Social Norms-Grounded Machine Ethics in Complex Narrative Situation

Tao Shen, Xiubo Geng, Daxin Jiang[*]
Microsoft Corporation
{shentao,xigeng,djiang}@microsoft.com

Abstract

Ethical judgment aims to determine if a person in a narrative situation acts under people's social norms under a culture, so it is crucial to understand actions in narratives and achieve machine ethics. Recent works depend on data-driven methods to directly judge the ethics of complex real-world narratives but face two major challenges. First, they cannot well handle dilemma situations due to a lack of basic knowledge about social norms. Second, they focus merely on sparse situation-level judgment regardless of the social norms involved during the judgment, leading to a black box. In this work, inspired by previous knowledge-grounded and -augmented paradigms, we propose to complement a complex situation with grounded social norms. Besides a norm-grounding knowledge model, we present a novel norm-supported ethical judgment model in line with neural module networks to alleviate dilemma situations and improve norm-level explainability. Empirically, our model improves state-of-the-art performance on two narrative judgment benchmarks.

1 Introduction

In natural language processing (NLP) literature, ethical judgment aims to determine if a person (e.g., narrator or someone else) in a narrative situation is morally wrong or correct (Lourie et al., 2021). For example, in a narrative situation "*I helped him but got taunted*", the narrator is morally good while the other is bad. It attracts more interest from academia and industry as it plays an indispensable role in human-centric applications and benefits a wide range of downstream tasks, e.g., dialogue systems and storytelling.

Recently, Forbes et al. (2020) propose a generative model, NORM TRANSFORMER, which can extract actions[1] from a narrative situation and then judge the ethics towards the actions. However, it can only generate action-level ethical judgment for simple narrative situations (i.e., sentences with limited events), so it usually fails to perform in complex narrative situations of many real-world applications. Here, "complex" is usually reflected in over-long narrative contexts (multiple paragraphs) and/or dilemma situations. Take Situation 1 in Figure 1 as a dilemma example: although we keep in mind that "*it's bad to punch others*", we cannot conclude the narrator is morally bad due to "*it's bad to steal*".

Situation 1: *I almost punch a friend who stole from me.*
Norm 1: *It's very bad to injure a person.*
Norm 2: *It's very bad to steal from others.*
Norm 3: *It's bad to betray a friend.*
Norm 4: *It's OK to want to take revenge.*
Situation 2: *When I call in a take out order and go pick it up myself, I don't tip.*
Norm 1: *It's good to always tip your server.*
Norm 2: *It's good to tip the driver when they pick you up.*
Norm 3: *It's bad to not want to leave a tip.*
Norm 4: *It's okay to expect a tip on takeout food.*

Figure 1: Two situations and the involved social norms. The first one is adapted from Forbes et al. (2020) while the second one is grounded by our model, where the norm in red is context-irrelevant to the situation.

On the contrary, Lourie et al. (2021) propose a data-driven method to directly judge complex narrative situations (e.g., real-life anecdotes from the Internet). Empowered by pre-trained language models (e.g., BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019)), the proposed method achieves satisfactory performance to boost its real-world applications but encounters two major challenges in the following.

First, complex situations often pose intricate storylines and character relationships, leading to more difficult moral *dilemmas* during ethical judgments. Again, it is difficult for machines to directly judge whether the narrator is morally wrong in Situation 1 of Figure 1 because it requires machines to im-

[*] Corresponding author.

[1] In this paper, "action" denotes a verb-centric "event" (Zhang et al., 2020) without subject, e.g., "*helped him*" and "*got taunted*".

Proceedings of the 29th International Conference on Computational Linguistics, pages 1333–1343
October 12–17, 2022.

ply multiple human-level social norms from the situation and understand their relations before making the final judgment. The social norms[2] can be regarded as unspoken commonsense rules about acceptable social behavior, which are crucial for an AI system to understand people's actions in narratives (Forbes et al., 2020). Second, human-curated labels of ethical judgment in complex situations are *sparse*: due to limited crowd-sourcing, the ethical judgment is labeled for a whole situation, i.e., at a very coarse level. As a result, a model learned on such sparse-labeled complex situations can only work as a black box to derive situation-level judgments, regardless of the involved social norms behind the judgments.

To overcome both the challenges, inspired by recent advances in knowledge-grounded/-augmented methods of open-domain (Wang et al., 2019) and commonsense (Lv et al., 2020) question answering (QA), we argue to complement complex narrative situations with action-level diverse social norms. Continue to take Situation 1 in Figure 1 as an example: although it is difficult to make a judgment based on the first two social norms, coupling with the other two diverse norms, "*It's bad to betray a friend*" and "*It's okay to want to take revenge*", can intuitively endorse a morally-okay judgment towards the narrator. Differing from previous knowledge-augmented methods that measure the relatedness of a query with grounded facts and then find an answer in the facts, our motivation is that the complementary social norms serve as supportive evidence to reduce moral dilemmas and promote norm-level explainability.

To this end, we propose a brand-new flexible ethical judgment framework with complementary social norms for complex narrative situations. First of all, to ground each event in a complex situation with diverse social norms and ensure grounding coverage given limited resources, we build a new norm-grounding knowledge model to generate social norms given a simple situation (e.g., a sentence in a complex situation) based on a pre-trained encoder-decoder backbone.

After grounding, we propose **N**orm-supported **E**thical Ju**d**gment (NEd) model in line with neural module networks (Liu et al., 2020) to complement a complex situation with grounded social norms. Specifically, built upon a pre-trained contextualizing encoder (e.g., RoBERTa), the model is composed of three neural modules: 1) *supportive alignment module* to softly and coarsely assign a sentence in the situation with its semantic-relevant norms, 2) *hierarchical integration module*, taking the alignment module's outputs as coarse evidence while operating at the token level, to enrich representations of events in the situation with those of social norms, and 3) *selective judgment module* to focus on key parts of the integration results and then make the final ethical judgment.

Our NEd model has certain merits: First, attributed to the alignment module, our model is robust to the errors (e.g., context-irrelevant/wrong norms, e.g., Norm 2 of Situation 2 in Figure 1) propagated from our norm-grounding knowledge model. Second, with a hierarchical (both norm-level and token-level) structure, our model can precisely enrich events in a situation with fine-grained ethical information, leading to superior judgment performance. Third, facilitated by intermediate outputs of the modules, our model is equipped with explainability in terms of human-understandable social norms. Lastly, our framework is general, flexible enough to various settings (see §4) of the ethical judgment and achieves new state-of-the-art performance on two benchmark datasets.

2 Related Work

In NLP, instead of theory-driven top-down approaches under prescriptive ethics (Bringsjord et al., 2006; Rossi and Mattei, 2019; Gert and Gert, 2020), recent works focus on data-driven bottom-up approaches with descriptive approaches (Balakrishnan et al., 2019; Wu and Lin, 2018) to achieve machine ethics (Rzepka and Araki, 2005; Anderson and Anderson, 2011). Ethical judgment, as an important task of machine ethics in NLP, is getting increased attention (Wolf et al., 2017; Schlesinger et al., 2018). Recent solutions (Lourie et al., 2021) depend on a data-driven paradigm but neglect the importance of the involved social norms during the judgment. But, how to explicitly integrate the social norms into ethical judgment in complex narratives is an open question.

Recently, several paradigms have been proposed to integrate additional knowledge, especially in open-domain and commonsense QA. Specifically, open-domain QA (Wang et al., 2019; Yang et al., 2019) retrieves related documents from large-scale

[2] "Norm" denotes to assign an ethical judgment to an "action", e.g., "*It's good to help others*" or ("*helping others*", good). It is a.k.a rules-of-thumb (RoT) in some literature (Forbes et al., 2020).

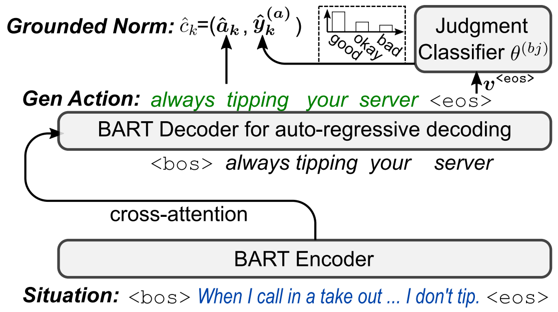

Grounded Norm: $\hat{c}_k$=($\hat{\boldsymbol{a}}_{\boldsymbol{k}}$, $\hat{\boldsymbol{y}}_{\boldsymbol{k}}^{(\boldsymbol{a})}$)
good
okay
bad
Judgment Classifier $\theta^{(bj)}$
$\boldsymbol{v}^{\texttt{<eos>}}$
Gen Action: always tipping your server <eos>
BART Decoder for auto-regressive decoding
<bos> always tipping your server
cross-attention
BART Encoder
Situation: <bos> When I call in a take out ... I don't tip. <eos>

Figure 2: Norm-grounding knowledge model.

corpus according to a query and then predict an answer based on the retrieved documents; commonsense QA (Lin et al., 2019; Lv et al., 2020) resorts to grounding on structured knowledge and then symbolically/latently derives the final answer based on the grounded facts. Though effective in their own fields, these integration paradigms are inapplicable here to fulfill our goal, i.e., enriching a complex situation with action-level, fine-grained social norms for less moral dilemmas and more human-level explainability.

3 Proposed Approach

This section begins with a general definition of ethical judgment task. Then, we present our norm-grounding knowledge model to ground events in a situation with diverse social norms (in §3.1 and illustrated in Figure 2). Next, given a grounded complex situation, we present our norm-supported ethical judgment (NEd) model as a flexible framework (in §3.2 and illustrated in Figure 3). Lastly, we detail training objectives to of our model in §3.3. This section begins with a general definition of ethical judgment task, followed by our norm-grounding knowledge model in §3.1 with Figure 2 and norm-supported ethical judgment (NEd) model in §3.2 with Figure 3.

Ethical Judgment. A complex narrative situation $\boldsymbol{S}$ consists of a sequence of sentences $\boldsymbol{S} = [\boldsymbol{s}_1, \dots, \boldsymbol{s}_m]$ where m denotes the number of sentences. Given a situation $\boldsymbol{S}$, *ethical judgment* aims to discriminate whether a person (e.g., the narrator or the other(s)) in $\boldsymbol{S}$ is morally wrong. Hence, ethical judgment is usually formulated as a classification problem, and the categories $\mathcal{Y}$ can be binary (i.e., {*good, bad*}), fine-grained (e.g., {*very good, good, okay, bad, very bad*}), etc.

3.1 Norm-Grounding Knowledge Model

Basically, we need to ground each event in a situation with diverse social norms. Traditional grounding methods (e.g., entity linking (Chen et al., 2020), event grounding (Du et al., 2021)) depend on lexical/semantic overlapping between a mention and entries in knowledge bases (e.g., corpus, graph). But, they are inapplicable to grounding an event with norms as events are expressed in free-form texts and knowledge base of norms is scarce, leading to low coverage and precision. In contrast, neural knowledge model (Bosselut et al., 2019) offers a novel solution: it learns from limited seed knowledge but leverage pre-trained language models to generalize more.

Following this line, we focus on crowd-sourced descriptions of norms and present a *norm-grounding knowledge model* as in Figure 2, to generate social norms given a simple situation. It is similar to GPT2-based NORM TRANSFORMER (Forbes et al., 2020) but differs in both target and base model.

First, we give a formal task definition to build a neural knowledge model for norm grounding. Here, we leverage training data from SOCIAL CHEMISTRY 101, which offers a sentence-level simple situation $\boldsymbol{s}$ in various scenarios and its corresponding diverse social norms $\mathcal{C} = \{c_1, \dots\}$, where each social norm c_k is composed of 1) an action $\boldsymbol{a}_k$ to describe one event in $\boldsymbol{s}$ and 2) its ethical judgment label $y_k^{(a)} \in \mathcal{Y}$, i.e., $c_k = (\boldsymbol{a}_k, y_k^{(a)})$.[3] Hence, the goal of norm grounding is to generate a set of diverse social norms $\mathcal{C}$ given a simple situation $\boldsymbol{s}$ (e.g., a sentence from a complex situation), i.e., $P(\mathcal{C}|s;\theta)$, which needs to cover all events in $\boldsymbol{s}$.

Then, we employ a pre-trained encoder-to-decoder model, BART (Lewis et al., 2020), as backbone to translate sentence-level situation $\boldsymbol{s}$ to an action $\boldsymbol{a}_k$ with its judgment $y_k^{(a)}$. That is, we define a conditional generation from $\boldsymbol{s}$ to $\boldsymbol{a}$, i.e.,

$$\hat{a}_t = \text{BART-EncDec}(\boldsymbol{s}, [\texttt{<bos>}, \boldsymbol{a}_{<t}]; \theta^{(bart)}),$$

where $\boldsymbol{s}$ is encoder input, $\boldsymbol{a} = [a_1, \dots]$ is the tokenized action $\boldsymbol{a}$, and $\hat{a}_t$ is the predicted token in t-th time step.

Next, to get an ethical judgment of the predicted action, we leverage the last states of the decoder

[3] Note 1) even a simple situation could contain multiple events, and 2) We take apart each *social norm* c_k with its *action* $\boldsymbol{a}_k$ and the corresponding *ethical judgment* $y_k^{(a)}$. This can highlight the semantics of the action of the norm and keep the judgment categorical.

(i.e., $\boldsymbol{h}^{(\texttt{<eos>})}$ – the embedding of end-of-sequence token <eos> in decoding):

$$\boldsymbol{y}^{(a)} = \text{softmax}(\text{MLP}(\boldsymbol{h}^{(\texttt{<eos>})}; \theta^{(bj)})), \quad (1)$$

where $\text{MLP}(\cdot; \theta^{(bj)})$ is $\theta^{(bj)}$-parameterized multilayer perceptron (MLP) to produce categorical distribution $\hat{\boldsymbol{y}}^{(a)} \in \mathbb{R}^{|\mathcal{Y}|}$ of ethical judgment towards the generated action $\hat{\boldsymbol{a}}$.

Training & Inference. We minimize an addition of negative log-likelihood of action generation and cross-entropy of ethical judgment. After trained, we use a sampling technique (Holtzman et al., 2020) to generate diverse norms: besides beam search (w/ size of 2), we use top-p=0.9 sampling during decoding and generate K social norms $\hat{\mathcal{C}} = \{(\hat{\boldsymbol{a}}_k, \hat{\boldsymbol{y}}_k^{(a)})_{k=1}^K\}$ in parallel for each situation $\boldsymbol{s}$ to ensure coverage/diversity.

3.2 Norm-Supported Ethical Judgment Model

After invoking norm-grounding knowledge model for each sentence $\boldsymbol{s}^j$ in a complex situation $\boldsymbol{S}$, we get social norms by

$$\hat{\mathcal{C}}^j = \{(\hat{\boldsymbol{a}}_k^j, \hat{\boldsymbol{y}}_k^{(a),j})_{k=1}^K\}, \forall j \in [1, m]. \quad (2)$$

Here, an open question remains about how to integrate these complementary norms into situation-level ethical judgment.

As an answer, we present a novel integration paradigm for ethical judgment, dubbed **N**orm-supported **E**thical ju**d**gment (NEd) model as in Figure 3. We introduce a concept of neural module network (Liu et al., 2020) because it can empower human-level explainability by visualizing intermediate outputs, consistent with our goal. It is noteworthy that, instead of considering variant combinations of neural modules as in (Andreas et al., 2016), we fix the neural architecture and focus more on the design of the modules as in (Liu et al., 2020).

First, we utilize a pre-trained Transformer encoder (e.g., RoBERTa) to embed a whole situation $\boldsymbol{S} = [\boldsymbol{s}^1, \ldots, \boldsymbol{s}^m]$ and the action $\hat{\boldsymbol{a}}_k^j$ of each generated social norm, i.e.,

$$\boldsymbol{U} = \text{Trans-Enc}([\boldsymbol{s}^1, \ldots, \boldsymbol{s}^m]; \theta^{(te)}), \quad (3)$$

$$\boldsymbol{V}_k^j = \text{Trans-Enc}(\hat{\boldsymbol{a}}_k^j; \theta^{(te)}), \quad (4)$$

where, $\forall j \in [1, m], k \in [1, K]$, and two encoders share parameters except for positional embeddings. $\boldsymbol{U}$ and $\boldsymbol{V}_k^j$ denote token-level representations.

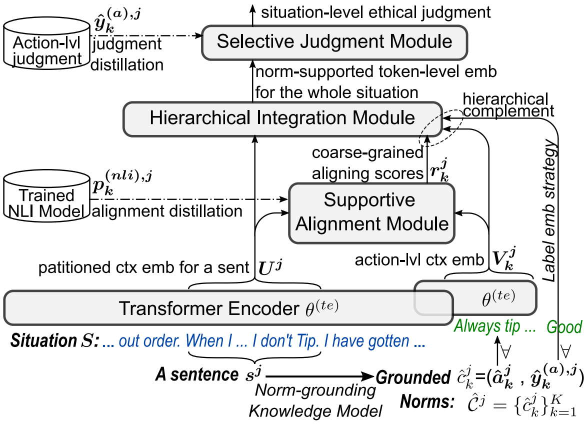

situation-level ethical judgment
Action-lvl judgment
$\hat{y}_k^{(a),j}$
judgment distillation
Selective Judgment Module
norm-supported token-level emb for the whole situation
hierarchical complement
Hierarchical Integration Module
coarse-grained aligning scores
r_k^j
Trained NLI Model
$p_k^{(nli),j}$
alignment distillation
Supportive Alignment Module
Label emb strategy
patitioned ctx emb for a sent
U^j
action-lvl ctx emb
V_k^j
Transformer Encoder $\theta^{(te)}$
$\theta^{(te)}$
Always tip ...
Good
Situation S: ... out order. When I ... I don't Tip. I have gotten ...
$\forall$
$\forall$
A sentence s^j
Norm-grounding Knowledge Model
Grounded $\hat{c}_k^j$=($\hat{a}_k^j$, $\hat{y}_k^{(a),j}$)
Norms: $\hat{\mathcal{C}}^j = \{\hat{c}_k^j\}_{k=1}^K$

Figure 3: **N**orm-supported **E**thical Ju**d**gment (NEd) model.

After situation-level long-term contextualizing, we partition $\boldsymbol{U}$ to sentence-level blocks to facilitate later integration:

$$\boldsymbol{U} = [\boldsymbol{U}^1, \ldots, \boldsymbol{U}^m], \forall \boldsymbol{U}^j \in \mathbb{R}^{d \times n^{(s)}}, \quad (5)$$

where $n^{(s)}$ is the number of tokens in a sentence. Built upon the above representations for 1) sentences in the situation and 2) actions of social norms, we propose three neural modules in the following to fulfill ethical judgment.

Supportive Alignment Module. It is crucial to measure if the action of a norm is not only consistent with at least one event in the sentence but also coherent to the context of the sentence. We first apply a mean-pooling to a sentence $\boldsymbol{U}^j$ and an action $\boldsymbol{V}_k^j$ to get one vector representation of each:

$$\boldsymbol{u}^j = \text{Pool}(\boldsymbol{U}^j) \text{ and } \boldsymbol{v}_k^j = \text{Pool}(\boldsymbol{V}_k^j), \quad (6)$$

where $\forall k \in [1, K]$. Then, following (Reimers and Gurevych, 2019), we represent their relationship by an interactive concatenation, i.e.,

$$\boldsymbol{o}_k^j = [\boldsymbol{u}^j; \boldsymbol{u}^j - \boldsymbol{v}_k^j; \boldsymbol{u}^j \odot \boldsymbol{v}_k^j; \boldsymbol{v}_k^j], \quad (7)$$

where $\forall k \in [1, K]$, $[\cdot;\cdot]$ denotes vector concatenation and $\odot$ denotes Hadamard product. Lastly, the relationship representation $\boldsymbol{o}_k^j$ is fed into an MLP with binary output, i.e.,,

$$\boldsymbol{r}_k^j = \text{softmax}(\text{MLP}(\boldsymbol{o}_k^j; \theta^{(al)})) \in \mathbb{R}^2, \quad (8)$$

where $(\boldsymbol{r}_k^j)_{[r=2]}$ denotes the relatedness intensity between sentence $\boldsymbol{s}^j \in \boldsymbol{S}$ and the action $\hat{\boldsymbol{a}}_k^j \in \mathcal{C}^j$. As a side benefit, such a module can also circumvent the errors propagated from norm-grounding knowledge model defined in §3.1.

Hierarchical Integration Module. After coarse alignment, we need to enrich each sentence in a complex situation with corresponding social norms to obtain diverse, supportive information. But we cannot integrate the norms in a straightforward manner (e.g., concatenation or addition) as a sentence contains multiple events, and we have no idea about which part is an event, not to mention how to align an event with actions. Hence, besides coarse norm-supportive alignment, we need to consider more fine-grained integration – operating at the token level and integrating in a sophisticated manner. Formally, we first equip a social norm's action $\boldsymbol{V}_k^j$ with its action-level judgment $\hat{\boldsymbol{y}}_k^{(a),j}$ to compose a complete representation of the norm (as mentioned in §3.1, we take each norm apart into action and judgment). That is

$$\tilde{\boldsymbol{V}}_k^j = \boldsymbol{V}_k^j + \boldsymbol{W}^{(jdg)} \hat{\boldsymbol{y}}_k^{(a),j}, \tag{9}$$

where $\boldsymbol{W}^{(jdg)} \in \mathbb{R}^{d \times |\mathcal{Y}|}$ denotes a weight matrix to identify judgment by following label embedding strategy (Wang et al., 2018), and the '$+$' here broadcasts along with sequence axis. Then, to achieve our hierarchical integration, we adapt one layer of the Transformer decoder by 1) we remove the self-attention but keep the cross-attention plus an MLP with residual connection and layer norm, 2) the cross-attention uses actions $\{\boldsymbol{V}_k^j\}_{k=1}^K$ as keys and their social norms $\{\tilde{\boldsymbol{V}}_k^j\}_{k=1}^K$ as values, and 3) we take the outputs $\{(\boldsymbol{r}_k^j)_{[r=2]}\}_{k=1}^K$ from the alignment module in Eq.(8) as norm-level gating values and apply them to cross-attention in a multiplicative manner. Thus, we define hierarchical integration operating on each s^j and its grounded norms $\hat{\mathcal{C}}^j$:

$$\bar{\boldsymbol{U}}^j = \sum\nolimits_k (\boldsymbol{r}_k^j)_{[r=2]} \tilde{\boldsymbol{V}}_k^j \text{softmax}((\boldsymbol{U}^j)^T \boldsymbol{V}_k^j / \sqrt{d})^T,$$
$$\tilde{\boldsymbol{U}}^j = \text{Layer-Norm}(\boldsymbol{U}^j + \bar{\boldsymbol{U}}^j; \theta^{(lm)}), \tag{10}$$

where $\forall j \in [1, m]$. Here for clear writing, we omit multi-head projections and an MLP after the attention, and please refer to (Vaswani et al., 2017) for their details. Next, given all enriched sentence representations $\{\tilde{\boldsymbol{U}}^j\}_{j=1}^m$ for sentences $[\boldsymbol{s}^1, \dots, \boldsymbol{s}^m]$, we re-unite them into token-level representations of the whole situation:

$$\tilde{\boldsymbol{U}} = [\tilde{\boldsymbol{U}}^1, \dots, \tilde{\boldsymbol{U}}^j], \tag{11}$$

where $[\cdot, \cdot]$ denotes concatenation along with sequence axis. Lastly, we apply one layer of Transformer encoder to $\tilde{\boldsymbol{U}}$ for long-term contextualized representations, i.e.,

$$\boldsymbol{E} = \text{Transformer-Layer}(\tilde{\boldsymbol{U}}; \theta^{(tl)}), \tag{12}$$

where $\boldsymbol{E}$ stands for the representations for all tokens in the situation $\boldsymbol{S}$, which have been integrated with precise, diverse, and supportive social norms in a hierarchical way.

Selective Judgment Module. Given $\boldsymbol{E}$, we first apply an attentive pooling (Liu et al., 2016; Lin et al., 2017), which aims at focusing on key parts of the integrated results, i.e.,

$$\boldsymbol{e} = \text{Attn-Pool}(\boldsymbol{E}) := \boldsymbol{E}\text{softmax}(\text{MLP}(\boldsymbol{E}; \theta^{(ap)})), \tag{13}$$

where $\boldsymbol{e} \in \mathbb{R}^d$, and $\text{MLP}(\cdot; \theta^{(ap)})$ is one-way out to represent the importance of each token. Lastly, we feed $\boldsymbol{e}$ into an MLP-based classifier, i.e.,

$$\hat{\boldsymbol{y}} = \text{softmax}(\text{MLP}(\boldsymbol{e}; \theta^{(cl)})),$$
$$\hat{y} = \arg\max \hat{\boldsymbol{y}}, \tag{14}$$

where $\hat{\boldsymbol{y}} \in \mathbb{R}^{|\mathcal{Y}|}$ is a categorical distribution over $\mathcal{Y}$, and $\hat{y}$ denotes the predicted judgment for the situation $\boldsymbol{S}$.

3.3 Training Objective

To train our proposed NEd model in an end-to-end fashion, we can define a cross-entropy loss for $\hat{\boldsymbol{y}}$ in Eq.(14), i.e.,

$$\mathcal{L}^{(main)} = -\sum\nolimits_{\boldsymbol{S} \in \mathcal{D}} \log \hat{\boldsymbol{y}}_{[y=y^*]}, \tag{15}$$

where $\hat{\boldsymbol{y}}_{[y=y^*]}$ denotes the probability of the gold label of $\boldsymbol{S}$ and $\mathcal{D}$ denotes training set. But there are many ethical judgment settings other than the simple classification. To exhibit our framework's flexibility to various settings, we will detail adapting procedure into two settings later in experiments. Besides the main loss $\mathcal{L}^{(main)}$, we design two distillation objectives to ensure they perform as expected. (i) $\mathcal{L}^{(ad)}$: *Alignment Distillation* aims at distilling semantic knowledge from a well-trained natural language inference (NLI) model to $\boldsymbol{r}_k^j$ (Eq.(8)) in the supportive alignment as a situation sentence is expected to entail the aligned social norms. (ii) $\mathcal{L}^{(jd)}$: *Judgment Distillation* aims at distilling action-level judgment knowledge $\hat{\boldsymbol{y}}_k^{(a),j}$ from our norm-grounding model into the *selective judgment module* by pooling each social norm's action plus the judge classifier, i.e., $\text{MLP}(\text{Pool}(\boldsymbol{V}_k^j); \theta^{(cl)})$.

Dataset	# Train	# Dev	# Test	# Tokens/Situation
ANECDOTES	27,766	2,500	2,500	410
DILEMMAS	23,596	2,340	2,360	10

Table 1: Statistics of two ethical judgment benchmarks.

Method (Macro-F1)	ANECDOTES		DILEMMAS	
	Dev	Test	Dev	Test
Prior	16.4	16.1	34.1	34.2
Sample	19.7	19.1	49.9	50.5
Style	16.5	16.2	55.0	52.4
BinaryNB	16.8	16.8	/	/
MultiNB	20.2	19.2	/	/
CompNB	23.4	22.9	/	/
Forest	16.4	16.1	/	/
Logistic	19.2	19.2	65.0	64.3
BERT$_{\text{large}}$	21.8	21.6	72.8	72.0
BERT$_{\text{large}}$ + Dirichlet	23.2	25.9	72.9	73.7
RoBERTa$_{\text{large}}$	27.8	30.5	75.7	74.6
RoBERTa$_{\text{large}}$ + Dirichlet	29.6	30.2	76.0	78.3
NEd-RoBERTa$_{\text{large}}$ **(ours)**	**41.20**	**37.32**	**76.91**	**78.59**
Human Performance	46.8	49.0	80.7	80.4

Table 2: Comparisons to state-of-the-art competitors on two benchmark datasets.

Method	**Bal-Acc**		**Macro-F1**	
	Dev	**Test**	**Dev**	**Test**
RoBERTa$_{\text{base}}$	26.62	28.14	27.84	29.59
RoBERTa$_{\text{base}}$ + Dirichlet	29.05	29.91	30.41	30.97
RoBERTa$_{\text{base}}$ + Soft	37.28	33.27	37.81	33.91
NEd-RoBERTa$_{\text{base}}$ **(ours)**	40.55	34.63	39.90	35.07
NEd-RoBERTa$_{\text{large}}$ **(ours)**	42.40	37.99	41.20	37.32

Table 3: Detailed comparisons on ANECDOTES.

Alignment Distillation. The *supportive alignment module* is designed to measure if the action of a social norm can provide supportive knowledge according to the context. This is similar to natural language inference (NLI) (Bowman et al., 2015) measuring if a premise is of *entailment*, *neutral*, or *contradiction* to a hypothesis, so it is intuitive to distill an NLI model to $\boldsymbol{r}_k^j$ in Eq.(8). Rather than direct distillation, it is noteworthy even if a norm's action contradicts the situation, the norm still can provide supports (e.g., in Figure 1, Situation 2 vs. its Norm 3). In formal, we first employ an NLI model trained on multi-genre natural language inference (Nangia et al., 2017) and pass a concatenation of $(\boldsymbol{s}^j, \hat{\boldsymbol{a}}_k^j)$ into the model to derive a three-categorical distribution $\bar{\boldsymbol{p}}_k^{(nli),j}$. Then, we merge *contradiction* and *entailment* to obtain a new distribution $\boldsymbol{p}_k^{(nli),j}$ over $\{$*neutral*, *non-neutral*$\}$. Thereby, we employ a Kullback–Leibler (KL) divergence between $\boldsymbol{r}_k^j$ and $\boldsymbol{p}_k^{(nli),j}$ as the training loss, i.e.,

$$\mathcal{L}^{(ad)} = \sum\nolimits_{\mathcal{D}} \sum\nolimits_{j=1}^{m} \sum\nolimits_{k=1}^{K} \text{KL-Div}(\boldsymbol{r}_k^j, \boldsymbol{p}_k^{(nli),j}).$$

Judgment Distillation. Since we have the contextual representations of each social norm's action as well as its action-level judgment, it is promising to distill such knowledge into the *selective judgment module*. To complete this, we apply a mean pooling to each $\boldsymbol{V}_k^j$ derived in Eq.(4) to get one vector representation of each action: $\boldsymbol{v}_k^j = \text{Pool}(\boldsymbol{V}_k^j)$, $\forall j \in [1, m], \forall k \in [1, K]$. The reason for not using the attentive pooling in Eq.(13) is that an action is only composed of a dozen of token so it is unnecessary for such a short sequence. Lastly, we use a KL divergence as the loss:

$$\mathcal{L}^{(jd)} = \sum\nolimits_{\mathcal{D},j,k} \text{KL-Div}(\text{softmax}(\text{MLP}(\boldsymbol{v}_k^j; \theta^{(cl)})), \hat{\boldsymbol{y}}_k^{(a),j}),$$

where $\theta^{(cl)}$ denotes the parameters defined in Eq.(14).

Overall Training Objective. Consequently, we can we can define the overall training objective as

$$\mathcal{L} = \mathcal{L}^{(main)} + g(\alpha) \cdot (\mathcal{L}^{(ad)} + \mathcal{L}^{(jd)}), \qquad (16)$$

where $g(\alpha)$ is exponential anneal of the weight.

4 Experiments

Datasets and Setting Adaptation. To exhibit our framework is flexible to various settings, we detail adaptations into two settings, corresponding to two datasets, i.e., ANECDOTES and DILEMMAS (Lourie et al., 2021). The statistics of the two benchmark datasets are listed in Table 1. For ANECDOTES, given a very complex narrative situation with multiple paragraphs, its goal is to discriminate $\{$*nobody wrong*, *narrator wrong*, *other wrong*, *everybody wrong*, or *more info needed*$\}$. We employ a binary judgment category and add another attentive pooling to the *selective judgment module*. We expect the two attention pooling mechanisms to focus on different parts of integrated results, which perform ethical judgments for "*narrator*" and "*other*" respectively. Besides, we add an MLP to calculate the probability of the result falling into *more info needed*. For DILEMMAS, given 2 situations (each w/ several events), its goal is to contrast their ethics and point out which one is more wrong. We use a binary judgment category and take the predicted probabilities of *bad* in $\hat{\boldsymbol{y}}$ (i.e., $\hat{\boldsymbol{y}}_{[y=\text{bad}]}$) as wrong scores.

Setups. To train our knowledge model, we use SOCIAL CHEMISTRY 101 (SC101) and employ BART$_{\text{large}}$. We optimize the model using mini-batch SGD with Adam optimizer, where learning rate is 10^{-5} with 5% warmup proportion, batch size is 16, the number of training epochs is 3. On the other hand, we initialize the backbone of our NEd model with either RoBERTa$_{\text{large}}$ or $_{\text{base}}$. Instead of hyperparameter tuning with Gaussian process optimization by Lourie et al. (2021), we set hyperparameters according to our experiences or early trials. We set K to 5/3 in two datasets. In Eq.(16), $\alpha \in [0, 1]$ is the ratio of training progress, and $g(x) = \exp(-\gamma \cdot x)$ is an exponential anneal where γ is a hype-parameter. We set γ to 10 to push the learning more incline to main loss. The base model is set with learning rate of 5×10^{-5} with 5% warmup, batch size of 32, number of epochs of 7. The large model is set with learning rate of 10^{-5} with 5% warmup, batch size of 16, number of epochs of 3. We run each model with 3 random seeds and evaluate on dev set every 500 steps during training; we report the best dev results as well as the corresponding test results. All experiments are run at one single Nvidia RTX6000 GPU. The codes are published at `https://github.com/taoshen58/NEd`.

Metrics. We use Macro-F1 (%) as our main metric to compare models. Compared to (Lourie et al., 2021), our work does not target ambiguity in ethics-related tasks and focus on making ethical judgment consistent with the majority. Thus, we rather use balanced-accuracy (Bal-Acc, %) as another metric and will also consider cross-entropy metric (Lourie et al., 2021) to verify our model's versatility.

4.1 Main Evaluations

Comparison with Competitors. In Table 2, we compare our large model, i.e., NEd-RoBERTa$_{\text{large}}$, with previous state-of-the-art competitors on ANECDOTES and DILEMMAS. BERT$_{\text{large}}$ and RoBERTa$_{\text{large}}$ denote fine-tuning the model with a classifying head. When coupled with "+ Dirichlet", they denote using a Dirichlet-multinomial layer to generalize $\mathrm{softmax}$ and enable models to express uncertainty over class probabilities (Lourie et al., 2021). It is observed, our model outperforms previous methods and improves the state-of-the-art performance by 5.1% and 0.3% on ANECDOTES and DILEMMAS, respectively. A potential reason for a marginal improvement on DILEMMAS is our performance has been too close to human performance (78.6% vs. 80.4%), making lifts difficult.

Method	Bal-Acc		Macro-F1	
	Dev	Test	Dev	Test
RoBERTa$_{\text{base}}$	72.82	72.47	72.80	72.41
RoBERTa$_{\text{base}}$ + Dirichlet	73.96	73.84	73.92	73.80
RoBERTa$_{\text{base}}$ + Soft	73.83	73.52	73.79	73.47
RoBERTa$_{\text{base}}$ + Norms	74.53	73.50	74.51	73.43
RoBERTa$_{\text{base}}$ + Filtered Norms	74.66	74.32	74.64	74.24
NEd-RoBERTa$_{\text{base}}$ **(ours)**	75.61	75.00	75.59	74.95
NEd-RoBERTa$_{\text{large}}$ **(ours)**	76.94	78.61	76.91	78.59

Table 4: Detailed comparisons with baselines on DILEMMAS.

Comparison with Baselines. As shown in Table 3 and 4, we also compare our model with various baselines. Here, we add a method named "RoBERTa$_{\text{base}}$ + Soft", which has already appeared in (Lourie et al., 2021) with only reported state-of-the-art dev results on ANECDOTES, as a strong baseline, and we re-implement it with a label-weighted loss to further boost its performance. As shown in Table 3, although "RoBERTa$_{\text{base}}$ + Soft" is significantly superior to "RoBERTa$_{\text{base}}$ + Dirichlet", it is still beaten by our NEd model. Similarly, as listed in Table 4, our model can surpass baselines on DILEMMAS by a large margin.

Comparison with Norm-augmented Methods. For more fair comparisons, we build two other RoBERTa-based baselines that also utilize our grounded social norms. Following common practice, we concatenate a situation $\boldsymbol{S}$ with its social norms, i.e., $[\boldsymbol{S}, \{\hat{\boldsymbol{a}}_k^j, \hat{y}_k^{(a),j}\}_{j,k}]$. As shown in the 2nd block of Table 4, "+ Norm" denotes directly concatenating all the grounded social norms while "+ Filtered Norms" denotes filtering out the norms with less relatedness to the situation, i.e., large neutral probability from NLI model > 0.2. These two methods are only applicable to DILEMMAS as they will lead to over-long ($\gg 512$) inputs on ANECDOTES. The table shows that our model still notably outperforms these two norm-augmented methods. Besides, the method with filtered norms performs better than direct concatenation, verifying the positive effects of the NLI model and the importance of our alignment distillation.

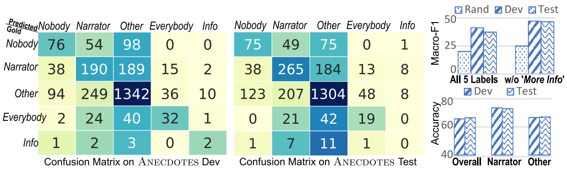
Performance on Controversiality. Lourie et al. (2021) use cross-entropy (Xentropy) as a metric to measure if a learned model can handle controversiality in judgments. Despite not being our

Method	Macro-F1		Xentropy	
	Dev	Test	Dev	Test
RoBERTa$_{\text{large}}$	75.7	74.6	0.578	0.577
RoBERTa$_{\text{large}}$ + Dirichlet	76.0	78.3	0.570	0.566
NEd-RoBERTa$_{\text{large}}$ **(ours)**	**76.91**	**78.59**	**0.5657**	**0.5652**

Table 5: Model comparison with the metric of cross-entropy.

Method	Macro-F1
NEd-RoBERTa$_{\text{large}}$ (full model)	**75.59**
◇ Removing *Alignment Module*	74.54
◇ Replacing *Alignment Module* with NLI Prior	75.15
◇ Removing Distillation Objectives	74.61
◇ Removing *Integration Module*	73.57
◇ Removing All Modules	72.80

Table 6: Ablation study on DILEMMAS Dev.



Gold \ Predicted	Nobody	Narrator	Other	Everybody	Info
Nobody	76	54	98	0	0
Narrator	38	190	189	15	2
Other	94	249	1342	36	10
Everybody	2	24	40	32	1
Info	1	2	3	0	2

Gold \ Predicted	Nobody	Narrator	Other	Everybody	Info
Nobody	75	49	75	0	1
Narrator	38	265	184	13	8
Other	123	207	1304	48	8
Everybody	0	21	42	19	0
Info	1	7	11	1	0

Figure 4: Insights into the performance on ANECDOTES.

Method	All (Ma-F1)	Easy	Hard
RoBERTa$_{\text{base}}$+Dirichlet	73.9	89.2	66.1
NEd-RoBERTa$_{\text{base}}$ (Δ)	75.6 (1.7)	90.05 (0.9)	68.15 (2.1)

Table 7: Evaluation on **easy** (cut-off examples) and **hard** (controversial dilemma examples). We split DILEMMAS Dev (**all**) into the two subsets according to if an example's consensus ratio > 0.8.

target, we report this metric to demonstrate our model's versatility. As in Table 5, all models include temperature calibration (Guo et al., 2017), and ours reaches competitive cross-entropy results compared to the specially designed Dirichlet layer.

Ablation Study. To check the contribution of each module, we conduct an ablation study in Table 6. Removing *Alignment Module*, equivalent to discarding coarse-grained integration, leads to a noticeable degeneration. And the degeneration will be alleviated when using prior coarse-grained weights (i.e., distillation targets from NLI) to replace the *Alignment Module*. Then, when removing the distillation objectives defined in Eq.(16), a notable performance decrease is observed, which verifies their importance. Next, we remove our *Integration Module*, degrading our model to RoBERTa plus our selective judgment module with distillation, resulting in a substantial decrease. Finally, we discard the only module, *selective judgment module*, from the last ablation, equal to RoBERTa baseline w/o all modules, leading to a further decrease.

4.2 Quantitative and Qualitative Analysis

Insights into ANECDOTES. Figure 2 shows a performance gap between dev and test on ANECDOTES. To dig this out, we illustrate their confusion matrices in Figure 4 (left & middle). It is shown that both recall and precision of label "*more info*" are 0 on the test, affecting macro metric. We further throw that label away, and the gap is largely narrowed, as in Figure 4 (upper-right). This exhibits that a distribution shift exists here and needs more efforts in the future. Besides the overall metric, our framework can derive the metrics w.r.t a specified person as in Figure 4 (bottom-right).

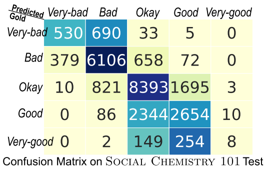
Handling Dilemmas Scenario. As shown in Table 7, we evaluate on two subsets of DILEMMAS Dev, which shows that our improvement on controversial dilemma examples is more significant than that on cut-off examples. These verify the superiority of our model in handling dilemmas scenarios.

Norm-Grounding Knowledge Model. We evaluate our knowledge model in Figure 5 (left), which shows low perplexity of action generation and high accuracy of action-level judgment on SC101 test. Although 5-categorical classification achieves 71%, as in Figure 5 (right) the misclassified examples mainly fall into its adjacent classes.

Evaluating Modules. In Figure 6, we give loss curves for Eq.(16): 1) the combined loss $\mathcal{L}$ is gradually close to the main loss $\mathcal{L}^{(main)}$ due to the exponential anneal, and 2) although learning rate of the two distillations are approaching to zero, their values do not increase significantly, verifying their objectives are consistent with our ethical judgment. Lastly, we test the performance of distilled modules in Figure 6 (right).

Case Study & Norm-level Explainability. As for Situation 1 in Figure 7, we show a case of ethical judgment in dilemma. It is observed our knowledge model precisely generates social norms consistent with the sentence and our hierarchical integration focuses on key norms to support the final ethical judgment. Meantime, the generated norms and alignment scores are human-understandable to intuitively explain the judgment from our model,

Metric	Dev	Test
Action PPL	6.80	6.94
Judge Accu	70.59	71.04



Gold \ Predicted	Very-bad	Bad	Okay	Good	Very-good
Very-bad	530	690	33	5	0
Bad	379	6106	658	72	0
Okay	10	821	8393	1695	3
Good	0	86	2344	2654	10
Very-good	0	2	149	254	8

Confusion Matrix on SOCIAL CHEMISTRY 101 Test

Figure 5: Evaluation of norm-grounding knowledge model.

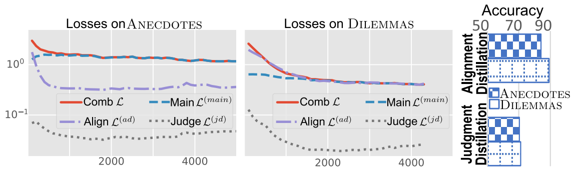


Figure 6: Loss descending curves with 0.99 moving average (left & middle) and module evaluation (right).

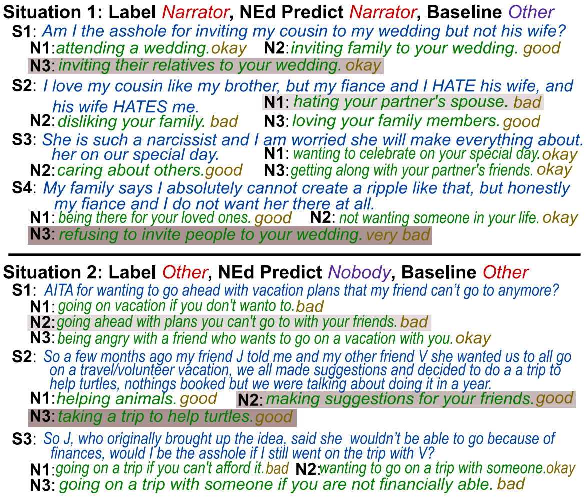
which notably improves norm-level explainability compared to the existing ethical judgment framework. In contrast, Situation 2 in Figure 7 shows an error case of our model. The possible reasons leading to the error are knowledge model incompetent in the long sentence (S2).

Diversity of Integration Scores. To investigate if the hierarchical (i.e., norm-level and token-level) attention scores are diverse to focus on different perspectives of social norms, we calculate the Pearson correlation coefficient between them, and the resulting is -0.081. And, we calculate the coefficient between $\boldsymbol{p}_{k\ [\text{neutral}]}^{(nli),j}$ from an NLI model and $\boldsymbol{r}_k^j$ from our alignment module, and the resulting is 0.565, verifying the learned module is not the same as the NLI model but learns latent alignments for our ethical judgment task.

Error Analysis & Limitations. We checked 20 error cases in ANECDOTES test and found that it is difficult for human beings to judge the ethics of complex situations even given the social norms, suggesting more explainable works in the future. Lastly, we also recognize model limitations: 1) although improving norm-level explainability, our pipeline inevitably leads to error propagation and affects the performance; and 2) our model must calculate deep contextualization for all social norms, resulting in large computation overheads.

5 Conclusion

In this paper, we propose an ethical judgment framework for complex narrative situations to reduce dilemma situations and improve norm-level explainability. These are achieved by our designed norm-grounding knowledge model and norm-supported ethical judgment (NEd) model. We conduct extensive experiments on two benchmark datasets to verify its superiority from both quantitative and qualitative perspectives.



Situation 1: Label *Narrator*, **NEd Predict** *Narrator*, **Baseline** *Other*

S1: *Am I the asshole for inviting my cousin to my wedding but not his wife?*
N1: *attending a wedding. okay* **N2**: *inviting family to your wedding. good*
N3: *inviting their relatives to your wedding. okay*

S2: *I love my cousin like my brother, but my fiance and I HATE his wife, and his wife HATES me.*
N1: *hating your partner's spouse. bad*
N2: *disliking your family. bad* **N3**: *loving your family members. good*

S3: *She is such a narcissist and I am worried she will make everything about. her on our special day.*
N1: *wanting to celebrate on your special day. okay*
N2: *caring about others. good* **N3**: *getting along with your partner's friends. okay*

S4: *My family says I absolutely cannot create a ripple like that, but honestly my fiance and I do not want her there at all.*
N1: *being there for your loved ones. good* **N2**: *not wanting someone in your life. okay*
N3: *refusing to invite people to your wedding. very bad*

Situation 2: Label *Other*, **NEd Predict** *Nobody*, **Baseline** *Other*

S1: *AITA for wanting to go ahead with vacation plans that my friend can't go to anymore?*
N1: *going on vacation if you don't wanto to. bad*
N2: *going ahead with plans you can't go to with your friends. bad*
N3: *being angry with a friend who wants to go on a vacation with you. okay*

S2: *So a few months ago my friend J told me and my other friend V she wanted us to all go on a travel/volunteer vacation, we all made suggestions and decided to do a a trip to help turtles, nothings booked but we were talking about doing it in a year.*
N1: *helping animals. good* **N2**: *making suggestions for your friends. good*
N3: *taking a trip to help turtles. good*

S3: *So J, who originally brought up the idea, said she wouldn't be able to go because of finances, would I be the asshole if I still went on the trip with V?*
N1: *going on a trip if you can't afford it. bad* **N2**: *wanting to go on a trip with someone. okay*
N3: *going on a trip with someone if you are not financially able. bad*

Figure 7: Case study of complex situations from ANECDOTES and grounded social norms. Texts with shadow denotes top-3 attended norms (N) and more dark denotes more intensive attention.

Ethics Statement. This work does not involve any sensitive data, but only public crowd-sourced corpora released in (Forbes et al., 2020; Lourie et al., 2021). Even the first situation and its social norms in Figure 1 (which may cause legal controversiality) are adapted from (Forbes et al., 2020). The resulting ethical judgment model can serve as a plug-in module to AI systems w/ language generation (e.g., dialogue system and chat-bot). First, it can filter unethical generated sentences. Second, it can perform ethical checks for massive user-posted and crowd-sourced data, thus reducing human-in-the-loop costs. Third, our model takes a step further to break down ethical judgments for norm-level transparency.

References

Michael Anderson and Susan Leigh Anderson. 2011. *Machine ethics*. Cambridge University Press.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 39–48. IEEE Computer Society.

Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. 2019. Incorporating behavioral constraints in online AI systems. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3–11. AAAI Press.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4762–4779. Association for Computational Linguistics.

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, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The Association for Computational Linguistics.

Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello. 2006. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intell. Syst.*, 21(4):38–44.

Shuang Chen, Jinpeng Wang, Feng Jiang, and Chin-Yew Lin. 2020. Improving entity linking by modeling latent entity type information. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7529–7537. AAAI Press.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2021. Excar: Event graph knowledge enhanced explainable causal reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2354–2363. 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 2020, Online, November 16-20, 2020*, pages 653–670. Association for Computational Linguistics.

Bernard Gert and Joshua Gert. 2020. The Definition of Morality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2020 edition. Metaphysics Research Lab, Stanford University.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. 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, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 2829–2839. Association for Computational Linguistics.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Jiangming Liu, Matt Gardner, Shay B. Cohen, and Mirella Lapata. 2020. Multi-step inference for reasoning over paragraphs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3040–3050. Association for Computational Linguistics.

Yang Liu, Chengjie Sun, Lei Lin, and Xiaolong Wang. 2016. Learning natural language inference using bidirectional LSTM model and inner-attention. *CoRR*, abs/1605.09090.

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. *CoRR*, 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 *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13470–13479. AAAI Press.

Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. 2020. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8449–8456. AAAI Press.

Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel R. Bowman. 2017. The repeval 2017 shared task: Multi-genre natural language inference with sentence representations. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP, RepEval@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, pages 1–10. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Francesca Rossi and Nicholas Mattei. 2019. Building ethically bounded AI. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 9785–9789. AAAI Press.

Rafal Rzepka and Kenji Araki. 2005. What statistics could do for ethics?: The idea of common sense processing based safety valve. In *AAAI Fall Symposium on Machine Ethics, Technical Report FS-05-06*, pages 85–87. AAAI.

Ari Schlesinger, Kenton P. O'Hara, and Alex S. Taylor. 2018. Let's talk about race: Identity, chatbots, and AI. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*, page 315. ACM.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. 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 5998–6008.

Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. 2018. Joint embedding of words and labels for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2321–2331. Association for Computational Linguistics.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage BERT: A globally normalized BERT model for open-domain question answering. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 5877–5881. Association for Computational Linguistics.

Marty J. Wolf, Keith W. Miller, and Frances S. Grodzinsky. 2017. Why we should have seen that coming: comments on microsoft's tay "experiment, " and wider implications. *SIGCAS Comput. Soc.*, 47(3):54–64.

Yueh-Hua Wu and Shou-De Lin. 2018. A low-cost ethics shaping approach for designing reinforcement learning agents. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1687–1694. AAAI Press.

Wei Yang, Yuqing Xie, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. Data augmentation for BERT fine-tuning in open-domain question answering. *CoRR*, abs/1904.06652.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. ASER: A large-scale eventuality knowledge graph. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 201–211. ACM / IW3C2.