

# SOCIAL CHEMISTRY 101: Learning to Reason about Social and Moral Norms

Maxwell Forbes<sup>†‡</sup> Jena D. Hwang<sup>‡</sup> Vered Shwartz<sup>†‡</sup> Maarten Sap<sup>†</sup> Yejin Choi<sup>†‡</sup>

<sup>†</sup>Paul G. Allen School of Computer Science & Engineering, University of Washington

<sup>‡</sup>Allen Institute for AI

{mbforbes, msap, yejin}@cs.washington.edu, {jenah, vereds}@allenai.org

[maxwellforbes.com/social-chemistry](https://maxwellforbes.com/social-chemistry)

## Abstract

Social norms—the unspoken commonsense rules about acceptable social behavior—are crucial in understanding the underlying causes and intents of people’s actions in narratives. For example, underlying an action such as “*wanting to call cops on my neighbor*” are social norms that inform our conduct, such as “*It is expected that you report crimes.*”

We present SOCIAL CHEMISTRY, a new conceptual formalism to study people’s everyday social norms and moral judgments over a rich spectrum of real life situations described in natural language. We introduce SOCIAL-CHEM-101, a large-scale corpus that catalogs 292k **rules-of-thumb** such as “*It is rude to run a blender at 5am*” as the basic conceptual units. Each rule-of-thumb is further broken down with 12 different dimensions of people’s judgments, including social judgments of good and bad, moral foundations, expected cultural pressure, and assumed legality, which together amount to over 4.5 million annotations of categorical labels and free-text descriptions.

Comprehensive empirical results based on state-of-the-art neural models demonstrate that computational modeling of social norms is a promising research direction. Our model framework, NEURAL NORM TRANSFORMER, learns and generalizes SOCIAL-CHEM-101 to successfully reason about previously unseen situations, generating relevant (and potentially novel) attribute-aware social rules-of-thumb.

## 1 Introduction

Understanding and reasoning about social situations relies on unspoken commonsense rules about *social norms*, i.e., acceptable social behavior (Haidt, 2012). For example, when faced with situations like “*wanting to call the cops on my neighbors*,” (Figure 1), we perform a rich variety of reasoning about about legality, cultural pressure,



Figure 1: This figure illustrates an intuitive subset of our formalism to reason about social norms in language. Our approach centers around Rules-of-Thumb (RoTs; text in colored tubes), which describe social expectations given a situation (text in the center hexagon). Rather than prescribing what is right or wrong, RoTs reveal ethical judgments about social propriety from varying perspectives.<sup>1</sup> Structured categorical (in smaller hexagons; e.g., “social judgment” and “cultural pressure”) annotations provide richer understanding. All RoTs shown here in tubes are generated by our NEURAL NORM TRANSFORMER conditioning on the center situation and the categorical types.

and even morality of the situation (here, “*reporting a crime*” and “*being friends with your neighbor*” are conflicting norms). Failure to account

<sup>1</sup>Note that the social identities of the participants of situations would further inform which social norms are most

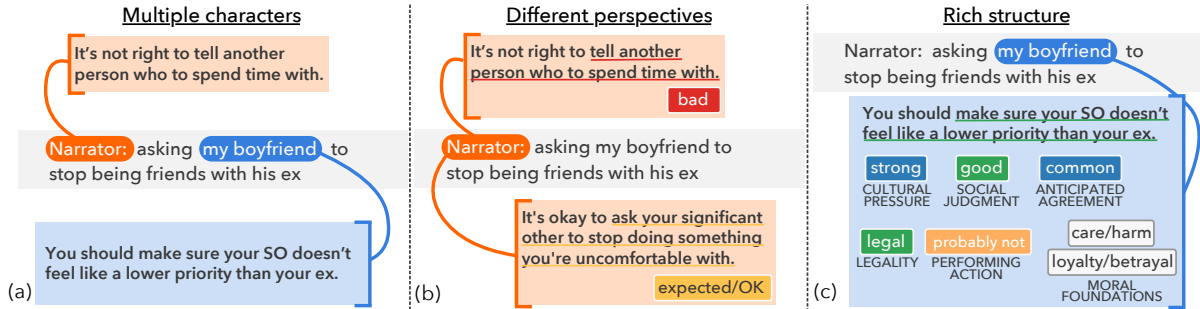


Figure 2: Three different slices of a complete annotation for a single situation, meant to illustrate our approach. Each **RoT** (text in colored boxes, e.g., “It’s not right to tell...”) is written for a particular real life **situation** (text in pale grey boxes, e.g., “asking my boyfriend to stop being ...”) and a specific **person** in that situation (“narrator” vs “my boyfriend”). (a) A situation often includes multiple people with distinct perspectives, evoking different (and possibly conflicting) RoTs. (b) Even a single person may have multiple, conflicting RoTs—key ingredients for moral dilemmas. (c) Each RoT is further broken down with categorical and free text annotations (shown in tiny colored buttons. e.g., “strong” for *cultural pressure*). The full definition of the low-level RoT attributes are in Figure 4.

for social norms could significantly hinder AI systems’ ability to interact with humans (Pereira et al., 2016).

In this paper, we introduce SOCIAL CHEMISTRY as a new formalism to study people’s social and moral norms over everyday real life situations. Our approach based on crowdsourced descriptions of norms is inspired in part by studies in *descriptive* or *applied* ethics (Hare et al., 1981; Kohlberg, 1976), which takes a *bottom-up* approach by asking people’s judgements on various ethical situations. This is in contrast to the *top-down* approach taken by *normative* or *prescriptive* ethics to prescribe the key elements of ethical judgements. The underlying motivation of our study is that we, the NLP field, might have a real chance to contribute to the studies of computational social norms and descriptive ethics through large-scale crowdsourced annotation efforts combined with state-of-the-art neural language models.

To that end, we organize *descriptive* norms via free-text *rules-of-thumb* (RoTs) as the basic conceptual units.

**Rule-of-Thumb (RoT)** — A descriptive cultural norm structured as the judgment of an action. For example, “It’s rude to run the blender at 5am.”

Each RoT is further broken down with 12 theoretically-motivated dimensions of people’s judgments such as social judgments of good and bad, theoretical categories of moral foundations,

relevant. For example, if the neighbors are African American, it might be worse to call the cops due to racial profiling (Eberhardt, 2020).

expected cultural pressure, and assumed legality. All together, these annotations comprise SOCIAL-CHEM-101, a new type of NLP resource that catalogs 292k RoTs over 104k real life situations, along with 365k sets of structural annotations, which break each RoT into 12 dimensions of norm attributes. Together, this amounts to over 4.5M categorical and free-text annotations.

We investigate how state-of-the-art neural language models can learn and generalize out of SOCIAL-CHEM-101 to accurately reason about social norms with respect to a previously unseen situation. We term this modeling framework NEURAL NORM TRANSFORMER, and find it is able to generate relevant (and potentially novel) rules-of-thumb conditioned on all attribute dimensions. Even so, this breadth of this task proves challenging to current neural models, with humans rating model’s adherence to different attributes from 0.28 to 0.91 micro-F1.

In addition, we showcase a potential practical use case of computational social norms by analyzing political news headlines through the lens of our framework. We find that our empirical results align with the *Moral Foundation Theory* of Graham et al. (2009); Haidt (2012) on how the moral norms of different communities vary depending on their political leanings and news reliability. Our empirical studies demonstrate that computational modeling of social norms is a feasible and promising research direction that warrants further investigation. SOCIAL-CHEM-101 provides a new resource to teach AI models to learn people’s norms, as well

as to support novel interdisciplinary research across NLP, computational norms, and descriptive ethics.

## 2 Approach

The study of social norms have roots in descriptive ethics and moral psychology. They tell us that social norms are culturally-sensitive standards of appropriate conduct. Alongside explicit laws and regulations that govern our society, social norms perform the role of providing guidelines on socially appropriate behaviors (Elster, 2006; Bowdery, 1941; Kohlberg, 1976) and are responsible for setting implicit expectations of what is socially right or wrong (Malle et al., 2014; Haidt, 2012; Hare et al., 1981). They influence a wide-range of social functions such as preserving biological needs to survival (e.g., refraining from harming or killing), maintaining social civility and order (e.g., maintaining politeness, recognizing personal space), and providing identity and belonging to a community (e.g., respecting the elderly). In turn, these social norms influence how we judge, communicate, and interact with each other.

**RoTs** Our aim is then to forefront these implicit expectations about social norms via RoTs. We formalize the definition of RoTs as situationally-relevant evaluative judgments of social norm, and posit that for any given **situation**, one or more RoTs will be evoked in the minds of the interpreter. Consider the following situation and its RoT.

### Punching someone.

RoT: It is unacceptable to injure a person.

Most readers can instantly recognize the situation is in violation of an unspoken social rule: “*Do not injure others.*” This rule is responsible for the series of natural questions that probe at the morality of the action, like “why did the narrator punch someone?” “was the action justified?” and “do I want to sympathize with the narrator?” The role of the RoT is to identify the unspoken rule in the situation by specifying the behavior or **action** (“injuring a person”) and its **acceptability judgment** (“it is unacceptable”). More complex situations can be associated with multiple RoTs, as seen in the example below:

### Punching a friend who stole from me.

- RoT 1: It is unacceptable to injure a person.
- RoT 2: People should not steal from others.
- RoT 3: It is bad to betray a friend.
- RoT 4: It is OK to want to take revenge.

The RoTs represent a variety of social norms that elaborate on various perspectives available in the

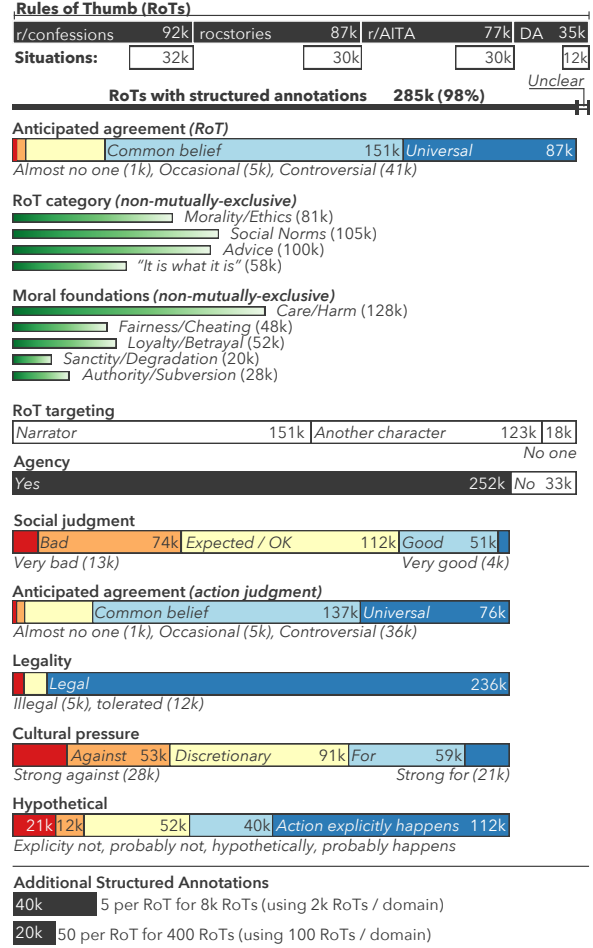


Figure 3: SOCIAL-CHEM-101 Dataset statistics. Bars are drawn to scale. Individual values for all of the different attributes are also given in Figure 4.

situation: RoTs about stealing (RoT 1) vs. punching (RoT 2), RoTs targeting the different characters in the situation (RoTs 1, 4 target the narrator; RoTs 2, 3 target narrator’s friend), and RoTs that elaborate on additional social interpretation implicit in the situation (RoT 3: theft from a friend is cast as an act of betrayal). Effectively, RoTs represent evaluative judgments about a social situation in light of unspoken but accepted social norms.<sup>2</sup> Figure 2 shows three subsets of a situation’s annotation to illustrate the perspectives RoTs capture.

**Cultural Scope of this study** We recognize that social norms are often culturally sensitive (Haidt et al., 1993; Kagan, 1984) and judgments of morality and ethics concerning individuality, community and society do not always hold universally (Shweder, 1990). While some situations (e.g.,

<sup>2</sup>Our definition of RoTs corresponds to the first of the two evaluative moral judgments defined in Malle et al. (2014).

“punching someone”) might have similar levels of acceptability across a number of cultures, others might have drastically varied levels depending on the culture of its participants (e.g., “kissing someone on the cheek as a greeting”). As a starting point, our study focuses on the socio-normative judgments of English-speaking cultures represented within North America. While we find some variation of judgments in our annotations (e.g., with respect to certain worker characteristics, see §A.6), extending this formalism to other countries and non-English speaking cultures remains a compelling area of future research.

### 3 SOCIAL-CHEM-101 Dataset

We obtained 104k source situations from 4 text domains (§3.1), for which we elicited 292k RoTs from crowd workers (§3.2). We then define a structured annotation task where workers isolate the central action described by the RoT and provide a series of judgments about the RoT and the action (§3.3). In total, we collect 365k structured annotations, performing multiple annotations per RoT for a subset of the RoTs to study the variance in annotations. Figure 3 illustrates our dataset statistics.

#### 3.1 Situations

We use a *situation* to denote the one-sentence prompt given to a worker as the basis for writing RoTs. We gather a total of 104k real life situations from four domains: scraped titles of posts in the subreddits *r/confessions* (32k) and *r/amitheasshole* (*r/AITA*, 30k), which largely focus on moral quandaries and interpersonal conflicts; 30k sentences from the ROCStories corpus (*rocstories*, Mostafazadeh et al., 2016); and scraped titles from the Dear Abby advice column archives<sup>3</sup> (*dearabby*, 12k).<sup>4</sup>

#### 3.2 Rules-of-Thumb (RoTs)

To collect RoTs, we provide workers with a situation as a prompt and them to write 1 – 5 RoTs inspired by that situation. From the 104k situations, we elicit a total of 292k RoTs. Despite RoTs averaging just 10 words, we observe that 260k/292k RoTs are unique across the dataset.

For the development of RoTs, we instruct the workers to produce RoTs that *explain the basics*

<sup>3</sup><https://www.uexpress.com/dearabby/archives>

<sup>4</sup>See Appendix A.1 for further data preprocessing details.

Figure 4: All attribute values for structured RoT annotations, with one complete example annotation filled in.

of social norms, just as one would instruct a five-year-old child on the ABCs of acceptable conduct. RoTs are to be:

1. **inspired by the situation**, to maintain a lower bound on relevance;
2. **self-contained**, to be understandable without additional explanation; and
3. structured as **judgment** of acceptability (e.g., good/bad, (un)acceptable, okay) and an **action** that is assessed.

In order to encourage RoT diversity, we also ask that an RoT should counterbalance *vagueness* against *specificity* so that RoTs generalize across multiple situations (e.g., “It is rude be selfish.”) without being too specific (e.g., “It is rude not to share your mac’n’cheese with your younger brother.”). We also ask workers to write RoTs illustrating *distinct ideas* and *avoid trivial inversions*



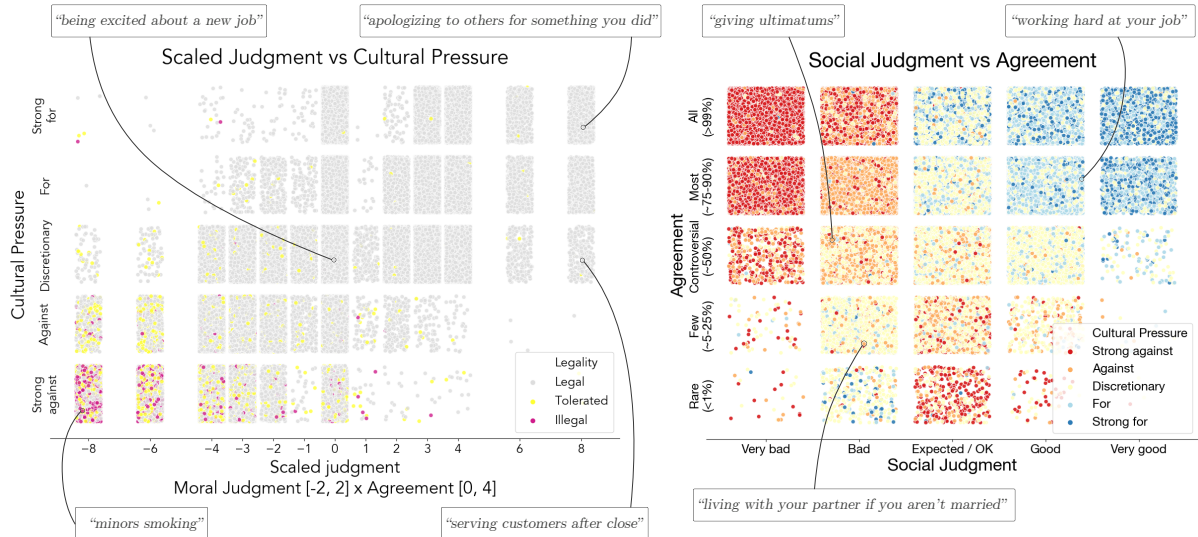


Figure 5: Plotting the distribution of RoTs in SOCIAL-CHEM-101 along axes of *moral judgment*, *agreement*, *cultural pressure*, and *legality*. **Left:** Moral judgment is scaled with agreement (how commonly held the belief is) and plotted against cultural pressure. Illegal activities fall in the bottom left: actions that are universally understood to be wrong and people feel negative cultural pressure for. **Right:** Moral judgment is plotted against agreement. Discretionary actions span a range of moral values (yellow ranging horizontally) and fringe beliefs often evoke strong negative cultural pressure even when morally neutral (bottom of plot).

to prevent low-information RoTs that rephrase the same idea or simply invert the judgement and action.

**Character Identification.** We ask workers to identify phrases in each situation that refer to people. For example, in a situation, like “*My brother chased after the Uber driver,*” workers mark the underlined spans. We collect three workers’ spans, calling each span a *character*. All characters identified become candidates for grounding RoTs and actions in the structured annotation. As such, we optimize for recall instead of precision by using the largest set of characters identified by any worker. We also include a *narrator* character by default.

### 3.3 RoT Breakdowns

We perform a structured annotation, which we term a *breakdown*, on each RoT. In an RoT breakdown, a worker isolates the underlying action contained in the RoT. Then, they assign a series of categorical attributes to both the RoT and the action. These categorical annotations allow for additional analyses and experiments relative to the text-only RoTs.

The attributes fall into two categories corresponding to the central annotation goals. The first goal is to tightly *ground* RoTs to their respective situations. The second goal is to partition *social* expectations using theoretically motivated categories.

A subset of the attributes are labeled on the RoT (e.g., “*It is expected that you report a crime*”), while others are on the action (e.g., “*reporting a crime*”). Figure 4 provides the complete set of labels available for an RoT breakdown.<sup>5</sup>

📍 **Grounding Attributes** We call three attributes *grounding attributes*. Their goal is to ground the RoT and action to the situation and characters. At the RoT-level, workers mark which character should heed the RoT with the **RoT Targeting** attribute. At the action level, workers first pick the **action’s best candidate** character, for whom the action is most relevant. However, since RoTs can identify actions that are both explicit and hypothetical in the situation, we additionally annotate whether the candidate character is explicitly **taking the action** in the situation.

👥 **Social Attributes** The second set of attributes characterize social expectations in an RoT. The first two social attributes both label **anticipated agreement**. For an RoT, this attribute asks how many people probably *agree* with the RoT as stated. At the action level, it asks what portion of people probably agree with the *judgment* given the *action*.

Four social attributes relate to the theoretical underpinnings of this work in §2. An RoT-level

<sup>5</sup>Workers are given the choice to mark the RoT as confusing, vague, or low quality, and move on (2% of RoTs).

attribute is the set of **Moral Foundations**, based on a well-known social psychology theory that outlines culturally innate moral reasoning (Haidt, 2012). The action-level attributes **legality** and **cultural pressure** are designed to reflect the two-coarse-grained categories proposed by the Social Norms Theory (Kitts and Chiang, 2008; Perkins and Berkowitz, 1986). Legality corresponds to prescriptive norms: what one ought to do. Cultural pressure corresponds to descriptive norms: what one is socially influenced to do. Finally, the **social judgment** aims to capture subjective moral judgment. A base judgment of what is good or bad is thought to intrinsically motivate social norms (Malle et al., 2014; Haidt et al., 1993).

The final two attributes provide a coarse categorization over RoTs and actions. The **RoT Category** attribute estimate distinctions between morality, social norms, and other kinds of advice. This aims to separate moral directives from tips or general world knowledge (e.g., “It is good to eat when you are hungry”). The attribute **agency** is designed to let workers distinguish RoTs that involve agentive action from those that indicate an experience (e.g., “It is sad to lose a family member”).

### 3.4 Analysis

We briefly highlight three key aspects of our formalism: social judgment, anticipated agreement, and cultural pressure. Figure 5 shows two plots partitioning RoTs based on these three attributes (with legality also highlighted in the left plot (a)).

In the left plot (Figure 5 (a)), the  $x$ -axis contains a new quantity, where social judgment ( $\in [-2, 2]$ ) is multiplied by agreement ( $\in [0, 4]$ ) to scale it.<sup>6</sup> The result is that  $x$  values range from universally-agreed bad actions (-8) to universally-agreed good actions (+8). Intuitively, the bottom-left group shows illegal actions, which are both “bad” (left  $x$ ) and people feel strong pressure not to do (bottom  $y$ ). The data are generally distributed in a line towards the top right, which are “good” (right  $x$ ) actions that people feel strong pressure to do (top  $y$ ).

However, the spread of the data in Figure 5 (a) illustrates the difference between morality and cultural pressure. There are a range of morally charged

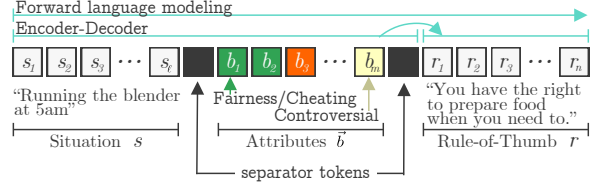


Figure 6: Illustration of modeling setup for the objective  $p(r|s, \vec{b}_r)$ .

actions, but for which people don’t feel cultural pressure (the horizontal range in  $x$  values across the central  $y = \text{Discretionary}$ ). Conversely, we observe actions that are morally neutral, but for which people do feel cultural pressure (the vertical range in  $y$  values along the middle  $x = 0$ ).

The right plot, Figure 5 (b), shows social judgment against agreement, colored by cultural pressure. At high levels of agreement (top of graph), cultural pressure (color) follows social judgment (horizontal changes in  $x$  values). However, for controversially-held judgments (lower  $y$  values), we see a range of cultural pressure. This includes morally good or bad actions that are still discretionary (middle  $y$  values), as well as morally neutral actions for which people feel strong cultural pressure (lower  $y$  values).

These plots illustrate two ways of stratifying actions along socially relevant dimensions. We anticipate considerable further dataset exploration remains.

## 4 Model

We investigate neural models based on pre-trained language models for learning various sub-tasks derived from SOCIAL-CHEM-101.

### 4.1 Training Objectives

Our main modeling formulation is straightforward. Given a situation ( $s$ ), we wish to model the conditional distribution of RoTs ( $r$ ), actions ( $a$ ), and set of attributes from the breakdown ( $\vec{b}$ ). We can partition the attributes  $\vec{b} = \{\vec{b}_r, \vec{b}_a\}$  into disjoint sets relevant to the RoT and action, and write

$$p(r, a, \vec{b}|s) = \underbrace{p(a, \vec{b}_a|r, \vec{b}_r, s)}_{\text{action transcription}} \times \underbrace{p(r, \vec{b}_r|s)}_{\text{RoT prediction}}. \quad (1)$$

Equation 1 allows us to model all components of interest given a situation  $s$ . However, the *action transcription* term is quite strongly conditioned, because actions are so closely related to their RoTs.

<sup>6</sup>Strict statisticians will note that plotting ordinal values numerically is an abuse of notation, much less scaling two values together. We present these graphs for illustrative purposes to observe the stratification of our dataset, not to make quantitative claims.

Objective		
RoT	Action	Interpretation
$p(r s)$	$p(a s)$	Text-only generation
$p(\vec{b}_r s)$	$p(\vec{b}_a s)$	Attribute prediction
$p(r s, \vec{b}_r)$	$p(a s, \vec{b}_a)$	Controlled generation
$p(\vec{b}_r s, r)$	$p(\vec{b}_a s, a)$	Attribute labeling
$p(r, \vec{b}_r s)$	$p(a, \vec{b}_a s)$	Model choice generation

Table 1: Generative model objectives corresponding to the training setups we consider. Each model (RoT or action) is trained on all objectives simultaneously.

In this paper, we instead focus our study of actions on a more difficult distribution that conditions only on the situation:

$$\underbrace{p(a, \vec{b}_a|r, \vec{b}_r, s)}_{\text{action transcription}} \xrightarrow{\text{omit RoT}} \underbrace{p(a, \vec{b}_a|s)}_{\text{action prediction}}. \quad (2)$$

We model both the *RoT prediction* (Eq. 1) and *action prediction* (Eq. 2) distributions with conditional forward language modeling. We tokenize all quantities ( $s, r, a, \vec{b}$ ), creating unique tokens for each attribute value  $b_i$ , and concatenate them together in a canonical order to form strings  $p(x_{\text{out}}|x_{\text{in}})$ . We then train to maximize the standard language modeling objective:

$$x = [x_{\text{in}}; x_{\text{out}}], \quad p(x) = \prod_{i=1}^n p(x_i|x_{<i}). \quad (3)$$

Both the *RoT prediction* (Eq. 1) and *action prediction* (Eq. 2) distributions have similar forms  $p(y, \vec{b}_y|s)$  for  $y \in \{r, a\}$ . We take advantage of this symmetry to study variations of both distributions. Inspired by recent work (Zellers et al., 2019), we construct permutations of our data that omit different fields while maintaining the canonical order. Table 1 shows the setups that we consider, and Figure 6 illustrates an example objective.

We train each model (either RoT or action) on all relevant objectives in Table 1 (i.e., one of the columns). Intuitively, this allows the model to condition on and generate a range of fields.<sup>7</sup> We can do this by simply treating each objective as defining a subset of the fields, as well as their ordering, for each data point. Then, we combine and shuffle all objectives’ views of the data.

<sup>7</sup>It is possible to remove the assumption that the situation is provided, which would allow the model to generate  $s$  as well. We leave such experiments for future work.

## 4.2 Architectures

We present results for the GPT and GPT-2 architectures (Radford et al., 2018, 2019), as well as two encoder-decoder language models (BART and T5, Lewis et al., 2019; Raffel et al., 2019). We train forward language models with loss over the entire sequence  $x$ , whereas encoder-decoder models only compute loss for the output sequence  $x_{\text{out}}$ . Collectively, we term these architectures trained on our objectives the NEURAL NORM TRANSFORMER.

## 5 Experiments and Results

### 5.1 Tasks

While we train each model on all (RoT or action) objectives at once, we pick two particular objectives to assess the models. The first is  $p(y, \vec{b}_y|s)$  — “*model choice*.” In this setting, each model is allowed to pick the most likely attributes  $\vec{b}_y$  given a situation  $s$ , and generate an RoT (or action)  $y$  that adheres to those attributes. This setup should be easier because a model is allowed to pick the conditions of its own generation ( $\vec{b}_y$ ).

The second setting is  $p(y|s, \vec{b}_y)$  — “*conditional*.” We provide models with a set of attributes  $\vec{b}_y$  that they must follow when generating an RoT (or action)  $y$ . This presents a more challenging setup, because models cannot simply condition on the set of attributes that they find most likely. We select sets of attributes  $\vec{b}_y$  provided by the human annotators for the situation  $s$  to ensure models are not tasked with generating from impossible constraints.

**Setup** We split our dataset into 80/10/10% train/dev/test partitions by situation, such that each domain’s situations are proportionally distributed. This guarantees previously unobserved dev and test situations. For all models we use top- $p$  decoding with  $p = 0.9$  (Holtzman et al., 2020).

**Baselines** We use a *Random RoT* baseline to verify the dataset diversity (selections should have low relevance to test situations) and evaluation setup (RoTs and actions should still be internally consistent). We also use a *BERT-Score* (Zhang et al., 2020) retrieval baseline that finds the most similar training situation. If attributes  $\vec{b}_y$  are provided, the retriever picks the RoT (or action) from the retrieved situation with the most similar attributes.

**Ablations** We report two model ablations. For *-Small*, we finetune GPT-2 Small with the same general architecture. For *-No pretrain*, we randomly

	→ <i>RoT</i>				→ <i>Action</i>							
	Category	Moral F.	Agree	Relevance	Agency	Judgment	Agree	Pressure	Legal	Taking	Relevance	
Random RoT	0.73	0.84	0.48	1.25	0.90	0.57	<b>0.55</b>	0.53	0.80	0.04	1.22	Model choice $p(y_i \hat{h}_y(s))$
BERT-Score (Z et al., 2020)	<b>0.76</b>	0.83	0.48	2.00	0.90	<b>0.64</b>	0.46	<b>0.61</b>	0.81	0.20	2.00	
GPT (R et al., 2018)	0.71	0.77	0.39	2.23	0.82	0.40	0.36	0.32	0.76	0.15	2.25	
BART (L et al., 2019)	0.69	0.79	<b>0.49</b>	2.60	<b>0.91</b>	0.55	0.54	0.46	0.80	0.18	2.52	
T5 (R et al., 2019)	0.62	<b>0.85</b>	0.42	<b>2.78</b>	0.78	0.36	0.36	0.23	0.56	0.23	<b>2.73</b>	
GPT-2 Small (R et al., 2019)	0.62	0.79	0.34	2.03	0.82	0.34	0.34	0.27	0.79	0.09	1.99	
GPT-2 XL - No pre-train	0.68	0.78	0.20	1.37	0.81	0.37	0.30	0.33	0.79	0.06	1.29	
GPT-2 XL	0.75	0.84	0.42	2.53	<b>0.91</b>	0.51	0.36	0.45	<b>0.82</b>	<b>0.32</b>	2.60	Controlled $p(y_i s, \hat{h}_y(s))$
Random RoT	0.59	0.75	<b>0.41</b>	1.20	0.84	0.27	0.28	0.21	0.74	0.01	1.19	
BERT-Score (Z et al., 2020)	0.66	0.78	<b>0.41</b>	2.00	0.87	0.40	<b>0.45</b>	0.34	<b>0.76</b>	0.16	1.97	
GPT (R et al., 2018)	0.64	0.79	0.36	2.21	0.83	0.46	0.36	0.38	0.74	0.17	2.26	
BART (L et al., 2019)	0.70	<b>0.81</b>	0.38	2.60	0.84	0.47	0.42	0.41	0.73	0.20	2.44	
T5 (R et al., 2019)	0.66	0.80	0.40	<b>2.77</b>	0.83	0.41	0.34	0.38	0.73	0.24	<b>2.79</b>	
GPT-2 Small (R et al., 2019)	0.64	0.78	0.30	2.10	0.78	0.38	0.30	0.27	0.71	0.10	1.97	
GPT-2 XL - No pre-train	0.67	0.79	0.23	1.35	0.83	0.36	0.32	0.26	0.73	0.04	1.33	
GPT-2 XL	<b>0.71</b>	0.79	0.38	2.65	<b>0.90</b>	<b>0.51</b>	0.38	<b>0.42</b>	0.74	<b>0.28</b>	2.54	

Table 2: Human evaluation results for conditionally generating RoTs and actions, either letting the models choose the attributes (top half), or providing the attributes as input constraints (bottom half). All columns are micro-F1 scores (0–1), except *Relevance* (1–3). **Takeaway:** While state-of-the-art models are able to generate relevant RoTs and actions that generally follow constraints (moderately high scores in some columns), correctly conditioning on a complete set of attributes remains challenging (several columns show poor model performance in bottom half).

Model	Ppl.	BLEU-4	Attr. $\mu F1$
→ <i>RoT</i>			
GPT	1.81	5.41	0.42
Bart-large	1.76	6.65	0.47
T5-large	1.94	<b>10.79</b>	0.34
GPT-2 Small	1.97	4.97	0.38
GPT-2 XL - No fine-tune	-	0.46	0.20
GPT-2 XL - No pre-train	2.54	4.39	0.42
GPT-2 XL	1.75	6.53	<b>0.53</b>
→ <i>Action</i>			
GPT	1.80	6.75	0.60
BART-Large	1.72	8.34	0.66
T5-Large	2.00	<b>8.93</b>	0.58
GPT-2 Small	1.94	6.62	0.56
GPT-2 XL - No fine-tune	-	0.25	0.52
GPT-2 XL - No pre-train	2.51	5.43	0.55
GPT-2 XL	1.73	7.98	<b>0.68</b>

Table 3: Test set performance by automatic metrics, including an attribute classifier. Perplexities are not comparable between encoder-decoder models (Bart and T5, loss on  $x_{out}$  only) and other models (loss on full sequence  $x$ ). **Takeaway:** Automatic metrics corroborate human evaluation results: while T5 is most adept at BLEU, GPT-2 XL more consistently adheres to attributes (Attr.  $\mu F1$ ).

initialize the model’s weights.<sup>8</sup>

## 5.2 Results

**Human Evaluation** Table 2 presents a human evaluation measuring how effective models are at generating RoTs and actions for both task settings. While most columns measure attribute adherence, the *Relevance* score is critical for distinguishing

<sup>8</sup>We omit the evaluation of an “out-of-the-box GPT2-XL” baseline (i.e. no fine-tuning) whose outputs predictably do not resemble RoTs or actions.

whether RoTs actually apply to the provided situation (e.g., see low scores for the *Random RoT* baseline). In both setups, T5’s generations rank as most tightly relevant to the situation. But in terms of correctly following attributes, GPT-2 is more consistent, especially in the *controlled* task setup (lower; top scores on 5/9 attributes). However, no model is able to achieve a high score on all columns in the bottom half of the table. This indicates that fully constrained conditional generation may still present a significant challenge for current models.

**Automatic Evaluation** We also provide automatic metrics of the generated outputs. We train attributes classifiers using RoBERTa (Liu et al., 2019), and use them to classify the model outputs.<sup>9</sup>

Table 3 presents test set model performance on perplexity, BLEU (Papineni et al., 2002), and attribute micro-F1 classifier score. The automatic metrics are consistent with human evaluation. T5 is a strong generator overall, achieving the highest BLEU score and the highest *relevance* score in §5.2. However, GPT-2 more consistently adheres to attributes, outperforming T5 in attribute  $F_1$  with nearly 20 points gap for RoTs, and over 10 points for actions.

## 6 Morality & Political Bias

To demonstrate a use case of our proposed formalism, we analyze the social norms and expectations evoked in news headlines from news sources of

<sup>9</sup>BERT and BART performed worse across attributes.



		Left (-) or Right (+)	Reliability
	Agreement	-0.015**	-0.008*
ROT Cat.	Morality / Ethics	-0.069***	-0.022***
	Social Norms	0.019***	-0.006*
	It is what it is	0.039***	-0.007**
	Advice	0.031***	0.033***
Moral F.	Care / Harm	-0.033***	-0.016***
	Authority / Subversion	<i>n.s.</i>	<i>n.s.</i>
	Fairness / Cheating	-0.050***	<i>n.s.</i>
	Loyalty / Betrayal	0.026***	-0.007**
	Sanctity / Degradation	0.014**	-0.017***

Table 4: Correlations between generated RoT attributes for headlines and the news source’s political leaning (left: neg., right: pos.) and reliability (controlled for political leaning). Results shown are significant after Holm-correction for multiple comparisons ( $p < 0.001$ : \*\*\*,  $p < 0.01$ : \*\*,  $p < 0.05$ : \*,  $p > 0.05$ : *n.s.*).

**Takeaway:** We see evidence that a model trained on the SOCIAL-CHEM-101 Dataset can naturally uncover moral and topical leanings in news sources, mirroring results found in previous news studies.

various political leanings and trustworthiness, using the NEURAL NORM TRANSFORMER (GPT-2 XL). Specifically, we generate ROTs and attributes for 50,000 news headlines randomly selected from Nørregaard et al. (2019), a large corpus of political headlines from 2018 paired with news source ratings of political leaning (5-point scale from left- to right-leaning) and factual reliability (5-point scale from least reliable to most reliable).<sup>10</sup>

Table 4 shows the correlations between RoT attributes and the political leaning and reliability of sources. Our results strongly corroborate findings by Graham et al. (2009), showing that liberal headlines evoke more “fairness” and “care,” while right-leaning headlines evoke more “sanctity” and “loyalty.” Furthermore, in line with findings by Volkova et al. (2017), more reliable news source tend to evoke more advice and less morality.

## 7 Related Work

Our formalism heavily draws from works in descriptive ethics and social psychology, but is also inspired by studies in social implicatures and cooperative principles in pragmatics (Kallia, 2004; Grice, 1975) and the theories of situationally-rooted evocation of frames (Fillmore and Baker, 2001).

Our work adds to the growing literature concerned with distilling reactions to situations (Vu et al., 2014; Ding and Riloff, 2016) as well as so-

<sup>10</sup>We use the MediaBias/FactCheck ratings: <https://mediabiasfactcheck.com>.

cial and moral dynamics in language (Van Hee et al., 2015). Commonly used for coarse-grained analyses of morality in text (Fulgoni et al., 2016; Volkova et al., 2017; Weber et al., 2018), Graham et al. (2009) introduce the Moral Foundations lexicon, a dictionary of morality-evoking words (later extended by Rezapour et al., 2019).

A recent line of work focused on representing social implications of everyday situations in free-form text in a knowledge graph (Rashkin et al., 2018; Sap et al., 2019). Relatedly, Sap et al. (2020) introduce Social Bias Frames, a hybrid free-text and categorical formalism to reason about biased implications in language. In contrast, our work formalizes a new type of reasoning around expectations of social norms evoked by situations.

Finally, concurrent works have developed rich and exciting resources studying similar phenomena. Tay et al. (2020) study *Would you rather?* questions, and Acharya et al. (2020) investigate ritual understanding across cultures. Hendrycks et al. (2020) study ethical questions, attempting to assign a real-valued utility to scenarios across a range of ethical categories. And Lourie et al. (2020) define the challenge of predicting the *r/AITA* task using the full posts. In contrast to these studies, our work addresses norms by distilling cultural knowledge to a new conceptual level of Rules-of-Thumb and corresponding structural annotations.

## 8 Conclusion

We present SOCIAL-CHEM-101, an attempt at providing a formalism and resource around the study of grounded social, moral, and ethical norms. Our experiments demonstrate preliminary success in generative modeling of structured ROTs, and corroborate findings of moral leaning in an extrinsic task. Comprehensive modeling of social norms presents a promising challenge for NLP work in the future.

## Acknowledgments

The authors would like to thank Nicholas Lourie, Rowan Zellers, and Chandra Bhagavatula. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE1256082, and in part by NSF (IIS-1714566), DARPA CwC through ARO (W911NF15-1-0543), DARPA MCS program through NIWC Pacific (N66001-19-2-4031), and the Allen Institute for AI.

## References

- Anurag Acharya, Kartik Talamadupula, and Mark A Finlayson. 2020. [An atlas of cultural commonsense for machine reasoning](#).
- George J Bowdery. 1941. Conventions and norms. *Philosophy of Science*, 8(4):493–505.
- Haibo Ding and Ellen Riloff. 2016. Acquiring knowledge of affective events from blogs using label propagation. In *AAAI*.
- Jennifer L Eberhardt. 2020. *Biased: Uncovering the hidden prejudice that shapes what we see, think, and do*. Penguin Books.
- Jon Elster. 2006. Fairness and norms. *Social Research*, pages 365–376.
- Charles J Fillmore and Collin F Baker. 2001. Frame semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop, NAACL*, volume 6.
- Dean Fulgoni, Jordan Carpenter, Lyle Ungar, and Daniel Preotiu-Pietro. 2016. An empirical exploration of moral foundations theory in partisan news sources. In *LREC*, pages 3730–3736.
- Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *J. Pers. Soc. Psychol.*, 96(5):1029–1046.
- Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.
- Jonathan Haidt. 2012. *The righteous mind: Why good people are divided by politics and religion*. Vintage.
- Jonathan Haidt, Silvia Helena Koller, and Maria G Dias. 1993. Affect, culture, and morality, or is it wrong to eat your dog? *Journal of personality and social psychology*, 65(4):613.
- Richard Mervyn Hare, Richard Mervyn Hare, Richard Mervyn Hare Hare, and Richard M Hare. 1981. *Moral thinking: Its levels, method, and point*. Oxford: Clarendon Press; New York: Oxford University Press.
- Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. [Aligning ai with shared human values](#).
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text de-generation](#). In *International Conference on Learning Representations*.
- Jerome Kagan. 1984. *The nature of the child*. Basic Books.
- Alexandra Kallia. 2004. Linguistic politeness: The implicature approach. *Multilingua*, 23(1/2):145–170.
- James A Kitts and Yen-Sheng Chiang. 2008. Encyclopedia of social problems.,
- Lawrence Kohlberg. 1976. Moral stages and moralization. *Moral development and behavior*, pages 31–53.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.
- Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2020. [Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes](#). *arXiv e-prints*.
- Bertram F Malle, Steve Guglielmo, and Andrew E Monroe. 2014. A theory of blame. *Psychological Inquiry*, 25(2):147–186.
- George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.
- Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. [A corpus and cloze evaluation for deeper understanding of commonsense stories](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.
- J Nørregaard, B D Horne, and S Adalı. 2019. NELA-GT-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *AAAI*. [www.aaai.org](http://www.aaai.org).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- Gonalo Pereira, Rui Prada, and Pedro A Santos. 2016. Integrating social power into the decision-making of cognitive agents. *Artificial Intelligence*, 241:1–44.
- H Wesley Perkins and Alan D Berkowitz. 1986. Perceiving the community norms of alcohol use among students: Some research implications for campus alcohol education programming. *International journal of the Addictions*, 21(9-10):961–976.

- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.
- Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A Smith, and Yejin Choi. 2018. Event2mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473.
- Rezvaneh Rezapour, Saumil H Shah, and Jana Diesner. 2019. Enhancing the measurement of social effects by capturing morality. In *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 35–45, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*.
- Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. ATOMIC: An atlas of machine commonsense for if-then reasoning. In *AAAI*.
- Richard A Shweder. 1990. In defense of moral realism: Reply to gabennesch. *Child Development*, 61(6):2060–2067.
- Yi Tay, Donovan Ong, Jie Fu, Alvin Chan, Nancy Chen, Anh Tuan Luu, and Christopher Pal. 2020. Would you rather? a new benchmark for learning machine alignment with cultural values and social preferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5369–5373.
- Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology-volume 1*, pages 173–180. Association for Computational Linguistics.
- Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015. [Detection and fine-grained classification of cyberbullying events](#). In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 672–680, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.
- Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *ACL*, pages 647–653, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Hoa Trong Vu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Acquiring a dictionary of emotion-provoking events. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 128–132.
- Su Wang, Greg Durrett, and Katrin Erk. 2018. Modeling semantic plausibility by injecting world knowledge. In *NAACL-HLT*.
- René Weber, J Michael Mangus, Richard Huskey, Fred-eric R Hopp, Ori Amir, Reid Swanson, Andrew Gordon, Peter Khooshabeh, Lindsay Hahn, and Ron Tamborini. 2018. Extracting latent moral information from text narratives: Relevance, challenges, and solutions. *Commun. Methods Meas.*, 12(2-3):119–139.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R’emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.
- Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. [Defending against neural fake news](#). In *NeurIPS*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.

## A Additional Dataset Details

### A.1 Situations

**Domains** We provide here a more thorough description how we collected situations from the four domains we consider. Figure 7 gives more example situations from each domain.

1. **r/amitheasshole** (30k) — The *Am I the Asshole?* (AITA) subreddit. This posts of this subreddit pose moral quandries, such as “AITA for wanting to uninvite an (ex?)-friend from my wedding for shit-talking our marriage?” We use the data from [Lourie et al. \(2020\)](#). They scrape the titles of posts, omitting the preamble (e.g., “AITA for”), normalizing to present tense, and filtering out administrative posts. We do not use any annotations provided by that community (where other posters vote who had the moral high ground).

<b>[r/amitheasshole]</b>
– telling my friend and her family to move out
– choosing to spend time with my friends or boyfriend rather than my family
– not wanting to hangout with sick girlfriend
– not wanting to do household chores
– banning my ex from my Spotify account
<b>[r/confessions]</b>
– My SO thinks I hate pickles, I like pickles but he LOVES pickles so I always pretend to hate them so he can have them.
– Best friend just got engaged.
– My girlfriend cheated and im cheating back on her
– I hate myself because I couldn't save my mother
– I'm scared of being a dad
<b>[rocstories]</b>
– Clark Ryder was proud of his job as a photojournalist.
– They had so many questions that I couldn't answer.
– Her husband surprised her on her birthday with plane tickets!
– She decided to wear slippers to protect her feet from Jason's toys.
– When he got to the assembled class he became very nervous.
<b>[dearabby]</b>
– Family of Six Tries Not to Be a Burden on Weekend Hosts
– Breakup Letter to Soldier Could Jeopardize Comrades in Arms
– Gentle Nudge Has Not Worked to Dislodge Mom From House
– Planning Helps Students Get Good Letter of Recommendation
– Man With Breast Cancer Experiences Extra Stress

Figure 7: Five randomly sampled situations from each of the four domains we consider.

2. **r/confessions** (32k) — The *Confessions* subreddit. This posts of this subreddit discuss personal stories, often with interpersonal conflicts, such as “*I feel threatened by women prettier than me.*” As with r/AITA, we scrape only the titles of these posts. This subreddit contains a high volume of hateful or disturbing content; we attempt to filter the worst of this using keywords, and also allow annotators to mark dark or disturbing items.
3. **rocstories** (30k) — The ROCStories corpus from (Mostafazadeh et al., 2016). ROCStories involve stories about everyday situations, and are generally less controversial than the other sources, e.g., “*They weren’t sure either so he started asking friends.*”. We select a subset of the sentences from ROCStories which are likely to involve two character references based on POS tagging (Toutanova et al., 2003), personal pronouns, and WordNet (Miller, 1995). We then randomly sample to pick 30k sentences.
4. **dearabby** (12k) — Titles of the Dear Abby advice column. These titles are usually information dense summaries of interpersonal situations written in the style of news headlines, e.g., “*Pushy Party Guests Make Themselves Too Much at Home.*” We scrape all of the titles found in the archives, and use heuristics to attempt to filter out all posts that do not match this style, such as announcements and holiday greetings.

We attempt to balance the number of situations collected for each domain. However, we are limited by the complete set of examples from dearabby (12k).

**Additional Labels** We allow annotators to mark each situation with any of the following labels that apply.

- **Unclear** The situation was too simple, vague, or confusing to understand what happened.
- **NSFW** The situation contains suggestive or adult content.
- **Dark / disturbing / controversial.** The situation contained content that may make folks uncomfortable, like suicide, torture, or abuse.

Annotators may pass on writing RoTs for a situation marked with any of those boxes, or they may still choose to do so. We keep all of the labels collected. They are included in the dataset as additional fields. For example, they could be used to omit certain training data to keep a model biased away from potentially controversial subjects.

## A.2 Character Identification

Our goal during character identification is to find the most descriptive phrase referring to each unique non-narrator person in the passage exactly once.

The reason for this goal is that always having a single, best reference to each person in the situation enables more consistent grounding.

While this goal is relatively straightforward, we find many edge cases arise. In cases where it is unclear if a person should be marked, our central criteria is **whether someone might write RoTs involving that person**. If so, that person should be included so they are a candidate for grounding. We found handling all of these edge cases complex enough to require human annotation instead of heuristics. We provide here the character identification guidelines that we give to the crowd worker annotators, along with an example illustrating each one.

### Character Identification Guidelines

- **Don’t include the (first person) narrator.** For example, “*I ate pizza*” would have no people highlighted.
- **Only include people.** For example, “*My horse George provides good conversation*” would have no people highlighted.
- **Only highlight each person once.** For example, “*I gave my brother a hug, I like him, he’s so nice*”, we would only include my brother, not “*him*” or “*he*.”
- **Highlight the most descriptive mention of a person.** For example, “*I can’t stand him, my brother is so mean.*”, we would pick my brother even though it comes after “*him*.”
- **Include the full phrase referring to the person.** Include words like “*a*”, “*the*”, “*my*”, and longer phrases. For example, “*The strange guy talked to my brother and my oldest uncle,*” we would pick The strange guy, my brother, and my oldest uncle, instead of just “*guy*”, “*brother*”, and “*uncle*.”



- **Don't include phrases where a generic person-looking word is used without referring to a particular person.** This often happens when describing a place or thing. For example, "*I walked into the men's room,*" we would not pick anything, because "*men's room*" is a generic phrase. Similarly, we would not pick anything for, "*I am a child,*" because "*child*" is just used as a description. But for, "*I walked into my brother's room*", we would pick my brother.
- **Include people used to refer to someone.** For example, "*My brother's girlfriend is so cool,*" we would pick both my brother and my brother's girlfriend.
- **Include pronouns (she, her, hers, etc.) if they're the most specific word available.** For example, in a sentence like "*I love him,*" we would pick him. However, for a sentence like, "*I love my brother, I can always talk to him.*" we would instead pick my brother because it's more specific.
- **Include pronouns like "they" and "them", also if they're the most specific word available.** For example, if we had the sentence "*They went to the party.*" we would pick they. However, if we had the sentence "*My friends went to the party and they had a good time.*" we would instead pick My friends since it is more specific.
- **Include plural first person pronouns (us, we, etc.) once.** For example, in a sentence like "*We went to the park.*" we would pick we. Or for a sentence like "*They spent hours talking to us and we had a good time.*" we would pick they and us.
- **Include other groups of people like "her siblings," "their class," and "his team."** For example, in a sentence like "*I talked to all of his uncles for a while.*" we would pick both his and his uncles.
- **Include proper names of people that aren't the narrator.** For example, in a sentence like "*Mary chased John at the park.*" we assume they are people (unless otherwise specified), and we would pick both Mary and John.
- **Include people with titles like "the policeman" and "the mailman."** For example, in the sentence "*I chased the store clerk.*" we would select the store clerk.
- **Include words like "someone" and "everyone."** For example, in the sentence "*I am going to dinner with someone.*" we would select someone.

### A.3 Rules-of-Thumb (RoTs)

This section provides more information on how RoTs are written. Figure 8 shows a sample of RoTs organized both by situation domain and topic.

As mentioned briefly in Section 3.2 of the paper body, we present workers with a series of guidelines for how to write RoTs. All RoT writing guidelines are in service of the goal that RoTs capture social, ethical, moral, and cultural norms. Unlike the guidelines for character identification, which are largely syntactic, the guidelines for writing RoTs

are semantic. This makes them more challenging both to define and check.

To motivate these guidelines, and to help readers intuitively characterize what RoTs are, we present the RoT writing guidelines here at greater length, annotated with examples and explanations. For each guideline (**in bold**), we provide an example situation (*in italics*) along with candidate RoTs that **violate** or **follow** the guideline.

### RoT Writing Guidelines

- **Explain the basics of good and bad behavior.** RoTs should describe cultural expectations, as if to a file-year-old child who doesn't yet know how the world works.
  - Example situation: *Not wanting to take tests to apply for college*
  - **Violates:** "Studies have shown people perform best on tests after sleeping at least seven hours"
  - **Follows:** "It's normal to be stressed out by exams"
  - **Why:** This broad guideline attempts to distinguish RoTs from encyclopedic knowledge. Instead, RoTs should contain everyday, common-sense knowledge about social norms and expectations.
- **Judgment and action.** An RoT must contain a judgment and an action.
  - Example situation: *Telling my husband he shouldn't buy his dream boat*
  - **Violates:** "Boats are expensive"
  - **Follows:** "It's mean to squash someone's dreams"
  - **Follows:** "People should be open to discussing big purchases with their spouses"
  - **Why:** Requiring an *action* helps ensure RoTs are about things people do. Requiring a *judgment* pushes statements to contain some information about norms and expectations.
- **Self-contained.** An RoT must be fully understandable on its own, without the situation it came from.
  - Example situation: *Being angry at my sister for not attending our fathers funeral because of his criminal history.*
  - **Violates:** "It makes them feel bad"
  - **Violates:** "The father caused emotional distress to his daughter and the narrator should not judge her actions too harshly."
  - **Follows:** "If someone commits serious crimes, it's OK for family to cut off contact with them."
  - **Why:** Without this requirement, RoTs would not naturally generalize to new situations, and would trend too specific. They would also could contain much less information, as much of the semantic content could be left in the situation and only referred to by the RoT.
- **Inspired by situation.** An RoT should be inspired by the situation it came from.
  - Example situation: *Wanting to uninvite a friend from my wedding.*
  - **Violates:** "It's rude to point at people you don't know"
  - **Follows:** "It's devastating to be excluded from a wedding you were invited to"

<p>[r/amitheasshole]  <b>Wanting to uninvite an (ex?)-friend from my wedding for shit-talking our marriage</b></p> <ul style="list-style-type: none"> <li>- When you are paying for a celebration, you are allowed to invite whoever you want.</li> <li>- It is reasonable to rescind an invitation to a wedding if someone is no longer your friend.</li> <li>- Telling someone they can't come to your wedding after they were already invited is tacky.</li> </ul>	<p>[r/confessions]  <b>I feel threatened by women prettier than me</b></p> <ul style="list-style-type: none"> <li>- It's bad to feel threatened.</li> <li>- It's normal to feel intimidated by others.</li> <li>- It's ok for someone to be prettier than you.</li> <li>- It's normal to compare yourself with others.</li> </ul>	<p>[dearabby]  <b>Pushy Party Guests Make Themselves Too Much at Home</b></p> <ul style="list-style-type: none"> <li>- You should respect other people's property.</li> <li>- You should admit to breaking something rather than converging it up.</li> <li>- It's OK to turn down an invitation if you're not interested in going.</li> <li>- It's rude to exclude others from a get-together.</li> </ul>	<p>[rocestories]  <b>They weren't sure either so he started asking friends.</b></p> <ul style="list-style-type: none"> <li>- It's okay to ask your friends about something you need to know.</li> <li>- It's understandable if you're uncertain of what to do.</li> <li>- You should ask for advice when you aren't sure what the right course of action is.</li> <li>- It's good to give your friend advice when they ask for it.</li> <li>- It's okay to be scared when you're not sure what to do.</li> </ul>
<ul style="list-style-type: none"> <li>- Trying to warn a coworker about the dangers of smoking is caring.</li> <li>- It's okay to ask someone not to smoke in your car.</li> <li>- It's wrong to pretend that you're smoking because it's unhealthy to smoke and you shouldn't idolize people that do.</li> <li>- You shouldn't accept cigarettes from friends when you don't smoke.</li> </ul>	<ul style="list-style-type: none"> <li>- You should not smoke inside.</li> <li>- It is bad to expose others to second hand smoke</li> <li>- It's bad to smoke.</li> <li>- It's bad for your health to smoke cigarettes.</li> <li>- You shouldn't smoke weed.</li> </ul>	<div>9/451 RoTs randomly sampled, searching for "smok*" across RoTs from all four domains.</div>	

Figure 8: **Top:** An example situation (bold) and corresponding RoTs (bullets) from each of the four domains we consider. **Bottom:** Random sample of RoTs about smoking, found by searching for *smok\** across the dataset.

- **Why:** Maintaining a link between RoT and situation allows for grounding RoTs during the structured annotation. Furthermore, since a different worker will likely provides the structural annotation for an RoT, relevance to the source situation helps ensure the worker understands the RoT's context and implications.
- **Balance Specificity and Vagueness.** An RoT should be inspired by, and relevant to, the provided situation. However, a rule-of-thumb should also give a general rule for how people behave in society, so should apply to more than just the given situation.
  - Example situation: *Not tipping my cashier last Tuesday*
  - **Violates:** “Not tipping a cashier last Tuesday is rude”
  - **Violates:** “It’s rude to be cheap”
  - **Follows:** “It’s usually OK not to tip cashiers in retail or grocery stores”
  - **Why:** This requirement can be the hardest to assess because of its subjectivity. RoTs that are too specific are usually slight modifications of the situation that include a judgment, and don’t describe underlying expectations. RoTs that are too vague often do describe norms, but the link to the situation can be so distant as to be misleading. Good RoTs may be somewhat specific, but explain both the underlying norms at play, and apply to other situations.
- **Distinct ideas.** When multiple RoTs are provided for a situation, each should contain a distinct idea. This includes inversions of the same idea.
  - Example situation: *Never taking out the trash*
  - **Violates:** “It’s irresponsible to avoid the chores you are assigned” *with* “It’s bad to not do chores you’re supposed to do”
  - **Violates:** “It’s irresponsible to avoid the chores you are assigned” *with* “It’s responsible to do the chores you are assigned”
  - **Follows:** “It’s irresponsible to avoid the chores you are assigned” *with* “You’re generally expected to have a clean home”
  - **Why:** This requirement is to prevent merely collecting paraphrases of the same RoT. Furthermore,

we assume that inversions are usually trivial semantic mutations, so they are also not worth collecting at scale.

## A.4 RoT Breakdowns

In this section, we provide more information about the structural RoT annotations, which we call “RoT Breakdowns.” In particular, we illustrate the potential values for each attribute with an example.

### A.4.1 RoT Categorization

RoT categories are originally designed to distinguish more desired annotation topics (morality/ethics, social norms) from less desired ones (advice and “it is what it is” statements). RoT categories are not mutually-exclusive, and the lines are not always clear. While we use all data regardless of RoT category in this paper’s experiments, future work using this dataset may consider filtering based on RoT category. Annotators are allowed to select from none through all, but are encouraged to select the 1 – 2 best.

#### Information

Type	🌐 Social
Annotated	RoT
Prompt	What category best fits this RoT?

Values		
Label	Description	Example
Morality/ Ethics	Rules or guiding principles of right or wrong conduct	<i>It's wrong to turn your back on your friends who need your help.</i>
Social Norms	Standards of appropriate behavior in a society; emphasizes social conventions	<i>It's good to shake hands with your opponent even if you lost.</i>
Advice	Prudent behaviors that are aimed at improving one's life	<i>It's good to take medicine your doctor prescribes.</i>
It is what it is	Describes how things are; avoids ethics, morality or social norms	<i>It is nice to be tall.</i>

### A.4.2 Moral Foundations

To simplify the annotation, we label *axes* of moral foundations. For example, *Care/Harm* are annotated together, rather than as *Care* and *Harm* separately. Other social attributes, such as *social judgment*, attempt to explicitly capture the goodness or badness of the act. Additionally, we omit the sixth moral foundation, *Liberty/Oppression*, which was later added to the formalism after the first five. Some examples are drawn from literature on Moral Foundations (Haidt, 2012). Note that these labels are not mutually exclusive, and annotators may choose from none through all.

Information		
Type	🧑 Social	
Annotated	RoT	
Prompt	Which moral foundations best fit this RoT?	
Values		
Label	Description	Example
Care/Harm	Preventing or inflicting pain or suffering	<i>It's mean to tell someone they aren't attractive as someone else.</i>
Fairness/Cheating	Notions of equity, justice, and rights	<i>It's wrong to copy someone else's answers in an exam.</i>
Loyalty/Betrayal	Obligations or concerns for group, family, and nation.	<i>It's rude for a cheerleader to boo their own team.</i>
Authority/Subversion	Submission and deference to traditions or legitimate authority.	<i>It's rude to walk away from your boss while they are talking to you.</i>
Sanctity/Degradation	Abhorrence for disgusting things or actions; purity	<i>It's disgusting to pee into a public pool full of people.</i>

### A.4.3 Action and Judgment

Each RoT should contain an action and a judgment. An essential step in the RoT breakdown is to identify the action so that it may be annotated with action attributes. We have annotators write both the action and the judgment, conjugating to the action to gerund tense, and rewriting when needed.

If the annotator is unable to clearly separate an action and a judgment from the RoT, they can simply mark the RoT as unclear and move on.


We provide here a few examples of RoTs along with corresponding actions and judgments. Some are straightforward, while others require more substantial rewriting.

- **RoT:** It's good to care for the elderly.
  - **Action:** caring for the elderly
  - **Judgment:** it's good
- **RoT:** It's okay to be angry if your friend talks to someone you used to date
  - **Action:** being angry when your friend talks to someone you used to date
  - **Judgment:** it's okay
- **RoT:** The holidays are expected to be especially difficult for those that are grieving.
  - **Action:** finding the holidays difficult while grieving
  - **Judgment:** expected

### A.4.4 Agency

It can be challenging to distinguish *agency* from *experience* in cases where the action involves thinking thoughts or feeling emotions. We provide the following additional examples to workers for these cases, and allow their discretion during the annotation:


- **Experience:** Feeling upset when someone calls you a name
- **Agency:** Being mad for two days when someone calls you a name
- **Agency:** Taking revenge on someone for calling you a name

Information		
Type	 Social	
Annotated	Action	
Prompt	Is the action <u>\$action</u> something you do or control, or is it something you experience?	
Values		
Label	Description	Example
Agency	You can do this	<i>Doing the dishes</i>
Experience	This happens to you; you have no control	<i>Being stuck in an earthquake</i>






### A.4.5 Social Judgment

For social judgment, we ask workers to transcribe the intent of RoT’s original judgment, rather than pick their own. Then, workers can mark their disagreement through their annotation of the *anticipated agreement* attribute. This allows us to faithfully capture controversial opinions, while also marking them as controversial.

#### Information

**Type**  Social  
**Annotated** Action  
**Prompt** Which best matches the RoT’s original judgment ( \$judgmentSpan ) of \$action ?


#### Values

Label	Description	Example
 Very bad	Spiteful, malicious, or insulting	<i>Slashing tires; Destroying someone’s phone</i>
 Bad	Frowned upon, rude, unkind, bad	<i>Littering; saying something mean to a friend</i>
 Expected / OK	Normal; generally expected by society	<i>Wearing clothes; showing up to work on time</i>
 Good	Kind of considerate	<i>Driving a friend to the airport</i>
 Very good	Generous, self-sacrificing, perhaps even heroic	<i>Buying groceries for a financially struggling neighbor</i>






### A.4.6 Anticipated Agreement

Anticipated agreement is theoretically continuously valued, compared to other attributes that are more naturally ordinal. Inspired by Wang et al. (2018), we bin this attribute into intuitive ranges, primarily to reduce cognitive load during annotation.

#### Information

**Type**  Social  
**Annotated** RoT, Action  
**Prompts** **RoT:** What portion of people probably agree that \$RoT ?  
**Action:** What portion of people probably agree that \$action is \$socialJudgment ?


#### Values

Label	Description	Example
 < 1%	Almost no one	<i>Murdering people is great</i>
 ~ 5–25%	People occasionally think this.	<i>Ghosts exist</i>
 ~ 50%	Controversial (people naturally disagree)	<i>It’s okay to have an abortion</i>
 ~ 75–90%	Common belief	<i>It’s okay to try again when you fail</i>
 > 90%	Universally true	<i>It’s bad to deliberately hurt your pets</i>




### A.4.7 Legality

Annotations of legality are usually straightforward. Whether an action is illegal but tolerated can vary depending on location and knowledge of the law.

#### Information

**Type**  Social  
**Annotated** Action  
**Prompt** Where you live, how legal is the action \$action ?


#### Values

Label	Description	Example
 Illegal	Legal consequences if caught	<i>Theft; murder</i>
 Depends/ Tolerated	Generally “illegal”, but often unenforced depending on circumstances	<i>Using a cellphone while driving</i>
 Legal	Not illegal	<i>Coughing without covering one’s mouth</i>

### A.4.8 Cultural Pressure

We provide instructions that cultural pressure could come from one’s family, friends, community, culture, or society at large. We ask annotators to evaluate cultural pressure according to their own feelings.

#### Information

**Type**  Social  
**Annotated** Action  
**Prompt** How much cultural pressure do you (or those you know) feel about \$action ?



#### Values

Label	Description	Example
<span style="color: red;">■</span> Strong pressure against	Culture frowns upon this action	<i>Intentionally harming an animal</i>
<span style="color: orange;">■</span> Pressure against	Culture generally discourages this action	<i>Spending money on jewelry if you can't afford it</i>
<span style="color: yellow;">■</span> Discretionary	Culture has little or nothing to say about this action	<i>Choosing to read before bed</i>
<span style="color: lightblue;">■</span> Pressure for	Culture generally encourages this action	<i>Being honest with people</i>
<span style="color: blue;">■</span> Strong pressure for	Culture strongly promotes this action	<i>Wearing clothes in public</i>

#### A.4.9 Taking Action

RoTs are written for a range of both hypothetical and actual actions related to the provided situation. Furthermore, sometimes the action is one that is explicitly not happening. This attribute labels how likely it is that the action is being taken by the relevant character. *Note: a subset of the  $\tau$ /AITA annotations were performed before the “probably not” label was introduced; for those, “hypothetical” is marked instead.*

#### Information

Type	Ⓢ Grounded
Annotated	Action
Prompt	Is <u>\$candidateCharacter</u> explicitly doing the action <u>\$action</u> ? Or is it the action <b>might</b> happen?

The upcoming examples use *narrator* and the following situation for context: *Not tipping the bartender at the club.*

#### Values

Label	Description	Example
<span style="color: red;">■</span> Explicitly not	It's explicitly written that they don't do this	<i>Tipping the bartender</i>
<span style="color: orange;">■</span> Probably not	Most likely not; they probably don't do this	<i>Enjoying the drinks</i>
<span style="color: yellow;">■</span> Hypothetical	We can't say / no evidence	<i>Going clubbing every day</i>
<span style="color: lightblue;">■</span> Probable	Most likely; hints are written	<i>Paying for drinks</i>
<span style="color: blue;">■</span> Explicit	It's written in the situation	<i>Going to the club</i>

#### A.5 Crowdsourcing

Workers undergo an extensive vetting process before working on RoTs. This includes a paid qualification (qual) with a quiz on each of the guidelines and a manual review of sample RoTs. Workers then pass the qual move to a staging pool where they can work on a small number of situations, and all of their RoTs are manually reviewed for adherence to the guidelines. After graduating from the staging pool, workers enter the main group of RoT writers and annotators. For every batch of data, we perform spot checks on the RoTs written and annotated by the main group, as well as send feedback to all of the workers answering any questions we receive. We continuously update the instructions with clarifications, new examples, and answers to questions.

#### A.6 Annotator Demographics

With an extensive qualification process, 137 workers participated in our tasks. Of those, 55% were women and 45% men. 89% of workers identified as white, 7% as Black. 39% were in the 30-39 age range, 27% in the 21-29 and 19% in the 40-49 age ranges. A majority (53%) of workers were single, and 35% were married. 47% of workers considered themselves as middle class, and 41% working class. In terms of education level, 44% had a bachelor's degree, 36% some college experience or an associates degree. Two-thirds (63%) of workers had no children, and most lived in a single (25%) or two-person (31%) household. Half (48%) our workers lived in a suburban setting, the remaining half was evenly split between rural and urban. Almost all (94%) of our workers had spent 10 or more years in the U.S.

#### A.7 Demographics and Annotations

We analyze the demographic variation in RoT and action annotations, using a set of 400 RoTs that were annotated by 50 workers each. In addition to the demographic variables described in §A.6, we also consider the political leaning of the state in which the worker resides (self-reported), by assigning each state a value based on the state-level voting patterns in the last four national elections (yielding five-point scale from 100% republican to 100% democratic).

For our analyses, we run a generalized linear model regressing the RoT categories on all  $z$ -scored demographic variables, and report the  $\beta$

	RoT Agree- ment	Action Agreement	Cultural Pressure	Social Judg- ment
Gender (M: 0, F: 1)	0.070***	0.104***	<i>n.s.</i>	<i>n.s.</i>
Urbanness	0.065***	0.085***	<i>n.s.</i>	<i>n.s.</i>
Education	0.022**	0.037***	<i>n.s.</i>	0.025***
Politics (rep: 0, dem: 1)	0.052***	0.075***	0.023**	<i>n.s.</i>
Household size	0.059***	0.080***	<i>n.s.</i>	<i>n.s.</i>
Social class	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>
Income	-0.027*	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>
Age	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>

Table 5: Correlations between worker demographics and categorical RoT annotations, Bonferroni corrected for multiple comparisons ( $p < 0.0001$ : \*\*\*,  $p < 0.001$ : \*\*,  $p < 0.01$ : \*).

coefficients from that model. In our action moral judgment analyses, we control for actions; for action agreement, we control for the action and the moral judgement; for the RoT agreement and action pressure, we control for individual RoTs. Our results for categorical RoT annotations are shown in Table 5.

**Agreement (RoT and Action)** The projection of how many people agree with the judgement is correlated with various demographic characteristics. Specifically, judgments of actions, being a woman and living in an urban setting was most strongly correlated with ascribing high agreement to the judgment. Other associations include higher education, household size, and inferred political leaning based on state of residency.

For RoT agreement, we find similar but weaker associations. Additionally, we find a small correlation between income and social class and ascribing higher agreement.

**Cultural Pressure** The only variable correlated with feeling culturally pressured is the political leaning of the state where workers are located, though the effect is small.

**Social Judgment** Similar to action agreement. Effects are somewhat weaker, but workers being women, highly educated, or younger are associated with selecting higher (better) judgment to actions.

## B Experimental details

**Generative Models** We use the Transformers package (Wolf et al., 2019) to implement our models. We train all the models for a single epoch with a batch size of 64, with the random seed 42. Each input and output sequence

is prefixed with a special token indicating its type (e.g. [attrs], [rot], [action]). We also define a special token for each attribute value (e.g. <morality-ethics>, <bad>, <all>, <against>). We initialize the special token embeddings with the embedding of their corresponding words, taking the average for multiword expressions. For example,  $\vec{v}_{\langle \text{bad} \rangle} = \vec{v}_{\text{bad}}$ ,  $\vec{v}_{\langle \text{morality-ethics} \rangle} = (\vec{v}_{\text{morality}} + \vec{v}_{\text{ethics}})/2$ .