms***
distilbert-base	78.0	22.1	52.4	23.6	77.0	49.4
bert-base	80.7	22.2	49.0	30.3	80.7	53.1
bert-large	82.6	19.4	53.5	30.9	82.9	52.0
roberta-large	92.5	43.7	49.1	23.1	91.4	53.8
albert-xxlarge-v2	**94.2**	**45.5**	**54.4**	27.8	**93.0**	**55.9**
gpt-neo-1.3B	83.0	30.3	50.8	30.4	82.4	42.8
gpt-neo-2.7B	86.2	38.2	51.2	**35.4**	85.0	46.5

Table 5: Accuracies of various pre-trained models on three variants of Moral Stories. We report metrics computed on the test data of the norm-distance split. The reported scores are those of the best performing hyper-parameter settings on the respective sub-task. On the *optional-ms* dataset all models achieved 50% (*ms*), 50% (*anti-ms*) and 100% on (*optional-ms*). See Appendix A.2 for details.

peak accuracy, e.g., DistilBERT suffers a loss of 8% on original Moral Stories. However, the models were able to outperform random guessing in all instances. When models were not initialized via pre-trained weights, but randomly, none of the considered settings learn meaningful representations. Rather, the models seem to simply predict the majority label ("moral").

5.3 Textual entailment

The leading approaches on several benchmarks assessing the normative knowledge of LMs, including ours, rely on fine-tuning on a custom-tailored corpus. Naturally, questions arise to what extent fine-tuning introduces new information into the models and whether it can be excluded from the experiments, i.e. through prompting. Related work reports significantly worse performance for prompting techniques as compared to fine-tuning (Jiang et al., 2021b; Hendrycks et al., 2021a). It remains unclear whether the accuracy of prompting is lower due to absence of normative information or simply due to the higher task complexity. Here, we want to show a possible connection of moral reasoning and natural language inference in a zero-shot paradigm (Yin et al., 2019).

Deciding whether a text entails a hypothesis in terms of natural language is the domain of textual entailment (Bowman et al., 2015; Nie et al., 2020). We propose a mapping of polarity of entailment and norm to complement our results. For example, considering the action "*X* eats a steak" with respect to the norm "It's bad to eat meat" implies *X* acted immorally, since *eating steak* is entailed by *eating meat*, which in turn is considered *bad*. Accordingly, we train a classifier to categorize norms as obligatory, impermissible or optional. The task turned out to be simple, since even small models (bert-base) achieved $\sim$ 98%. Next, we apply a textual entailment model (Nie et al., 2020), whose task is to determine whether an action satisfies the behavior as described by some norm.

We consider two scenarios. At first, the textual entailment component is applied as is, representing a true zero-shot setting. The problem is that the model input is slightly different to that of the original task. To counter the issue we also fine-tune it on corresponding extracts of Moral Stories devoid of the judgment aspect. E.g., we take into account only *eating steak* as premise and *eating meat* as hypothesis, but not the full norm.

The results of both approaches are shown in Table 6. In the zero-shot setting the pipeline performs comparably to a fine-tuned bert-large model with full access to the data. With fine-tuning enabled, textual entailment achieves second best scores in three out of four cases.

6 Discussion

We used concepts of standard deontic logic to derive norms contrary to those of Moral Stories. Overall, SDL can only be viewed as one of many possible frameworks that could be used. We reiterate that we explicitly do not adopt SDL for reasoning purposes, but only for its clear-cut definitions of operators, which we deem transferable to the natural language domain. Here, we intuitively map definitions of Moral Stories to those of deontic logic. Specifically, we interpret human judgment as salient indicators. Experiments on polarity clas-

Model	Pre-trained				Randomly initialized			
	ms	***anti-ms***	***o.-ms***	***confl.-ms***	***ms***	***anti-ms***	***o.-ms***	***confl.-ms***
distilbert-base	70.0	65.1	99.6	78.2	50.0	50.0	100.0	66.7
bert-base	75.4	74.0	99.7	83.0				
bert-large	78.4	77.2	**99.8**	85.1	⋮	⋮	⋮	⋮
roberta-large	89.1	86.4	99.5	91.7				
albert-xxlarge-v2	**90.8**	**88.1**	99.6	**92.8**	50.0	50.0	100.0	66.7
TE-no-fine-tune	78.2	76.7	99.2	84.7				
TE-fine-tuned	90.1	87.9	99.2	92.4				

Table 6: Fine-tuning on the union of Moral Stories and its derivations, called *conflicting-ms*. The two lines at the bottom refer to the approaches based on textual entailment, which naturally require previous training.

sification (see Section 5.3) indicate that the mapping indeed results in distinct classes. One major difference of SDL and natural language is the way negation is handled. While the former provides an exhaustive operationalization of negation, the latter is much more nuanced (Jiang et al., 2021a). Here, we effectively contract a multitude of textual judgments (e.g. "It is good to") into three equivalence classes. These, in turn, are then associated with SDL operators.

7 Conclusion

We investigated the abilities of language models to simultaneously represent opposing sets of norms in the context of a moral action classification setting. Based on notions from deontic logic, we derived two such sets from the Moral Stories benchmark and ran extensive evaluations on a range of architectures. Our results suggest that fine-tuning on just one of the sets imposes a strong bias onto the models, in the sense that the left out norms are severely misrepresented. Further, when subjected to highly conflicting norms, we found pre-training to play an essential role for models to adapt well. Models that were not pre-trained and thus are not affected by possible bias towards specific norms were found to collapse to random guessing. However, contrary to intuition, with pre-training enabled, the models were able to reconcile even most inconsistent normative settings. Finally, we propose one option to factor out the reasoning aspect of the task into textual entailment. The approach performs on par to the best fine-tuned model.

Limitations

The strongest limitation in our paper is drawing our conclusions for de-biasing PLMs for individual social subgroups from experiments on synthetically built datasets. On the positive side the creation of datasets by norm inversion from often used real world datasets leads to a high rater agreement in terms of syntactic correctness. Whether the specific form and a possible inner coherence of real world norms for specific social subgroups would have made a difference, remains, however, an open question. The necessary size of respective datasets for both pre-training and fine-tuning makes their collection difficult and is thus left for future work.

In line with recent works, our experiments make heavy use of fine-tuning. Although others have also investigated probing techniques, there are more options to adapt PLMs. For example, model editing tools have shown recent success in changing factual knowledge in PLMs (De Cao et al., 2021). Whether methods targeted at factual information can be adapted to the moral knowledge is unclear. Although our work provides insight into the adaptability of LMs to diverging social norms, we do not investigate the consequences of introducing contradictory statements into the models for downstream tasks – additional efforts are required. To this end, future research might leverage existing tools, e.g. LAMA (Petroni et al., 2019), to assess the impact of charging LMs with specific social norms. Moreover, we compare fine-tuning performances of pre-trained vs. randomly initialized models on the same range of hyper-parameters. While longer training on non-pre-trained instances could improve results, we decided to keep the computational costs fixed across both experiments, possibly giving an advantage to the pre-trained cases. Finally, our work only considers one specific natural language, due to the required datasets missing for other societies. However, we deem the presented methods transferable to other languages, given that a reasonable mapping to deontic logic operators is possible.

References

Prithviraj Ammanabrolu, Liwei Jiang, Maarten Sap, Hanna Hajishirzi, Yejin Choi, and Noah A. Smith. 2022. Aligning to normative values in morally informed game environments.

Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2022. Probing pre-trained language models for cross-cultural differences in values.

Cristina Bicchieri. 2005. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large scale autoregressive language modeling with meshtensorflow.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506.

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.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 698–718. 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.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437.

Dylan Hadfield-Menell, Stuart J. Russell, Pieter Abbeel, and Anca D. Dragan. 2016. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3909–3917.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Mantas Mazeika, Andy Zou, Sahil Patel, Christine Zhu, Jesus Navarro, Dawn Song, Bo Li, and Jacob Steinhardt. 2021b. What would jiminy cricket do? towards agents that behave morally. *NeurIPS*.

Jonathan Ho and Stefano Ermon. 2016. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4565–4573.

John N. Hooker and Tae Wan N. Kim. 2018. Toward non-intuition-based machine and artificial intelligence ethics: A deontological approach based on modal logic. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 130–136, New York, NY, USA. Association for Computing Machinery.

Laurence R. Horn and Heinrich Wansing. 2020. Negation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2020 edition. Metaphysics Research Lab, Stanford University.

Liwei Jiang, Antoine Bosselut, Chandra Bhagavatula, and Yejin Choi. 2021a. "I'm not mad": Commonsense implications of negation and contradiction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4380–4397, Online. Association for Computational Linguistics.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. 2021b. Delphi: Towards machine ethics and norms.

Jared Kaplan, Sam McCandlish, T. J. Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *ArXiv*, abs/2001.08361.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Vivek Kulkarni, Shubhanshu Mishra, and Aria Haghighi. 2021. LMSOC: An approach for socially sensitive pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2967–2975, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

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*, volume 35, pages 13470–13479.

Paul McNamara and Frederik Van De Putte. 2022. Deontic Logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2022 edition. Metaphysics Research Lab, Stanford University.

Md Sultan Al Nahian, Spencer Frazier, Mark Riedl, and Brent Harrison. 2020. Learning norms from stories: A prior for value aligned agents. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 124–130.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Shrimai Prabhumoye, Brendon Boldt, Ruslan Salakhutdinov, and Alan W Black. 2021. Case study: Deontological ethics in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3784–3798, Online. Association for Computational Linguistics.

Arthur N. Prior and Norman Prior. 1955. *Formal Logic*. Oxford University Press.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3505–3506. ACM.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A. 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(3):258–268.

Nate Soares. 2018. The value learning problem. In *Artificial intelligence safety and security*, pages 89–97. Chapman and Hall/CRC.

G. H. von Wright. 1951a. Deontic Logic. *Mind*, LX(237):1–15.

G. H. von Wright. 1951b. *An Essay in Modal Logic*. Amsterdam: North-Holland Pub. Co.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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 3914–3923, Hong Kong, China. Association for Computational Linguistics.

A Supplementary Material

A.1 RoT-generator: Supplementary details

Hyper-parameter search The best performing models found during hyper-parameter search are shown in Table 9. We used the *DeepSpeed* (Rasley et al., 2020) framework on top of the *transformers* (Wolf et al., 2020) library for mixed-precision training and general speed-ups. The following hyper-parameters where kept constant: Number of warm-up steps: 0, gradient norm: 0, weight decay: 0, optimizer: AdamW, model input length: 128. During hyper-parameter search we considered batch sizes $\{16, 32, 64, 128\}$ and learning rates $\{1\text{e-}5, 3\text{e-}5, 5\text{e-}5\}$.

Rater-agreement We measure pairwise rater agreement in Table 7. Three graduate students were asked to assess generation correctness of pairs of original and generated norms. The average rate of

A+B	A+C	B+C	A+B+C
0.93	0.915	0.945	0.895

Table 7: Pairwise rater agreement in percent of equal rating for raters A,B and C.

correct generations per rater is given in Table 8. We

A	B	C
0.935	0.965	0.95

Table 8: Percentage of correct generations per rater.

also compute interrater-agreement based on Krippendorff's α (0.264) (Krippendorff, 2011). The comparatively low score is due to the heavy skewness of the rating distribution and rater's rarely disagreeing on the same samples. Rather, we found raters to only agree on incorrectness in two cases.

A.2 Classification

Hyper-parameter search We used the same parameter ranges as in the generation cases. Tables 10, 11 and 12 show the corresponding best performing parameter settings and complement Table 5 in the main paper.

Model	#parameters	loss		BLEU-4		ROUGE-L		Hyper-parameters
		eval	test	eval	test	eval	test	
bart-base	406M	0.019	0.019	89.57	89.63	95.49	95.46	bs 16, lr 3e-5
bart-large	139M	0.033	0.034	89.88	90.00	95.62	95.62	bs 16, lr 3e-5
t5-base	220M	0.019	0.018	89.01	89.10	95.33	95.34	bs 16, lr 5e-5
t5-small	60M	0.022	0.022	88.19	88.33	94.95	94.96	bs 16, lr 5e-5

Table 9: Best performing generation models after hyper-parameter search. Parameter column reports batch size (bs) and learning rate (lr).

	ms	***anti-ms***	***o.-ms***	***contra_ms***	hyperparameters
distilbert-base-uncased	78.0	22.1	52.4	50.8	bs 32, lr 5e-5
bert-base-uncased	80.7	22.2	49.0	50.6	bs 16, lr 5e-5
bert-large-uncased	82.6	19.4	53.5	51.8	bs 128, lr 3e-5
roberta-large	92.5	43.7	49.1	61.8	bs 128, lr 3e-5
albert-xxlarge-v2	**94.2**	**45.5**	**54.4**	**64.7**	bs 32, lr 1e-5
EleutherAI/gpt-neo-1.3B	83.0	30.3	50.8	54.7	bs 32, lr 1e-5
EleutherAI/gpt-neo-2.7B	86.2	38.2	51.2	58.5	bs 16, lr 1e-5

Table 10: Results of hyper-parameter search for models trained on Moral Stories.

	ms	***anti-ms***	***o.-ms***	***contra_ms***	hyperparameters
distilbert-base-uncased	23.6	77.0	49.4	50.0	bs 64, lr 3e-5
bert-base-uncased	30.3	80.7	53.1	54.7	bs 32, lr 5e-5
bert-large-uncased	30.9	82.9	52.0	55.3	bs 16, lr 1e-5
roberta-large	23.1	91.4	53.8	56.1	bs 16, lr 1e-5
albert-xxlarge-v2	27.8	**93.0**	**55.9**	**58.9**	bs 32, lr 1e-5
EleutherAI/gpt-neo-1.3B	30.4	82.4	42.8	51.9	bs 32, lr 1e-5
EleutherAI/gpt-neo-2.7B	**35.4**	85.0	46.5	55.6	bs 16, lr 1e-5

Table 11: Results of hyper-parameter search for models trained on *anti-ms*.

	ms	***anti-ms***	***o.-ms***	***contra_ms***	hyperparameters
distilbert-base-uncased	**50.0**	**50.0**	**100.0**	**66.7**	bs 32, lr 5e-5
bert-base-uncased	**50.0**	**50.0**	**100.0**	**66.7**	bs 32, lr 5e-5
bert-large-uncased	**50.0**	**50.0**	**100.0**	**66.7**	bs 32, lr 5e-5
roberta-large	**50.0**	**50.0**	**100.0**	**66.7**	bs 32, lr 5e-5
albert-xxlarge-v2	**50.0**	**50.0**	**100.0**	**66.7**	bs 32, lr 5e-5
EleutherAI/gpt-neo-1.3B	**50.0**	**50.0**	**100.0**	**66.7**	bs 32, lr 5e-5
EleutherAI/gpt-neo-2.7B	**50.0**	**50.0**	**100.0**	**66.7**	bs 32, lr 5e-5

Table 12: Results of hyper-parameter search for models trained on *optional-ms*. All hyper-parameter configurations achieved the same result, most likely due to only one label being present.