SCRUPLES: A Corpus of Community Ethical Judgments on 32,000 Real-life Anecdotes

Nicholas Lourie,[1] Ronan Le Bras,[1] Yejin Choi [1,2]
[1]Allen Institute for AI, WA, USA,
[2]Paul G. Allen School of Computer Science & Engineering, University of Washington, WA, USA
{nicholasl, ronanlb, yejinc}@allenai.org

Abstract

As AI systems become an increasing part of people's everyday lives, it becomes ever more important that they understand people's ethical norms. Motivated by *descriptive ethics*, a field of study that focuses on *people's descriptive* judgments rather than *theoretical prescriptions* on morality, we investigate a novel, data-driven approach to machine ethics.

We introduce SCRUPLES, the first large-scale dataset with 625,000 ethical judgments over 32,000 real-life anecdotes. Each anecdote recounts a complex ethical situation, often posing moral dilemmas, paired with a distribution of judgments contributed by the community members. Our dataset presents a major challenge to state-of-the-art neural language models, leaving significant room for improvement. However, when presented with simplified moral situations, the results are considerably more promising, suggesting that neural models can effectively learn simpler ethical building blocks.

A key take-away of our empirical analysis is that norms are not always clean-cut; many situations are naturally divisive. We present a new method to estimate the best possible performance on such tasks with inherently diverse label distributions, and explore likelihood functions that separate *intrinsic* from *model* uncertainty. Data and code are available at https://github.com/allenai/scruples.

`Title.` **Closing the door in a salespersons face?**
`Text.` The other day a salespersons knocked on our door and it was obvious he was about to sell something (insurance company uniform with flyers). I already have insurance that I'm happy with so I already knew I would've said no to any deals (although I would probably say no to any door to door salesperson)
I opened the door and before he could get a word out I just said "Sorry, not interested" and closed the door and went about my day.
My friend thinks that I was rude and should've let him at least introduce himself, whereas I feel like I saved us both time - he doesn't waste time trying to push a sale he won't get and I don't waste time listening to the pitch just to say "no thanks".
So, AITA?

`Type.` HISTORICAL
`Label.` OTHER
`Scores.` AUTHOR: 0, OTHER: 7, EVERYBODY: 0, NOBODY: 0, INFO: 0

Figure 1: An example from the dev set. Labels describe who the community views as in the wrong (i.e., the salesperson). Table 2 describes the labels in detail.

1 Introduction

State-of-the-art techniques excel at syntactic and semantic understanding of text, reaching or even exceeding human performance on major language understanding benchmarks (Devlin et al. 2019; Lan et al. 2019; Raffel et al. 2019). However, reading between the lines with pragmatic understanding of text still remains a major challenge, as it requires understanding social, cultural, and ethical implications. For example, given *"closing the door in a salesperson's face"* in Figure 1, readers can infer what is not said but implied, e.g., that perhaps the house call was unsolicited. When reading narratives, people read not just what is stated literally and explicitly, but also the rich non-literal implications based on social, cultural, and moral conventions.

Beyond narrative understanding, AI systems need to understand people's norms, especially ethical and moral norms, for safe and fair deployment in human-centric real-world applications. Past experiences with dialogue agents, for example, motivate the dire need to teach neural language models the ethical implications of language to avoid biased and unjust system output (Wolf, Miller, and Grodzinsky 2017; Schlesinger, O'Hara, and Taylor 2018).

However, machine ethics poses major open research challenges. Most notably, people must determine what norms to build into systems. Simultaneously, systems need the ability to anticipate and understand the norms of the different communities in which they operate. Our work focuses on the latter, drawing inspiration from *descriptive ethics*, the field of study that focuses on *people's descriptive* judgements, in contrast to *prescriptive ethics* which focuses on *theoretical prescriptions* on morality (Gert and Gert 2017).

As a first step toward computational models that predict communities' ethical judgments, we present a study based on people's diverse ethical judgements over a wide spectrum of social situations shared in an online community. Perhaps unsurprisingly, the analysis based on real world data quickly reveals that ethical judgments on complex real-life scenarios can often be divisive. To reflect this real-world challenge accurately, we propose predicting the *distribution* of normative judgments people make about real-life anecdotes. We formalize this new task as WHO'S IN THE WRONG? (WHO), predicting which person involved in the given anecdote would be considered in the wrong (i.e., breaking ethical norms) by a given community.

Ideally, not only should the model learn to predict clean-cut ethical judgments, it should also learn to predict if and when people's judgments will be divisive, as moral ambiguity is an important phenomenon in real-world communities. Recently, Pavlick and Kwiatkowski (2019) conducted an extensive study of annotations in natural language inference and concluded that diversity of opinion, previously dismissed as annotation "noise", is a fundamental aspect of the task which should be modeled to accomplish better language understanding. They recommend modeling the distribution of responses, as we do here, and found that existing models do not capture the kind of uncertainty expressed by human raters. Modeling the innate ambiguity in ethical judgments raises similar technical challenges compared to clean-cut categorization tasks. So, we investigate a modeling approach that can separate intrinsic and model uncertainty; and, we provide a new statistical technique for measuring the noise inherent in a dataset by estimating the best possible performance.

To facilitate progress on this task, we release a new challenge set, SCRUPLES:[1] a corpus of more than 32,000 real-life anecdotes about complex ethical situations, with 625,000 ethical judgments extracted from reddit.[2] The dataset proves extremely challenging for existing methods. Due to the difficulty of the task, we also release DILEMMAS: a resource of 10,000 actions with normative judgments crowd sourced from Mechanical Turk. Our results suggest that much of the difficulty in tackling SCRUPLES might have more to do with challenges in understanding the complex narratives than lack of learning basic ethical judgments.

Summarizing our main contributions, we:

- Define a novel task, WHO'S IN THE WRONG? (WHO).
- Release a large corpus of real-life anecdotes and norms extracted from an online community, reddit.
- Create a resource of action pairs with crowdsourced judgments comparing their ethical content.
- Present a new, general estimator for the best possible score given a metric on a dataset.[3]
- Study models' ability to predict ethical judgments, and assess alternative likelihoods that capture ambiguity.[4]

[1] Subreddit Corpus Requiring Understanding Principles in Life-like Ethical Situations

[2] https://reddit.com: A large internet forum.

[3] Try out the estimator at https://scoracle.apps.allenai.org.

[4] Demo models at https://norms.apps.allenai.org.

2 Datasets

SCRUPLES has two parts: the ANECDOTES collect 32,000 real-life anecdotes with normative judgments; while the DILEMMAS pose 10,000 simple, ethical dilemmas.

2.1 ANECDOTES

The ANECDOTES relate something the author either did or considers doing. By design, these anecdotes evoke norms and usually end by asking if the author was in the wrong. Figure 1 illustrates a typical example.

Each anecdote has three main parts: a *title*, body *text*, and label *scores*. Titles summarize the story, while the text fills in details. The scores tally how many people thought the participant broke a norm. Thus, after normalization the scores estimate the probability that a community member holds that opinion. Table 2 provides descriptions and frequencies for each label. Predicting the label distribution from the anecdote's title and text makes it an instance of the WHO task.

In addition, each story has a *type*, *action*, and *label*. Types relate if the event actually occurred (**HISTORICAL**) or only might (**HYPOTHETICAL**). Actions extract gerund phrases from the titles that describe what the author did. The label is the highest scoring class.

SCRUPLES offers 32,766 anecdotes totaling 13.5 million tokens. Their scores combine 626,714 ethical judgments, and 94.4% have associated actions. Table 1 expands on these statistics. Each anecdote exhibits high lexical diversity with words being used about twice per story. Moreover, most stories have enough annotations to get some insight into the distribution of ethical judgments, with the median being eight.

Source To study norms, we need representative source material: real-world anecdotes describing ethical situations with moral judgments gathered from a community. Due to reporting bias, fiction and non-fiction likely misrepresent the type of scenarios people encounter (Gordon and Van Durme 2013). Similarly, crowdsourcing often leaves annotation artifacts that make models brittle (Gururangan et al. 2018; Poliak et al. 2018; Tsuchiya 2018). Instead, SCRUPLES gathers community judgments on real-life anecdotes shared by people seeking others' opinions on whether they've broken a norm. In particular, we sourced the raw data from a subforum on `reddit`[5], where people relate personal experiences and then community members vote in the comments on who they think was in the wrong.[6] Each vote takes the form of an initialism: **YTA**, **NTA**, **ESH**, **NAH**, and **INFO**, which correspond to the classes, **AUTHOR**, **OTHER**, **EVERYONE**, **NO ONE**, and **MORE INFO**. Posters also title their anecdotes and label if it's something that happened, or something they might do. Since all submissions are unstructured text, users occasionally make errors when providing this information.

Extraction Each anecdote derives from a forum post and its comments. We obtained the raw data from the Pushshift Reddit Dataset (Baumgartner et al. 2020) and then used

[5] https://reddit.com/r/AmItheAsshole

[6] SCRUPLES v1.0 uses the data from 11/2018–4/2019.

	INSTANCES	ANNOTATIONS	ACTIONS	TOKENS	TYPES
train	27,766	517,042	26,217	11,424,463	59,605
dev	2,500	52,433	2,344	1,021,008	19,311
test	2,500	57,239	2,362	1,015,158	19,168
total	32,766	626,714	30,923	13,460,629	64,476

Table 1: Dataset statistics for the ANECDOTES. Tokens combine stories' titles and texts. Token types count distinct items.

CLASS	MEANING	FREQUENCY
AUTHOR	author is wrong	29.8%
OTHER	other is wrong	54.4%
EVERYBODY	everyone is wrong	4.8%
NOBODY	nobody is wrong	8.9%
INFO	need more info	2.1%

Table 2: Label descriptions and frequencies from dev. Frequencies tally individual judgments (not the majority vote).

	PRECISION	RECALL	F1	SPAM
Comment	0.99	0.95	0.97	20.5%
Post	1.00	0.99	0.99	56.2%

Table 3: Filtering metrics. Spam is the negative class. The accuracy on comments and posts is 95% and 99%.

rules-based filters to remove undesirable posts and comments (e.g. for being deleted, from a moderator, or too short). Further rules and regular expressions extracted the title, text, type, and action attributes from the post and the label and scores from the comments. To evaluate the extraction, we sampled and manually annotated 625 posts and 625 comments. Comments and posts were filtered with an F1 of 97% and 99%, while label extraction had an average F1 of 92% over the five classes. Tables 3 and 4 provide more detailed results from the evaluation.

Each anecdote's individual components are extracted as follows. The *title* is just the post's title. The *type* comes from a tag that the subreddit requires titles to begin with ("AITA", "WIBTA", or "META").[7] We translate AITA to **HISTORICAL**, WIBTA to **HYPOTHETICAL**, and discard META posts. A sequence of rules-based text normalizers, filters, and regexes extract the *action* from the title and transform it into a gerund phrases (e.g. "not offering to pick up my friend"). 94.4% of stories have successfully extracted actions. The *text* corresponds to the post's text; however, users can edit their posts in response to comments. To avoid data leakage, we fetch the original text from a bot that preserves it in the comments, and we discard posts when it cannot be found. Finally, the *scores* tally community members who expressed a given label. To improve relevance and independence, we only consider comments replying directly to the post (i.e., *top-level* comments). We extract labels using regexes to match variants of initialisms used on the site, and resolve multiple matches using rules.

[7]Respectively: "am I the a-hole", "would I be the a-hole", and "meta-post" (about the subreddit).

CLASS	PRECISION	RECALL	F1
AUTHOR	0.91	0.91	0.91
OTHER	0.99	0.94	0.96
EVERYONE	1.00	0.91	0.96
NO ONE	1.00	0.86	0.92
MORE INFO	0.93	0.78	0.85
HISTORICAL	1.00	1.00	1.00
HYPOTHETICAL	1.00	1.00	1.00

Table 4: Metrics for extracting labels from comments and post types from post titles.

`Action 1.` *telling a mom and Grandma to try to keep their toddler quiet in the library*
`Action 2.` *putting parsley on my roommates scrambled eggs*
`Label.` **ACTION 1**
`Scores.` **ACTION 1**: 5, **ACTION 2**: 0

Figure 2: A random example from the DILEMMAS (dev). Labels identify the action crowd workers saw as less ethical.

2.2 DILEMMAS

Beyond subjectivity (captured by the distributional labels), norms vary in importance: while it's good to say "thank you", it's imperative not to harm others. So, we provide the DILEMMAS: a resource for normatively ranking actions. Each instance pairs two actions from the ANECDOTES and identifies which one crowd workers found less ethical. See Figure 2 for an example. To enable transfer as well as other approaches using the DILEMMAS to solve the ANECDOTES, we aligned their train, dev, and test splits.

Construction. For each split, we made pairs by randomly matching the actions twice and discarding duplicates. Thus, each action can appear at most two times in DILEMMAS.

Annotation. We labeled each pair using 5 different annotators from Mechanical Turk. The dev and test sets have 5 extra annotations to estimate human performance and aid error analyses that correlate model and human error on dev. Before contributing to the dataset, workers were vetted with Multi-Annotator Competence Estimation (MACE)[8] (Hovy et al. 2013).[9] MACE assigns reliability scores to workers based on inter-annotator agreement. See Paun et al. (2018) for a recent comparison of different approaches.

[8]Code at https://github.com/dirkhovy/MACE
[9]Data used to qualify workers is provided as extra train.

3 Methodology

Ethics help people get along, yet people often hold different views. We found communal judgments on real-life anecdotes reflect this fact in that some situations are clean-cut, while others can be divisive. This inherent subjectivity in people's judgements (i.e., moral ambiguity) is an important facet of human intelligence, and it raises unique technical challenges compared to tasks that can be defined as clean-cut categorization, as many existing NLP tasks are often framed.

In particular, we identify and address two problems: estimating a performance target when human performance is imperfect to measure, and separating innate moral ambiguity from model uncertainty (i.e., a model can be *certain* about the inherent moral *ambiguity* people have for a given input).

3.1 Estimating the BEST Performance

For clean-cut categorization, human performance is easy to measure and serves as a target for models. In contrast, it's difficult to elicit distributional predictions from people, making human performance hard to measure for ethical judgments which include inherently divisive cases. One solution is to ensemble many people, but getting enough annotations can be prohibitively expensive. Instead, we compare to an oracle classifier and present a novel Bayesian estimator for its score, called the BEST performance,[10] available at https://scoracle.apps.allenai.org.

The Oracle Classifier To estimate the best possible performance, we must first define the oracle classifier. For clean-cut categorization, an oracle might get close to perfect performance; however, for tasks with innate variability in human judgments, such as the descriptive moral judgments we study here, it's unrealistic for the oracle to always guess the label a particular human annotator might have chosen. In other words, for our study, the oracle can at best know how people annotate the example on average.[11] Intuitively, this corresponds to ensembling infinite humans together.

Formally, for example i, if N_i is the number of annotators, Y_{ij} is the number of assignments to class j, p_{ij} is the probability that a random annotator labels it as class j, and $Y_{i:}$ and $p_{i:}$ are the corresponding vectors of class counts and probabilities, then the gold annotations are multinomial:

$$Y_{i:} \sim \mathrm{Multinomial}(p_{i:}, N_i)$$

The oracle knows the probabilities, but not the annotations. For cross entropy, the oracle gives $p_{i:}$ as its prediction, $\hat{p}_{i:}$:[12]

$$\hat{p}_{ij} := p_{ij}$$

We use this oracle for comparison on the evaluation data.

The BEST Performance Even if we do not know the oracle's predictions (i.e., each example's label distribution), we *can* estimate the oracle's performance on the test set. We present a method to estimate its performance from the gold annotations: the BEST performance.

[10]Bayesian Estimated Score Terminus

[11]This oracle is often called *the Bayes optimal classifier*.

[12]For hard-labels, we use the most likely class. This choice isn't optimal for all metrics but matches common practice.

SCENARIO	RELATIVE ERROR		
	ACCURACY	F1 (MACRO)	XENTROPY
`Anecdotes`	0.1%	0.6%	0.1%
`3 Annotators`	1.1%	3.1%	1.1%
`Mixed Prior`	1.1%	0.8%	0.4%

Table 5: BEST's relative error when estimating the oracle score in simulations. Anecdotes simulates the ANECDOTES, 3 Annotators simulates 3 annotators per example, and Mixed Prior simulates a Dirichlet mixture as the true prior.

Since $Y_{i:}$ is multinomial, we model $p_{i:}$ with the conjugate Dirichlet, following standard practice (Gelman et al. 2003):

$$\begin{aligned} p_{i:} &\sim \mathrm{Dirichlet}(\alpha) \\ Y_{i:} &\sim \mathrm{Multinomial}(p_{i:}, N_i) \end{aligned}$$

Using an empirical Bayesian approach (Murphy 2012), we fit the prior, α, via maximum likelihood, $\hat{\alpha}$, and estimate the oracle's loss, ℓ, as the expected value over the posterior:

$$s := \mathbb{E}_{p|Y,\hat{\alpha}}[\ell(p, Y)]$$

In particular, for cross entropy on soft labels:

$$s = \sum_i \mathbb{E}_{p_{i:}|Y_{i:},\hat{\alpha}} \left[\sum_j \frac{Y_{ij}}{N_i} \log p_{ij} \right]$$

Simulation Experiments To validate BEST, we ran three simulation studies comparing its estimate and the true oracle score. First, we simulated the ANECDOTES' label distribution using a Dirichlet prior learned from the data (`Anecdotes`). Second, we simulated each example having three annotations, to measure the estimator's usefulness in typical annotation setups (`3 Annotators`). Last, we simulated when the true prior is not a Dirichlet distribution but instead a mixture, to test the estimator's robustness (`Mixed Prior`). Table 5 reports relative estimation error in each scenario.

3.2 Separating Controversiality from Uncertainty

Most neural architectures confound model uncertainty with randomness intrinsic to the problem.[13] For example, softmax predicts a single probability for each class. Thus, 0.5 could mean a 50% chance that everyone picks the class, or a 100% chance that half of people pick the class. That singular number conflates model uncertainty with innate controversiality in people's judgements.

To separate the two, we modify the last layer. Instead of predicting probabilities with a softmax

$$\hat{p}_{ij} := \frac{e^{z_{ij}}}{\sum_k e^{z_{ik}}}$$

and using a categorical likelihood:

$$-\sum_i \sum_j Y_{ij} \log \hat{p}_{ij}$$

[13]Model uncertainty and intrinsic uncertainty are also often called *epistemic* and *aleatoric* uncertainty, respectively (Gal 2016).

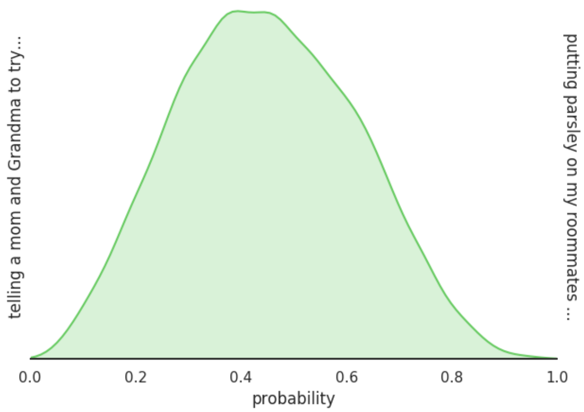

Figure 3: The model's distribution for the chance that someone judges action 2 as more unethical in Figure 2.

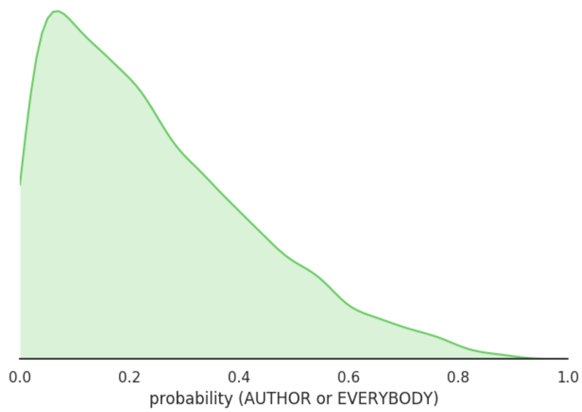

Figure 4: The model's distribution for the chance that someone judges the author as in the wrong in Figure 1.

We make activations positive with an exponential

$$\hat{\alpha}_{ij} \coloneqq e^{z_{ij}}$$

and use a Dirichlet-Multinomial likelihood:

$$-\sum_i \log \frac{\Gamma(N_i)\Gamma(\sum_j \hat{\alpha}_{ij})}{\Gamma(N_i + \sum_j \hat{\alpha}_{ij})} \prod_j \frac{\Gamma(Y_{ij} + \hat{\alpha}_{ij})}{\Gamma(Y_{ij})\Gamma(\hat{\alpha}_{ij})}$$

In practice, this modification requires two changes. First, labels must count the annotations for each class rather than take majority vote; and second, a one-line code change to replace the loss with the Dirichlet-Multinomial one.[14]

With the Dirichlet-Multinomial likelihood, predictions encode a distribution over class probabilities instead of singular point estimates. Figures 3 and 4 visualize examples from the DILEMMAS and ANECDOTES. Point estimates are recovered by taking the mean predicted class probabilities:

$$\mathbb{E}_{p_{ij}|\hat{\alpha}_{i:}}(p_{ij}) = \frac{\hat{\alpha}_{ij}}{\sum_j \hat{\alpha}_{ij}} = \frac{e^{z_{ij}}}{\sum_j e^{z_{ij}}}$$

Which is mathematically equivalent to a softmax. Thus, Dirichlet-multinomial layers generalize softmax layers.

3.3 Recommendations

Synthesizing results, we propose the following methodology for NLP tasks with labels that are naturally distributional:

Metrics Rather than evaluating hard predictions with metrics like F1, experiments can compare distributional predictions with metrics like total variation distance or cross-entropy, as in language generation. Unlike generation, classification examples often have multiple annotations and can report cross-entropy against soft gold labels.

Calibration Many models are poorly calibrated out-of-the-box, so we recommend calibrating model probabilities via temperature scaling before comparison (Guo et al. 2017).

[14]Our PyTorch implementation of this loss may be found at https://github.com/allenai/scruples.

ASPECT	CLEAN-CUT	AMBIGUOUS
labels	hard	soft or counts
prediction	point	distribution
last layer	softmax	Dirichlet-multinomial
metrics	accuracy, f1, etc.	xentropy, total variation distance, etc.
target score	human	BEST

Table 6: Comparison of clean-cut vs. ambiguous tasks.

Target Performance Human performance is a reasonable target on clean-cut tasks; however, it's difficult to elicit human judgements for distributional metrics. Section 3.1 both defines an oracle classifier whose performance provides the upper bound and presents a novel estimator for its score, the BEST performance. Models can target the BEST performance in clean-cut or ambiguous classification tasks; though, it's especially useful in ambiguous tasks, where human performance is misleadingly low.

Modeling Softmax layers provide no way for models to separate label controversiality from model uncertainty. Dirichlet-multinomial layers, described in Section 3.2, generalize the softmax and enable models to express uncertainty over class probabilities. This approach draws on the rich tradition of generalized linear models (McCullagh and Nelder 1989). Other methods to quantify model uncertainty exist as well (Gal 2016).

Table 6 summarizes these recommendations. Some recommendations (i.e., targeting the BEST score) could also be adopted by clean-cut tasks.

4 Experiments

To validate SCRUPLES, we explore two questions. First, we test for discernible biases with a battery of feature-agnostic and stylistic baselines, since models often use statistical cues

Baseline	F1 (macro)		Cross Entropy	
	dev	test	dev	test
`Prior`	0.164	0.161	1.609	1.609
`Sample`	0.197	0.191	NaN	NaN
`Style`	0.165	0.162	1.609	1.609
`BinaryNB`	0.168	0.168	1.609	1.609
`MultiNB`	0.202	0.192	1.609	1.609
`CompNB`	0.234	0.229	1.609	1.609
`Forest`	0.164	0.161	1.609	1.609
`Logistic`	0.192	0.192	1.609	1.609
`BERT`	0.218	0.216	1.081	1.086
`+ Dirichlet`	0.232	0.259	1.059	1.063
`RoBERTa`	0.278	**0.305**	1.043	1.046
`+ Dirichlet`	**0.296**	0.302	**1.027**	**1.030**
`Human`	0.468	0.490	–	–
`Best`	0.682	0.707	0.735	0.742

Table 7: Baselines for ANECDOTES. The best scores are in bold. Calibration smooths models worse than the uniform distribution to it, giving a cross-entropy of 1.609.

Baseline	F1 (macro)		Cross entropy	
	dev	test	dev	test
`Prior`	0.341	0.342	0.693	0.693
`Sample`	0.499	0.505	NaN	NaN
`Length`	0.511	0.483	NaN	NaN
`Style`	0.550	0.524	0.691	0.691
`Logistic`	0.650	0.643	0.657	0.660
`BERT`	0.728	0.720	0.604	0.606
`+ Dirichlet`	0.729	0.737	0.595	0.593
`RoBERTa`	0.757	0.746	0.578	0.577
`+ Dirichlet`	**0.760**	**0.783**	**0.570**	**0.566**
`Human`	0.807	0.804	–	–
`Best`	0.848	0.846	0.495	0.498

Table 8: Baselines for DILEMMAS. The best scores are bold.

to solve datasets without solving the task (Poliak et al. 2018; Tsuchiya 2018; Niven and Kao 2019). Second, we test if ethical understanding challenges current techniques.

4.1 Baselines

The following paragraphs describe the baselines at a high level.

Feature-agnostic Feature-agnostic baselines use only the label distribution, ignoring the features. `Prior` predicts the class probability for each label, and `Sample` assigns all probability to one class drawn from the label distribution.

Stylistic Stylistic baselines probe for stylistic artifacts that give answers away. `Style` applies a shallow classifier to a suite of stylometric features such as punctuation usage. For the classifier, the ANECDOTES use gradient boosted decision trees (Chen and Guestrin 2016), while the DILEMMAS use logistic regression. The `Length` baseline picks multiple choice answers based on their length.

Lexical and N-Gram These baselines apply classifiers to bag-of-n-grams features, assessing the ability of lexical knowledge to solve the tasks. `BinaryNB`, `MultiNB`, and `CompNB` (Rennie et al. 2003) apply Bernoulli, Multinomial, and Complement Naive Bayes, while `Logistic` and `Forest` apply logistic regression and random forests (Breiman 2001).

Deep Lastly, the deep baselines test how well existing methods solve SCRUPLES. `BERT` (Devlin et al. 2019) and `RoBERTa` (Liu et al. 2019) fine-tune powerful pretrained language models on the tasks. In addition, we try both `BERT` and `RoBERTa` with the Dirichlet-multinomial likelihood (`+ Dirichlet`) as described in Section 3.2.

4.2 Training and Hyper-parameter Tuning

All models were tuned with Bayesian optimization using scikit-optimize (Head et al. 2018).

Shallow models While the feature-agnostic models have no hyper-parameters, the other shallow models have parameters for feature-engineering, modeling, and optimization. These were tuned using 128 iterations of Gaussian process optimization with 8 points in a batch (Chevalier and Ginsbourger 2013), and evaluating each point via 4-fold cross validation. For the training and validation metrics, we used cross-entropy with hard labels. All shallow models are based on scikit-learn (Pedregosa et al. 2011) and trained on Google Cloud n1-standard-32 servers with 32 vCPUs and 120GB of memory. We tested these baselines by fitting them perfectly to an artificially easy, hand-crafted dataset. Shallow baselines for the ANECDOTES took 19.6 hours using 32 processes, while the the DILEMMAS took 1.4 hours.

Deep models Deep models' hyper-parameters were tuned using Gaussian process optimization, with 32 iterations and evaluating points one at a time. For the optimization target, we used cross-entropy with soft labels, calibrated via temperature scaling (Guo et al. 2017). The training loss depends on the particular model. Each model trained on a single Titan V GPU using gradient accumulation to handle larger batch sizes. The model implementations built on top of PyTorch (Paszke et al. 2017) and transformers (Wolf et al. 2019).

Calibration Most machine learning models are poorly calibrated out-of-the-box. Since cross-entropy is our main metric, we calibrated each model on dev via temperature scaling (Guo et al. 2017), to compare models on an even footing. All dev and test results report calibrated scores.

4.3 Results

Following our goal to model norms' distribution, we compare models with cross-entropy. RoBERTa with a Dirichlet likelihood (`RoBERTa + Dirichlet`) outperforms all other models on both the ANECDOTES and the DILEMMAS. One explanation is that unlike a traditional softmax layer trained on hard labels, the Dirichlet likelihood leverages all annotations without the need for a majority vote. Similarly, it can separate the controversiality of the question from the model's uncertainty, making the predictions more expressive (see Section 3.2). Tables 7 and 8 report the results. You can demo the model at https://norms.apps.allenai.org.

Label-only and stylistic baselines do poorly on both the DILEMMAS and ANECDOTES, scoring well below human and BEST performance. Shallow baselines also perform poorly on the ANECDOTES; however, the bag of n-grams logistic ranker (`Logistic`) learns some aspects of the DILEMMAS task. Differences between shallow models' performance on the ANECDOTES versus the DILEMMAS likely come from the role of lexical knowledge in each task. The ANECDOTES consists of complex anecdotes: participants take multiple actions with various contingencies to justify them. In contrast, the DILEMMAS are short with little narrative structure, so lexical knowledge can play a larger role.

5 Analysis

Diving deeper, we conduct two analyses: a controlled experiment comparing different likelihoods for distributional labels, and a lexical analysis exploring the DILEMMAS.

5.1 Comparing Different Likelihoods

Unlike typical setups, Dirichlet-multinomial layers use the full annotations, beyond just majority vote. This distinction should especially help more ambiguous tasks like ethical understanding. With this insight in mind, we explore other likelihoods leveraging this richer information and conduct a controlled experiment to test whether training on the full annotations outperforms majority vote.

In particular, we compare with cross-entropy on averaged labels (`Soft`) and label counts (`Counts`) (essentially treating each annotation as an example). Both capture response variability; though, `Counts` weighs heavily annotated examples higher. On the ANECDOTES, where some examples have thousands more annotations than others, this difference is substantial. For datasets like the DILEMMAS, with fixed annotations per example, the likelihoods are equivalent.

Tables 9 and 10 compare likelihoods on the ANECDOTES and DILEMMAS, respectively. Except for `Counts` on the ANECDOTES, likelihoods using all annotations consistently outperform majority vote training in terms of cross-entropy. Comparing `Counts` with `Soft` suggests that its poor performance may come from its uneven weighting of examples. `Dirichlet` and `Soft` perform comparably; though, `Soft` does better on the less informative, hard metric (F1). Like `Counts`, `Dirichlet` weighs heavily annotated examples higher; so, reweighting them more evenly may improve its score.

5.2 The Role of Lexical Knowledge

While the ANECDOTES have rich structure—with many actors under diverse conditions—the DILEMMAS are short and simple by design: each depicts one act with relevant context.

BASELINE	F1 (MACRO)	CROSS ENTROPY
`BERT`	0.218	1.081
`+ Soft`	0.212	1.053
`+ Counts`	0.235	1.074
`+ Dirichlet`	0.232	1.059
`RoBERTa`	0.278	1.043
`+ Soft`	**0.346**	**1.027**
`+ Counts`	0.239	1.045
`+ Dirichlet`	0.296	**1.027**

Table 9: Likelihood comparisons on the ANECDOTES (dev). Soft uses cross-entropy on soft labels, Counts uses cross-entropy on label counts, and Dirichlet uses a Dirichlet-multinomial layer. The best scores are in bold.

BASELINE	F1 (MACRO)	CROSS ENTROPY
`BERT`	0.728	0.604
`+ Soft`	0.725	0.594
`+ Counts`	0.728	0.598
`+ Dirichlet`	0.729	0.595
`RoBERTa`	0.757	0.578
`+ Soft`	**0.764**	**0.570**
`+ Counts`	0.763	**0.570**
`+ Dirichlet`	0.760	**0.570**

Table 10: Likelihood comparisons on the DILEMMAS (dev). Soft uses cross-entropy on soft labels, Counts uses cross-entropy on label counts, and Dirichlet uses a Dirichlet-multinomial layer. The best scores are in bold.

To sketch out the DILEMMAS' structure, we extracted each action's root verb with a dependency parser.[15] Overall, the training set contains 1520 unique verbs with "wanting" (14%), "telling" (7%), and "being" (5%) most common. To identify root verbs significantly associated with either class (more and less ethical), we ran a two-tailed permutation test with a Holm-Bonferroni correction for multiple testing (Holm 1979). For each word, the likelihood ratio of the classes served as the test statistic:

$$\frac{P(\text{word}|\text{less ethical})}{P(\text{word}|\text{more ethical})}$$

We tested association at the 0.05 level of significance using 100,000 samples in our Monte Carlo estimates for the permutation distribution. Table 11 presents the verbs most significantly associated with each class, ordered by likelihood ratio. While some evoke normative tones ("lying"), many do not ("causing"). The most common verb, "wanting", is neither positive nor negative; and, while it leans towards more ethical, this still happens less than 60% of the time. Thus, while strong verbs, like "ruining", may determine the label, in many cases additional context plays a major role.

To investigate the aspects of daily life addressed by the DILEMMAS, we extracted 5 topics from the actions via Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003), using

[15] `en_core_web_sm` from spaCy: https://spacy.io/.

VERB	LR	BETTER	WORSE	TOTAL
ordering	0.10	31	3	34
confronting	0.49	105	51	156
asking	0.56	1202	676	1878
trying	0.58	303	175	478
wanting	0.76	3762	2846	6608
ghosting	2.56	77	197	274
lying	2.75	48	132	180
visiting	2.91	22	64	86
ruining	5.11	18	92	110
causing	5.18	11	57	68

Table 11: Verbs significantly associated with more or less ethical choices from the DILEMMAS (train). LR is the likelihood ratio, Better and Worse count the times the verb was the better or worse action, and Total is their sum.

TOPIC	TOP WORDS
1	wanting, asking, family, dog, house
2	gf, parents, brother, breaking, girlfriend
3	telling, friend, wanting, taking, girlfriend
4	going, mother, friend, giving, making
5	friend, getting, girl, upset, ex

Table 12: Top 5 words for the DILEMMAS' topics (train), learned through LDA (Blei, Ng, and Jordan 2003).

the implementation in scikit-learn (Pedregosa et al. 2011). The main hyper-parameter was the number of topics, which we tuned manually on the DILEMMAS' dev set. Table 12 shows the top 5 words from each of the five topics. Interpersonal relationships feature heavily, whether familial or romantic. Less apparent from Table 12, other topics like retail and work interactions are also addressed.

6 Related Work

From science to science-fiction, people have long acknowledged the need to align AI with human interests. Early on, computing pioneer I.J. Good raised the possibility of an "intelligence explosion" and the great benefits, as well as dangers, it could pose (Good 1966). Many researchers have since cautioned about super-intelligence and the need for AI to understand ethics (Vinge 1993; Weld and Etzioni 1994; Yudkowsky 2008), with several groups proposing research priorities for building safe and friendly intelligent systems (Russell, Dewey, and Tegmark 2015; Amodei et al. 2016).

Beyond AI safety, *machine ethics* studies how machines can understand and implement ethical behavior (Waldrop 1987; Anderson and Anderson 2011). While many acknowledge the need for machine ethics, few existing systems understand human values, and the field remains fragmented and interdisciplinary. Nonetheless, researchers have proposed many promising approaches (Yu et al. 2018).

Efforts principally divide into top-down and bottom-up approaches (Wallach and Allen 2009). *Top-down* approaches have designers explicitly define ethical behavior. In contrast, *bottom-up* approaches learn morality from interactions or examples. Often, top-down approaches use symbolic methods such as logical AI (Bringsjord, Arkoudas, and Bello 2006), or preference learning and constraint programming (Rossi 2016; Rossi and Mattei 2019). Bottom-up approaches typically rely upon supervised learning, or reinforcement and inverse reinforcement learning (Abel, MacGlashan, and Littman 2016; Wu and Lin 2018; Balakrishnan et al. 2019). Beyond the top-down bottom-up distinction, approaches may also be divided into *descriptive* vs. *normative*. Komuda, Rzepka, and Araki (2013) compare the two and argue that descriptive approaches may hold more immediate practical value.

In NLP, fewer works address general ethical understanding, instead focusing on narrower domains like hate speech detection (Schmidt and Wiegand 2017) or fairness and bias (Bolukbasi et al. 2016). Still, some efforts tackle it more generally. One body of work draws on *moral foundations theory* (Haidt and Joseph 2004; Haidt 2012), a psychological theory explaining ethical differences in terms of how people weigh a small set of *moral foundations* (e.g. care/harm, fairness/cheating, etc.). Researchers have developed models to predict the foundations expressed by social media posts using lexicons (Araque, Gatti, and Kalimeri 2019), as well as to perform supervised moral sentiment analysis from annotated twitter data (Hoover et al. 2020).

Moving from theory-driven to data-driven approaches, other works found that word vectors and neural language representations encode commonsense notions of normative behavior (Jentzsch et al. 2019; Schramowski et al. 2019). Lastly, Frazier et al. (2020) utilize a long-running children's comic, *Goofus & Gallant*, to create a corpus of 1,387 correct and incorrect responses to various situations. They report models' abilities to classify the responses, and explore transfer to two other corpora they construct.

In contrast to prior work, we emphasize the task of building models that can predict the ethical reactions of their communities, applied to real-life scenarios.

7 Conclusion

We introduce a new task: WHO'S IN THE WRONG?, and a dataset, SCRUPLES, to study it. SCRUPLES provides simple ethical dilemmas that enable models to learn to reproduce basic ethical judgments as well as complex anecdotes that challenge existing models. With Dirichlet-multinomial layers fully utilizing all annotations, rather than just the majority vote, we're able to improve the performance of current techniques. Additionally, these layers separate model uncertainty from norms' controversiality. Finally, to provide a better target for models, we introduce a new, general estimator for the best score given a metric on a classification dataset. We call this value the BEST performance.

Normative understanding remains an important, unsolved problem in natural language processing and AI in general. We hope our datasets, modeling, and methodological contributions can serve as a jumping off point for future work.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable feedback. In addition, we thank Mark Neumann, Maxwell Forbes, Hannah Rashkin, Doug Downey, and Oren Etzioni for their helpful feedback and suggestions while we developed this work. This research was supported in part by NSF (IIS-1524371), the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1256082, DARPA CwC through ARO (W911NF15-1- 0543), DARPA MCS program through NIWC Pacific (N66001-19-2-4031), and the Allen Institute for AI.

Ethics Statement

Ethical understanding in NLP, and machine ethics more generally, are critical to the long-term success of beneficial AI. Our work encourages developing machines that anticipate how communities view something ethically, a major step forward for current practice. That said, one community's norms may be inappropriate when applied to another. We urge practitioners to consider the norms of their users' communities as well as the consequences and appropriateness of any model or dataset before deploying it. The code, models, and data in this work engage in an active area of research, and should not be deployed without careful evaluation. We hope our work contributes towards developing robust and reliable ethical understanding in machines.

References

Abel, D.; MacGlashan, J.; and Littman, M. L. 2016. Reinforcement learning as a framework for ethical decision making. In *Workshops at the thirtieth AAAI conference on artificial intelligence*.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* .

Anderson, M.; and Anderson, S. L., eds. 2011. *Machine Ethics*. Cambridge University Press. doi:10.1017/CBO9780511978036.

Araque, O.; Gatti, L.; and Kalimeri, K. 2019. MoralStrength: Exploiting a Moral Lexicon and Embedding Similarity for Moral Foundations Prediction. *Knowl. Based Syst.* 191: 105184.

Balakrishnan, A.; Bouneffouf, D.; Mattei, N.; and Rossi, F. 2019. Incorporating Behavioral Constraints in Online AI Systems. In *AAAI*.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media* 14(1): 830–839.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan): 993–1022.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, 4349–4357.

Breiman, L. 2001. Random Forests. *Machine Learning* 45: 5–32.

Bringsjord, S.; Arkoudas, K.; and Bello, P. 2006. Toward a General Logicist Methodology for Engineering Ethically Correct Robots. *IEEE Intelligent Systems* 21: 38–44.

Chen, T.; and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 785–794. New York, NY, USA: ACM. ISBN 978-1-4503-4232-2.

Chevalier, C.; and Ginsbourger, D. 2013. Fast Computation of the Multi-Points Expected Improvement with Applications in Batch Selection. In *Revised Selected Papers of the 7th International Conference on Learning and Intelligent Optimization - Volume 7997*, LION 7, 59–69. New York, NY, USA. ISBN 978-3-642-44972-7.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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*, 4171–4186. Association for Computational Linguistics.

Frazier, S.; Nahian, M. S. A.; Riedl, M. O.; and Harrison, B. 2020. Learning Norms from Stories: A Prior for Value Aligned Agents. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* .

Gal, Y. 2016. *Uncertainty in Deep Learning*. Ph.D. thesis, University of Cambridge.

Gelman, A.; Carlin, J.; Stern, H.; and Rubin, D. 2003. *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis. ISBN 9781420057294.

Gert, B.; and Gert, J. 2017. The Definition of Morality. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2017 edition.

Good, I. J. 1966. Speculations Concerning the First Ultraintelligent Machine. volume 6 of *Advances in Computers*, 31 – 88. Elsevier.

Gordon, J.; and Van Durme, B. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, 25–30. ACM.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1321–1330. JMLR.org.

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S.; and Smith, N. A. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, 107–112.

Haidt, J. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Pantheon Books.

Haidt, J.; and Joseph, C. 2004. Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues. *Daedalus* 133(4): 55–66. ISSN 00115266.

Head, T.; MechCoder; Louppe, G.; Shcherbatyi, I.; fcharras; Vinícius, Z.; cmmalone; Schröder, C.; nel215; Campos, N.; Young, T.; Cereda, S.; Fan, T.; rene rex; Shi, K. K.; Schwabedal, J.; carlosdanielcsantos; Hvass-Labs; Pak, M.; SoManyUsernamesTaken; Callaway, F.; Estève, L.; Besson, L.; Cherti, M.; Pfannschmidt, K.; Linzberger, F.; Cauet, C.; Gut, A.; Mueller, A.; and Fabisch, A. 2018. scikit-optimize/scikit-optimize: v0.5.2. doi:10.5281/zenodo.1207017.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* 65–70.

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. *Social Psychological and Personality Science* .

Hovy, D.; Berg-Kirkpatrick, T.; Vaswani, A.; and Hovy, E. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1120–1130.

Jentzsch, S.; Schramowski, P.; Rothkopf, C.; and Kersting, K. 2019. Semantics derived automatically from language corpora contain human-like moral choices. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 37–44.

Komuda, R.; Rzepka, R.; and Araki, K. 2013. Aristotelian Approach and Shallow Search Settings for Fast Ethical Judgment. *International Journal of Computational Linguistics Research* 4(1): 14–22.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *ArXiv* abs/1909.11942.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M. V.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L. S.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv* abs/1907.11692.

McCullagh, P.; and Nelder, J. 1989. *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis. ISBN 9780412317606.

Murphy, K. P. 2012. *Machine Learning: A Probabilistic Perspective*. The MIT Press. ISBN 0262018020.

Niven, T.; and Kao, H.-Y. 2019. Probing Neural Network Comprehension of Natural Language Arguments. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 4658–4664.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic Differentiation in PyTorch. In *NIPS Autodiff Workshop*.

Paun, S.; Carpenter, B.; Chamberlain, J.; Hovy, D.; Kruschwitz, U.; and Poesio, M. 2018. Comparing Bayesian Models of Annotation. *Transactions of the Association for Computational Linguistics* 6: 571–585.

Pavlick, E.; and Kwiatkowski, T. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics* 7(0): 677–694. ISSN 2307-387X.

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. *Journal of Machine Learning Research* 12: 2825–2830.

Poliak, A.; Naradowsky, J.; Haldar, A.; Rudinger, R.; and Van Durme, B. 2018. Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 180–191.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *ArXiv* abs/1910.10683.

Rennie, J. D. M.; Shih, L.; Teevan, J.; and Karger, D. R. 2003. Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, 616–623. AAAI Press.

Rossi, F. 2016. Moral Preferences. In *10th Workshop on Advances in Preference Handling (MPREF) at AAAI*.

Rossi, F.; and Mattei, N. 2019. Building ethically bounded AI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9785–9789.

Russell, S.; Dewey, D.; and Tegmark, M. 2015. Research priorities for robust and beneficial artificial intelligence. *Ai Magazine* 36(4): 105–114.

Schlesinger, A.; O'Hara, K. P.; and Taylor, A. S. 2018. Let's talk about race: Identity, chatbots, and AI. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 315. ACM.

Schmidt, A.; and Wiegand, M. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10.

Schramowski, P.; Turan, C.; Jentzsch, S.; Rothkopf, C. A.; and Kersting, K. 2019. BERT has a Moral Compass: Improvements of ethical and moral values of machines. *ArXiv* abs/1912.05238.

Tange, O. 2011. GNU Parallel - The Command-Line Power Tool. *;login: The USENIX Magazine* 36(1): 42–47. doi:10.5281/zenodo.16303. URL http://www.gnu.org/s/parallel.

Tsuchiya, M. 2018. Performance Impact Caused by Hidden Bias of Training Data for Recognizing Textual Entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Vinge, V. 1993. The Coming Technological Singularity. *Whole Earth Review* .

Waldrop, M. M. 1987. A Question of Responsibility. *AI Magazine* 8(1): 28. doi:10.1609/aimag.v8i1.572. URL https://www.aaai.org/ojs/index.php/aimagazine/article/view/572.

Wallach, W.; and Allen, C. 2009. Moral Machines: Teaching Robots Right from Wrong. Oxford University Press. doi:10.1093/acprof:oso/9780195374049.001.0001.

Weld, D.; and Etzioni, O. 1994. The First Law of Robotics (a Call to Arms). In *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence*, AAAI'94, 1042–1047. AAAI Press.

Wolf, M. J.; Miller, K.; and Grodzinsky, F. S. 2017. Why we should have seen that coming: comments on Microsoft's tay experiment, and wider implications. *ACM SIGCAS Computers and Society* 47(3): 54–64.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; and Brew, J. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* abs/1910.03771.

Wu, Y.-H.; and Lin, S.-D. 2018. A low-cost ethics shaping approach for designing reinforcement learning agents. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Yu, H.; Shen, Z.; Miao, C.; Leung, C.; Lesser, V. R.; and Yang, Q. 2018. Building Ethics into Artificial Intelligence. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 5527–5533.

Yudkowsky, E. 2008. Artificial intelligence as a positive and negative factor in global risk. *Global catastrophic risks* 1(303): 184.