MOKA: Moral Knowledge Augmentation for Moral Event Extraction

Xinliang Frederick Zhang[1], **Winston Wu**[2], **Nick Beauchamp**[3], and **Lu Wang**[1]

[1]Computer Science and Engineering, University of Michigan, Ann Arbor, MI
[2]Department of Computer Science, University of Hawaii at Hilo, Hilo, HI
[3]Department of Political Science, Northeastern University, Boston, MA
[1]{xlfzhang,wangluxy}@umich.edu
[2]wswu@hawaii.edu, [3]n.beauchamp@northeastern.edu

Abstract

News media employ moral language to create memorable stories, and readers often engage with the content that align with their values. Moral theories have been applied to news analysis studying moral values in isolation, while the intricate dynamics among *participating entities* in shaping moral events have been overlooked. This is mainly due to the use of obscure language to conceal evident ideology and values, coupled with the insufficient moral reasoning capability in most existing NLP systems, where LLMs are no exception. To study this phenomenon, we first annotate a new dataset, MORAL EVENTS, consisting of 5,494 structured annotations on 474 news articles by diverse US media across the political spectrum. We further propose MOKA, a moral event extraction framework with MOral Knowledge Augmentation, that leverages knowledge derived from moral words and moral scenarios. Experimental results show that MOKA outperforms competitive baselines across three moral event understanding tasks. Further analyses illuminate the selective reporting of moral events by media outlets of different ideological leanings, suggesting the significance of event-level morality analysis in news. Our datasets and codebase are available at https://github.com/launchnlp/MOKA.

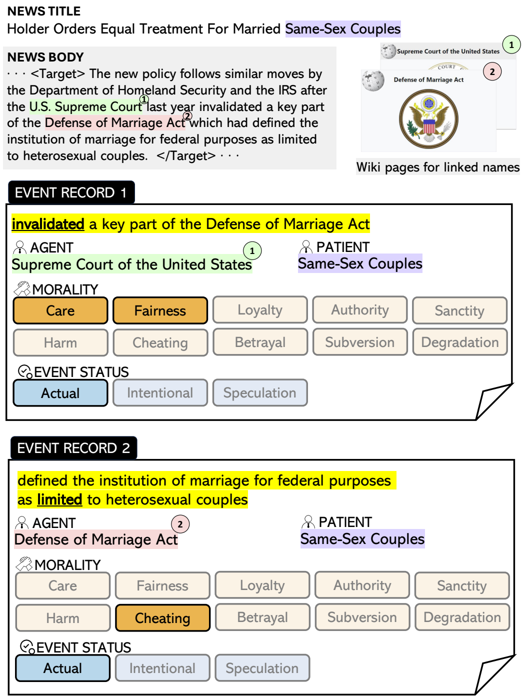

Figure 1: Sample moral events annotations for the target sentence from MORAL EVENTS dataset. Event participants are annotated per Wikipedia pages if applicable. In each event record, the **event trigger** is a single word in an event span, and it might embody multiple moralities. Moral event extraction is challenging due to several reasons: implicit participants (e.g. same-sex couples in Event Record 1) may not be mentioned in the target sentence, and understanding the relations among the participants is necessary to correctly infer the morality.

1 Introduction

News media often frame news stories to further a particular ideological viewpoint, in order to create narratives that engage target readers with aligned values (Scheufele, 1999). Due to the profound connection between ideology and moral values (Haidt and Graham, 2007; Haidt et al., 2009), such narratives wield significant influence by tactfully presenting ideological content (Lakoff, 2010) and reshaping readers' opinions (Haidt et al., 2009; Feinberg and Willer, 2015), without overtly revealing their inherent bias. However, existing work predominantly focuses on examining moral values at a coarse-grained level, often confined to document-level analysis (Haidt et al., 2009; Fulgoni et al., 2016; Mokhberian et al., 2020), or in isolated, context-agnostic examination (Frimer et al., 2019), which neglects the importance of how participating entities would alter readers' moral perspectives. This phenomenon is attributed to the fact that existing computational tools, incapable of subtle inference and discerning tacit moral language, are grappling with the challenge posed by media outlets that frequently obscure their inherent ideology and values. Large language models (LLMs) are no exception. Despite their superior performance on numerous tasks, they still encounter formidable challenges in the face of moral scenario reasoning (Jiang et al., 2021), lack a firm moral stance (Zhou et al., 2023; Krügel et al., 2023) and yield suboptimal performances, e.g., ChatGPT performs merely comparably to a dictionary-based approach as demonstrated in Table 2. Therefore, there is an imperative need for a novel approach to moral understanding – embarking on the development of NLP models equipped with enhanced moral reasoning capabilities to navigate and address the obstacles that impede a grasp of morality.

Moral Foundation Theory (MFT; Haidt and Graham, 2007; Graham et al., 2009, 2013), a prominent social psychology theory that posits five moral foundations with each containing two polarities of virtue and vice, e.g., `Care/Harm` or `Fairness/Cheating`, has been widely used for analyzing both mainstream news (Hopp et al., 2021) and social media (Lin et al., 2018; Hoover et al., 2020; Trager et al., 2022) content. In particular, news tells complex stories that contain multiple people and events along with interactions among them. **Events**, as the building blocks of media narratives, capture such relationships. The participants in the events, the ordering of them, and the selection of events themselves have been shown to be useful for crafting impactful news articles (White and Ventola, 2002; Van Dijk, 2013; Bourgeois et al., 2018). In this work, we study morality and moral reasoning at the level of event, enabling fine-grained structural analysis, capturing the nuances of relationships between participants performing moral actions, and uncovering the deeper layers of ethical dimensions intrinsic to news narratives. To this end, we first propose the concept of **moral events**, which capture the interaction among moral participants, such as moral agents and moral patients (Gray and Wegner, 2009), as demonstrated in Figure 1. We then study the problem of *structured* **moral event extraction**, which enables analyzing at a fine-grained level how the choice of events in news articles and the context in which the events occur together carry moral implications, form effective news stories, and sway readers' perceptions.

The task of moral event extraction in news articles presents two unique challenges. First, moral judgment in news reporting is obscure, though mostly can be revealed with the help of context and background knowledge. Morality often involves a subjective judgment of right-doing vs. wrong-doing, but news must maintain some semblance of objectivity (Westerståhl, 1983). Therefore, when a news article cannot openly praise or criticize someone, the writer often subtly informs the reader of their opinion towards the subject by selecting which actions the subject takes toward other subjects. Second, identifying the moralities requires an understanding of the relationship between participants in the moral action, which can be learned from external moral knowledge. For example, in Figure 1, correct identification of Event Record 1 needs to take into account the longer context (e.g., the title) or background knowledge (that the Defense of Marriage Act governs same-sex couples) to identify the patient who is affected by the *invalidated* event. Identifying the morality of *invalidated* requires knowing that although the act of invalidating usually has a negative connotation and might imply `Harm` on a surface level, in this case it is actually beneficial to the patient and in fact embodies `Fairness` and `Care` towards same-sex couples. These challenges motivate our approach to moral event extraction: a modeling framework that incorporates external moral knowledge at various granularities.

Our paper makes the following contributions. First, we define a new schema for *moral events* which is grounded in MFT and existing work in linguistics. We then propose the task of *moral event extraction*: given unstructured text, detect morality-bearing event triggers, extract participating entities, and identify underlying moralities.

Second, to support this task, we collect a large dataset **MORAL EVENTS**, consisting of moral event annotations of news articles from media outlets with diverse ideologies in the United States. This dataset is unique in that the annotations are conducted on multiple news articles about the same story, allowing us to analyze differences in how news outlets with different political leanings report moral events. The annotations of moral agents and moral patients go beyond surface mentions and syntactic constraints, capturing implicit participants in moral actions.

Third, we develop **MOKA**, a generative framework for moral event extraction with **MO**ral **K**nwoledge **A**ugmentation. Capitalizing on the recent success of retrieval-augmented language understanding (Lewis et al., 2020; Févry et al., 2020; Izacard et al., 2022), MOKA identifies and integrates knowledge derived from different levels of granularity, including moral word-based examples and scenario-based moral knowledge. Additionally, to support MOKA pre-training, we crawl a bank of 344k moral word-based examples, MORALITY BANK, in accordance with experts-validated morality lexicons (Graham et al., 2009; Frimer et al., 2019). Extensive experiments showcase the usefulness and robustness of MOKA over strong baselines, including SOTA models on the event extraction task and ChatGPT (*gpt-3.5-turbo*). Our results show that external moral knowledge is essential for language models to excel at moral event extraction and ethics-related moral reasoning in general. Our further investigation on the moral event reporting reveals several interesting findings, including prominent left-right asymmetries where `Right-to-Left` moral events are more prevalent than the reverse regardless of underlying morality and media ideology as well as the unique behavior of Centrist media, which overwhelmingly focuses on moral events enabled by right-leaning entities. These insights illuminate the importance of event-level morality analysis in news.

2 Related Work

There are a few studies in understanding morality in news, though they remain relatively limited in scope. Many studies in the social sciences have linked political ideology with morality (Graham et al., 2009; Wolsko et al., 2016). In the news, studies have shown that moral and emotional news spreads faster through social media (Brady et al., 2017). In addition, one's morality can also determine views on topics often reported in the news, such as vaccine hesitancy (Amin et al., 2017), violent protests (Mooijman et al., 2018), and war (Koleva et al., 2012). In this section, we focus on related work for morality tasks in NLP and methods for moral foundation prediction and event extraction.

2.1 NLP Benchmarks for Morality

There has been a recent surge of interest in NLP research centered around morality, including moral norms, ethical judgment, and social bias. Most work on investigating morality is based on Moral Foundation Theory (MFT; Haidt and Graham, 2007; Graham et al., 2009, 2013), a prominent social psychology theory that posits five moral foundations, each with two polarities of virtue and vice: `Care/Harm`, `Fairness/Cheating`, `Loyalty/Betrayal`, `Authority/Subversion` and `Sanctity/Degradation`. Under this theory, these are claimed to be culturally-universal moral foundations upon which people intuitively base their virtues and institutions (Hu et al., 2020).

Many recently annotated morality **datasets** are limited to social media text, including Twitter (Brady et al., 2017; Johnson and Goldwasser, 2018; Wang and Inbar, 2021; Stranisci et al., 2021; Hoover et al., 2020) and Reddit (Lourie et al., 2021; Alhassan et al., 2022; Trager et al., 2022). Others combine social media text with crowdsourced data to study morality-related topics such as offensiveness (Sap et al., 2020), rules of thumbs (Forbes et al., 2020), knowledge of ethics (Hendrycks et al., 2021), branching narratives (Emelin et al., 2021), and judgments on everyday situations (Jiang et al., 2021). Whereas, the language of social media is usually short and lacks rich context. On the contrary, only a handful of existing work studies morality in news articles at the word-level (Mokhberian et al., 2020) or topic-level (Fulgoni et al., 2016; Shahid et al., 2020). However, their analysis is at a coarse-grained level or on a limited number of news topics. In contrast, we collect a high-quality corpus of moral events from a wide range of news sources across the US political spectrum, that supports the study of how the interplay of events and moralities is used to craft effective news articles.

Most similar to our work are morality frames (Roy et al., 2021) and the eMFD (Hopp et al., 2021). Although morality frames also capture participants in moral actions, they do not account for implicit patients affected by the moral action, and whose mention does not occur in the text span. The eMFD released a set of news articles coded for morality as a supplement to their morality lexicon, while they only annotate text spans and embodied moralities. In contrast, our work contains fine-grained structured event annotations including participants and linguistic features such as negation and factuality.

2.2 Moral Foundation Prediction

Many papers surveyed above also present experiments on classifying or predicting morality, a task also known as **moral foundation prediction**. Some treat the prediction of morality as a categorical classification task, accomplished by fine-tuning pre-trained language models (Lin et al., 2018; Alhassan et al., 2022). Others reformulate classification as template-based natural language generation

(Jiang et al., 2021; Forbes et al., 2020). Moreover, while existing work focuses on predicting a moral label at the context-agnostic word-level (Graham et al., 2009; Frimer et al., 2019) or document-level (Haidt et al., 2009; Mokhberian et al., 2020), our models predict fine-grained structured moral events using the combination of the context in which these events occur and external moral knowledge, allowing us to capture the nuances of moral actions involving different participants as well as deeply understand the role in which morality plays in shaping the narrative of news media.

2.3 Event Extraction

Our work follows a long line of research in event extraction, including two important tasks: event detection (ED) and event argument extraction (EAE). ED is defined as identifying event trigger that best describes an event, i.e., change of state (Chen et al., 2018; Lou et al., 2021; Deng et al., 2021), while EAE has the goal of extracting a phrase from text that mentions an event-specific attribute labeled with a specific argument role (Du and Cardie, 2020a; Li et al., 2021; Parekh et al., 2022).

ED is commonly modeled as sequence labeling (Li et al., 2021), question answering (Du and Cardie, 2020b), or template-based conditional generation (Hsu et al., 2022). On the other hand, three major approaches have been developed to tackle the more challenging EAE task: sequence labeling (Chen et al., 2015; Nguyen et al., 2016; Yang et al., 2019; Du and Cardie, 2020a) where global features have been incorporated to constrain the inference (Lin et al., 2020); question answering (Du and Cardie, 2020b; Tong et al., 2022), where models additionally incorporate ontology knowledge about argument roles; and generative models for structured extraction (Li et al., 2021; Yang et al., 2021; Lu et al., 2021; Du and Ji, 2022). Most recently, an emerging area of research is using large language models (e.g., chatGPT) to perform EAE, but the performance is not on par with specialized EE systems (Li et al., 2023).

Our work proposes a new understanding task, *moral event extraction*, a combination of ED and EAE with special focus on morality-bearing events. Unlike conventional event extraction, where each event type has its own event schema, we define a universal schema for moral events grounded in Moral Foundations Theory and linguistics. On the modeling side, to the best of our knowledge, we are the first that manages to explicitly incorporate *external moral knowledge* for event extraction tasks, and moral understanding in general.

	Virtue	Vice	Proportion
Care/Harm	1,348	2,060	51.6%
Fairness/Cheating	531	453	14.9%
Loyalty/Betrayal	329	257	8.9%
Authority/Subversion	1,140	418	23.6%
Sanctity/Degradation	19	46	1.0%
Total	3,367	3,234	100.0%

Table 1: Distribution of moralities in moral event annotations in MORAL EVENTS. Each event has four attributes including agent, patient, event status and morality (Figure 1). The five moral foundations stem from MFT (Haidt and Graham, 2007). Numbers in *Virtue* and *Vice* columns are raw counts of annotated moralities.

3 MORAL EVENTS Collection and Annotation

3.1 Moral Event Schema

We define a new structured schema for a **moral event** which represents a moral action. A moral event encompasses moral agents, moral patients, a morality-bearing event span and event trigger, embodied morality, and event status. A moral event is performed or enabled by **moral agents** and affects **moral patients**. Moral agents and patients usually possess moral agency, the capability of doing things right or wrong (Gray and Wegner, 2009), and a moral event may have multiple moral agents and patients. The **moral event span** is a contiguous sequence of words in the text which concisely contains the event and carries stand-alone meaning. This span embodies one or more **moralities** in MFT: a moral evaluation will arise when the patient is harmed or helped by the action enabled by the agent (McPherson, 1984; Gray and Wegner, 2009; Hopp et al., 2021). Note that moral patients may be *implicit*: they do not have to be mentioned in the event span. In line with ACE 2005 (Doddington et al., 2004) and the LDC annotation guideline[1], the moral event also includes an **event trigger** that can best characterize the moral event.

To assist the investigation into the linguistic phenomenon of moral events, an event also has an **event status** which describes the factuality of an event, i.e., whether an event is *actual* or *non-actual* (Saurí and Pustejovsky, 2009; Lee et al., 2015). We further divide *non-actual* into *intentional* and

[1] www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf

speculative events, where *intentional* describes an event that is being planned or intended to happen, while *speculative* represents an event that may happen, usually speculated by someone who is not an event participant (Demner-Fushman et al., 2008; Kolhatkar et al., 2019; Mahany et al., 2022).

3.2 Annotation Process

We create our dataset, MORAL EVENTS, using the following process. We first sample a subset of 87 news stories from *SEESAW* (Zhang et al., 2022), a dataset consisting of U.S. news articles published between 2012 and 2021, where each story contains 3 articles on the same event but reported by media of different ideology. To complement the coverage and supplement this set with recent news, we further collect a new set of news article triplets from AllSides.com focusing on important issues in 2021 and 2022, including abortion, gun control, and public health. We extract text from these articles using Newspaper,[2] and clean all articles by removing boilerplate text and embedded tweets

Next, we hire six college students with native English proficiency to annotate moral events. Each article is annotated by at least two people to ensure annotation quality, and each annotator has access to all three articles in a story to maintain a non-biased view. We list major steps of annotating a single article, with a detailed protocol in Appendix C.

1. The annotator first reads an article and then identifies agency-bearing entities that are participants in moral events. An entity may be of type Person, Organization, Geo-Political, or Other.[3] Entities are coded by their canonical name, i.e., the name listed in Wikipedia, For example, mentions of “President Trump” or “Trump” are coded as “Donald Trump”.
2. For each sentence, the annotator identifies moral events and their attributes following the event schema defined in §3.1.
3. Finally, the annotator determines the 5-way ideological leaning of the article.

After an article is annotated (i.e., first pass), we proceed to the second pass to improve the annotation quality. Specifically, we employ two distinct approaches: (a) an article is **revised** by a second person who corrects existing annotations and adds missing ones; (b) a second person annotates the full article from scratch following the procedures above, and a third person conducts an adjudication process to **merge** and resolve annotation conflicts. 83% of articles are revised with approach (a) while 17% adopt approach (b). This two-pass annotation style effectively balances quality and efficiency.

[2] https://github.com/codelucas/newspaper/

[3] *Other* includes religions (e.g., People of Faith) and topics (e.g., Homeland Security) among others.

3.3 Statistics

MORAL EVENTS includes 474 news articles from 158 stories, published by 63 different media outlets (26 left, 18 center, and 19 right). On average, each article contains 28.3 sentences and 658.6 words, and 9.8 entities and 11.6 events, underscoring the annotation density of MORAL EVENTS. In total, there are 5,494 **event annotations**. The articles cover 38 salient topics reported from 2012 to 2022 (inclusive), and includes 1,952 distinct entities (see Table A1 for top-30 frequent entities). **Entity types** cover People (62.4%), Organization (20.4%), Geo-Political (9.6%), and Others (7.6%), showing the diversity of entity annotations.

3.4 Quality Control

We ensured the quality of the annotations at multiple steps in the collection process. We briefly present the key points below, with detailed information in Appendix A. All annotators participated in training sessions before and during the annotation process. We have found high inter-annotator agreements (shown in Table A2) measured by Krippendorff’s alpha (Krippendorff, 2011) on the annotated attributes of the moral events, where 1 is complete agreement, -1 is complete disagreement, and 0 is chance agreement. The high agreement for merged articles validates our decision to revise annotations. In addition, for merged articles using approach (b), annotator agreement on the article’s ideological leaning is 0.76. Upon comparing annotated article leanings with AllSides’ media-level ideology labels, we find a significant improvement in the matching rate, rising from 70.9% to 76.4%, after the adjudication process.

4 Models

We now define the task of *moral event extraction*, where models must extract structured moral events from unstructured texts. In the same vein as mainstream event extraction, we decompose moral event extraction into event detection and event argument extraction sub-tasks, but with a special focus on morality. To tackle these tasks, we develop a new

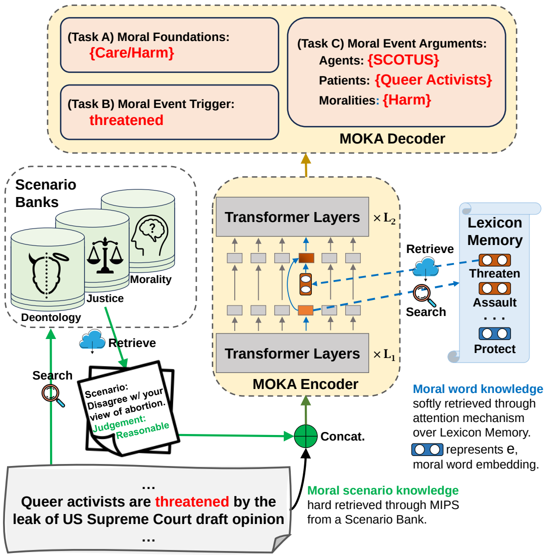

Figure 2: Overview of our moral knowledge augmented generative framework, MOKA, for (downstream) moral event extraction. The left part displays the process of retrieving and combining relevant scenarios through the moral-scenario augmentation module. The right part shows how the moral-word augmentation module integrates moral word knowledge through attention-based soft retrieval. Embeddings for *Threaten* and *Assault* are colored in red and they embody `Harm`, while *Protect* is in blue and it embodies `Care`. The agent "Supreme Court of the United States" is shortened as "SCOTUS" in the diagram. {} indicates there can be multiple instances for one field. The inputs to MOKA are a 4-sentence segment and a full moral event span for tasks A and C, but a 4-sentence segment only for task B.

framework, MOKA (Figure 2), which incorporates external moral knowledge into pre-trained language models at two levels: lexical-based moral word knowledge (§4.1) and example-based moral scenario knowledge (§4.2). We instantiate MOKA with *Flan-T5-large*, noting that MOKA is also compatible with models of different architectures (encoder-decoder and causal LMs).

Moral Knowledge Augmentation. To harness moral reasoning, it is critical to have a priori knowledge of necessary moral principles, just like a person of practical wisdom would (Leibowitz, 2014; Schwartz and Sharpe, 2011). However, large language models' access to moral facts is usually limited due to the lack of moral knowledge seen in the pretraining corpus (Jiang et al., 2021), although injecting morality into models has long been a question for debate (Wallach and Allen, 2008; Bigman and Gray, 2018; Awad et al., 2018).

Models with a retrieval mechanism to access explicit non-parametric memory can provide provenance for their decision-making process and thus perform more robustly (Lewis et al., 2020). These retrieval mechanisms have so far been mainly investigated for certain knowledge-intensive tasks, such as entity-intensive question answering (Glass et al., 2022; Chen et al., 2023). Hence, our work takes the first step to marry a retrieval component with moral knowledge to improve moral event understanding.

4.1 Moral Word Knowledge

MORALITY BANK Construction. Unlike open domains where an existing knowledge base (KB) is always available such as the WikiData,[4] no such KB exists in the realm of moralities. Seeing that, we hypothesize that *an utterance embodies a morality if it contains a mention which embodies a specific value*, where a moral mention is an occurrence of a moral word.[5] Therefore, we union two validated morality lexicons, MFD (Graham et al., 2009) and MFD2.0 (Frimer et al., 2019), and obtain 891 *moral words*. We scrape example sentences that contain at least one moral mention from four authoritative online dictionaries.[6] We limit the sentence length between 5 and 80 words, totaling $334k$ sentences. 95% of example sentences are used for pre-training and the rest for validation. Samples from `Morality Bank` are shown in Table A5.

Lexicon Memory Access. Similar to Févry et al. (2020) and Verga et al. (2021), we define the *Lexicon Memory* $\mathbf{E}$ as a matrix containing an embedding for each moral word. For each word, we initialize and freeze its embedding, $\mathbf{e}$, by averaging the contextualized representations of its mentions in all example sentences. During encoding, a moral mention is tagged with a special token pair (<Morality>, </Morality>). Access to the Lexicon Memory is triggered when encountering the morality special tokens as follows. MOKA takes as a query vector $\mathbf{h}_q$, the averaged representation of the special token pair and the moral mention in between. $\mathbf{h}_q$ is then used to retrieve relevant moral knowledge $\mathbf{h}_m$ from the Lexicon Memory via a single-head attention mechanism, $\mathbf{h}_m = \text{Attn}(\mathbf{h}_q, \mathbf{E})$, where $\text{Attn}(\cdot,\cdot)$ is the same cross-attention mechanism as implemented in

[4] www.wikidata.org/wiki/Wikidata:Main_Page

[5] A moral word is a unique entry in the morality lexicon and the base form of moral mentions. For example, mentions like *threatening* and *threatened* are mapped to *Threaten* entry.

[6] Cambridge (UK & US sites), Merriam-Webster, Dictionary.com, and YourDictionary.com which itself is a compilation of 16 authoritative sources.

Vaswani et al. (2017). Finally, $\mathbf{h}_m$ is added to $\mathbf{h}_q$, the sum of which is normalized before being fed to the next layer. Following Févry et al. (2020), we interleave standard transformer layers with the Lexicon Memory access layer at a lower layer, which is 8th layer ($L_1 = 8$ and $L_2 = 16$ in Figure 2).

Moral Word Knowledge Pre-training. The pre-training objective is a combination of language modeling ($\mathcal{L}_{LM}$), morality prediction ($\mathcal{L}_{MV}$), and moral word linking ($\mathcal{L}_{MWL}$) and moral label association ($\mathcal{L}_{MLA}$), each is described below. Language modeling is employed to train MOKA to denoise corrupted sentences, to familiarize itself with moral language usage. Morality prediction is introduced to provide a direct signal to train MOKA to uncover the morality embodied in a morality-bearing input sentence. Note, to prevent MOKA from learning shortcuts, the *seed word* used to scrape the input sentence is always masked.[7] Lastly, we propose two new training objectives to train the memory access mechanism effectively. For each moral mention in the text, moral word linking objective guides MOKA to identify the corresponding moral word by learning to maximize the attention score over the correct entry, e.g., *Threaten* entry in Figure 2. Furthermore, we design a **moral label association** objective to promote MOKA's capability of associating a moral mention and its embodied morality(s). To this end, MOKA is trained to, for each morality embodied by the mention, maximize the summation of attention scores over all moral words in $\mathbf{E}$ that share the same morality. To support the training of moral words that might be associated with more than one morality, we use *multi-label margin loss*, as described in Equation (4). This objective has merits in flattening scores over target moralities and mitigating saturated gradients compared to standard cross-entropy. Detailed mathematical formulations of Lexicon Memory access are in Appendix B.

Our work differs from existing work using entity memory (Févry et al., 2020; Verga et al., 2021) in three aspects. First, moral concepts and stances are more abstract than concrete entities. No KB exists in the context of morality, so we curate MORALITY BANK, transforming morality-bearing sentences into a structured knowledge base. In addition, unlike entity memory which can utilize entity-linking tools out-of-the-box, we rely on designed objectives $\mathcal{L}_{MWL}$ and $\mathcal{L}_{MLA}$ to enable memory access.

4.2 Moral Scenario Knowledge

Moral Scenario Bank Compilation. Although fundamental theoretical moral theories are prescriptive and rule-based, we depart from this approach and adopt example-based, descriptive moral scenarios. As pointed out by Jiang et al. (2021), while human kinds can directly understand abstract moral principles without the need for interaction with concrete moral scenarios, those principles are too perplexing for machines. Consequently, we guide MOKA to develop its moral sense by immersing it in real-world moral scenarios. To achieve this, we compile a suite of Moral Scenario Banks by incorporating three large-scale ethics-related datasets: `Delphi` (Jiang et al., 2021), `Social Chemistry` (Forbes et al., 2020), and `ETHICS` (Hendrycks et al., 2021). We convert them into 7 moral scenario banks, and show detailed statistics and examples in Table A6.

Scenario Retrieval. For each moral scenario bank, we convert (scenario, label) pairs into a key-value pairs. Then, we encode all keys into dense vectors using the Flan-T5 encoder, which ensures an isomorphic embedding space between searching and reasoning. To implement efficient maximum inner-product search (MIPS), we create a ScaNN index (Guo et al., 2020) and search top-K relevant scenario pairs using dot product similarity between the query and keys, i.e., scenarios. We set $K = 3$ in this study. Retrieved scenario pairs are concatenated together with the input, which is then fed into MOKA encoder, as shown in Figure 2.

Moral Scenario Knowledge Pre-training. We pre-train MOKA on moral scenario banks to improve its moral reasoning by guiding it to navigate the complex interplay of diverse moral principles within real-world scenarios. The task is formulated as given an input *scenario* and a set of relevant scenario pairs in the form of (scenario, label),[8] such as ("enjoying your life with your family", "morally good"), MOKA is trained to generate a desired output.

To further improve the digestion of retrieved scenarios and enhance the encoder's moral reasoning capabilities, we introduce a new pre-training objective – Retrieved Label Masking (RLM). Specifically, we randomly mask out the label of one re-

[7] We also experimented with not always masking a seed word, which yields inferior results.

[8] During pre-training, the input and retrieved scenarios are from the same source, so we remove the identical one from the retrieved set.

trieved example and apply MLM objective to recover this label. Through explicitly training the encoder to discern the associated moral label, it can help MOKA from collapsing to simply memorizing retrieved labels and making trivial inferences.[9]

This approach is in line with retrieval augmented generation, in which existing work mainly empowers a language model with a retriever to fetch knowledge items in text form from an external knowledge bank (Lewis et al., 2020; Fan et al., 2021; Shi et al., 2022; Izacard et al., 2022). However, most existing work is limited to the use of a single knowledge source, where the only exception is a contemporaneous work by Pan et al. (2023). On the contrary, MOKA embraces multiple moral knowledge sources under different moral scenarios.

4.3 Downstream Moral Event Extraction

Figure 2 depicts the flow when MOKA is deployed for downstream moral event extraction. At fine-tuning stage, we first take the input passage to retrieve K-scenarios (K=3) from the moral scenario bank MOKA is pre-trained on. We then combine retrieved scenarios with the original input to create a moral knowledge-enriched input. Next, we tag moral mentions on the fly, and then follow the Lexcion Memory Access steps outlined in §4.1 to integrate moral word knowledge. Eventually, MOKA is trained to generate an end-task-specific output with three training objectives: $\mathcal{L}_{\text{FT}} = \mathcal{L}_{CE} + \mathcal{L}_{MWL} + \mathcal{L}_{MLA}$, where $\mathcal{L}_{CE}$ is a standard cross-entropy loss applied to the decoder, and $\mathcal{L}_{MWL}$ and $\mathcal{L}_{MLA}$ are the same memory-access losses as described in §4.1.

5 Experiments

5.1 Tasks and Datasets

We conduct holistic evaluations on three different moral event extraction sub-tasks using two datasets, our curated MORAL EVENTS and eMFD corpus (Hopp et al., 2021). The input in all tasks is a 4-sentence document which includes a *target sentence*, a preceding sentence, a succeeding sentence, and a title.[10]

Task A: Moral foundation prediction. Given a document and *one* moral event span, the goal of models is to make a 5-way judgment on the moral foundation for the given moral event.

Task B: Moral event trigger detection. Similar to mainstream event trigger detection, given a document, the models are expected to detect moral event triggers from the target sentence.

Task C: Moral event argument extraction. Conditioned on a document and *one* moral event span, the model output is in a triplet format including identifications of moral agents and patients in addition to a 10-way morality prediction. This task requires profound moral reasoning skills to excel and understand the interplay between participants and moralities.

We include eMFD corpus (Hopp et al., 2021) as another benchmark, which does not annotate event attributes but moral foundations only, marking it only applicable to Task A. Also, since each document might embody more than one foundation or morality, we follow existing research on approaching multi-label classification with generative models (Yang et al., 2018; Yue et al., 2021; Chai et al., 2022) by consistently linearizing foundations or moralities as a sequence in our experiments.

Model	MORAL EVENTS F1	MORAL EVENTS Acc.	eMFD Corpus F1	eMFD Corpus Acc.
Baselines				
Dictionary-based counting (Brady et al.)	45.8	56.8	33.0	52.0
RoBERTa (large; Liu et al.)	63.6	82.6	28.7	69.0
POLITICS (base; Liu et al.)	62.7	82.4	29.0	68.8
ChatGPT (zero-shot; Li et al.)	41.2	69.9	31.9	66.9
ChatGPT (few-shot; Li et al.)	46.9	75.6	30.5	69.1
Flan-T5 (large; Chung et al.)	62.0	83.6	25.4	68.4
MOKA with moral word knowledge augmentation only				
Pretrain on `Morality Bank` only	63.6	83.9	27.3	69.0
+ moral word linking	63.9	83.9	27.8	69.0
+ moral label association	64.0	83.9	28.5	69.1
MOKA with moral scenario knowledge augmentation only				
Delphi (moral judgement; Jiang et al.)	63.7	84.1	30.4	70.4
w/ RLM	62.3	83.8	30.1	70.3
Deontology (Hendrycks et al.)	62.5	83.6	30.5	70.5
w/ RLM	62.2	83.5	30.4	70.4
Social chem (foundation; Forbes et al.)	62.2	83.7	32.4	70.6
w/ RLM	64.1	84.0	32.5	70.7
MOKA with dual moral knowledge augmentation				
Delphi (moral judgement)	63.3	83.6	32.9	70.7
w/o moral label association	63.9	84.1	32.1	70.6
Deontology	64.0	84.0	32.9	70.8
w/o moral label association	64.2	84.0	**34.3**	**71.1**
Social chem (foundation)	**65.3**	**84.3**	33.7	71.0
w/o moral label association	64.1	84.0	33.4	71.0
Improvements over best baseline	2.7%	0.8%	3.9%	2.9%

Table 2: Weighted F1 and accuracy results on MORAL EVENTS and eMFD Corpus (Hopp et al., 2021) for task A (average of 5 runs). Best results are in **bold**. Our MOKAs that outperform all baselines are highlighted on a scale of 5 red shades based on the relative improvements over the *strongest baseline*. Our MOKAs augmented with dual moral knowledge achieves better performances across the board by non-trivial margins. Full results and color scheme explanations refer to Table A4.

[9] In pertaining, we find that retrieved labels match the output label 60% of the time.

[10] eMFD corpus does not contain title, so we include 2 preceding sentences instead.

Model	Task B	Task C				
	Trigger EM	Morality F1	Agent EM	Agent F1	Patient EM	Patient F1
Baselines						
DEGREE (base; Hsu et al.)	45.5	53.0	47.3	58.6	30.1	39.2
DEGREE (large; Hsu et al.)	46.2	54.2	49.2	60.3	30.5	40.3
ChatGPT (zero-shot; Li et al.)	19.5	39.5	30.3	49.8	12.3	23.2
ChatGPT (few-shot; Li et al.)	32.1	38.1	34.2	51.4	20.1	30.6
Flan-T5 (large; Chung et al.)	46.2	53.8	47.5	59.4	30.8	41.2
MOKA with moral word knowledge augmentation only						
Pretrain on `Morality Bank` only	45.3	54.6	47.5	59.9	31.2	41.7
+ moral word linking	45.6	55.9	47.6	59.8	**31.5**	41.7
+ moral word linking + moral label association	46.2	57.0	48.3	60.2	31.3	41.9
MOKA with moral scenario knowledge augmentation only						
Delphi (moral judgement; Jiang et al.)	47.0	57.5	48.5	60.4	30.9	41.4
w/ RLM	47.4	55.6	48.5	60.3	31.2	41.5
Deontology (Hendrycks et al.)	46.1	54.8	49.0	60.9	30.9	41.6
w/ RLM	47.2	56.0	**49.5**	61.2	31.3	**42.1**
Social chem (foundation; Forbes et al.)	46.7	56.5	48.9	**61.4**	31.0	41.4
w/ RLM	47.5	56.0	48.8	60.5	31.0	41.7
MOKA with dual moral knowledge augmentation						
Delphi (moral judgement)	47.4	56.8	48.1	60.3	30.2	40.5
w/o moral label association	46.7	57.2	47.6	60.0	30.2	40.5
Deontology	46.8	**58.2**	47.9	60.3	30.9	41.1
w/o moral label association	**48.1**	57.3	48.2	61.0	30.7	41.1
Social chem (foundation)	46.5	58.1	48.4	61.0	30.5	40.8
w/o moral label association	46.7	57.7	48.2	60.5	30.0	40.1
Improvements over best baseline	4.1%	7.4%	0.6%	1.8%	2.3%	2.2%

Table 3: Results on MORAL EVENTS for tasks B and C, and breakdown of performances by event attributes (average of 5 runs). Best results are in **bold**. Our MOKAs that outperform all baselines are highlighted on a scale of 5 red shades based on the relative improvements over the *strongest baseline*. Our MOKAs augmented with dual moral knowledge achieve consistently better performance on trigger detection and morality inference, while the best results on participant extractions are reached by single-knowledge variants. Full results and color scheme explanations refer to Table A7.

We split MORAL EVENTS by chronological order, and take as the test set 90 news articles published in the 2nd half of 2022. We sample a subset of articles from the eMFD corpus and partition them randomly on the article level. Table A3 shows the detailed statistics of splits on both datasets.

5.2 Baselines

For the well-established Task A, following Alhassan et al. (2022), we compare with encoder-only models: RoBERTa (Liu et al., 2019) and POLITICS (Liu et al., 2022), a RoBERTa-based model that is continually trained on 3 million news articles with triplet loss. We also include a dictionary approach as a baseline (Brady et al., 2017), where the moral foundation is determined by the presence of moral words defined in validated morality lexicons (Graham et al., 2009; Buttrick et al., 2020).[11] On the other hand, Task B and C are two novel tasks introduced in this work to study different aspects of moral events, we thus follow the literature on event extraction (Parekh et al., 2022; Li et al., 2023), and compare with a SOTA baseline, DEGREE (Hsu et al., 2022), a BART-based model that extracts event records using natural language templates. For all three tasks, we compare with Flan-T5-Large model (Chung et al., 2022), which are fine-tuned on the downstream evaluation sets only, and adapt prompts designed in Li et al. (2023) to test the capability of ChatGPT, *gpt-3.5-turbo*. Our adapted prompts are shown in Table A8.

5.3 MOKA Variants

We consider three variants of MOKA. First, with **moral word knowledge augmentation only**, we experiment with pertaining on `Morality Bank` only with $\mathcal{L}_{LM}$ and $\mathcal{L}_{MV}$ training objectives. We then gradually add our designed objectives: $\mathcal{L}_{MWL}$ and $\mathcal{L}_{MLA}$. For **moral scenario knowledge augmentation only**, we connect MOKA with one Scenario Bank at a time,[12] and test the effectiveness of our suggested RLM objective. Putting all together, with the final **dual moral knowledge augmentation** variant, we always use the models that are pre-trained with the RLM objective. We further examine MOKA's efficacy with and without $\mathcal{L}_{MLA}$ in the moral word pre-training stage.

[11] To prevent trivially predicting all foundations, we consider the top-3 moral foundations based on counting frequency.

[12] We also experimented with conflating all Scenario Banks together. However, this did not improve performance, but led to a slower training time due to the considerably larger index.

5.4 Results

Evaluation Metrics. We report accuracy and weighted F1 for moral foundation prediction in Task A and morality inference in Task C. For trigger detection (Task B), We consider Trigger F1-score, the same criteria as in prior works (Wadden et al., 2019; Lin et al., 2020). For participants extraction (i.e., agents and patients) in Task C, we follow QA (Rajpurkar et al., 2016, 2018) and EE (Du and Cardie, 2020a; Tong et al., 2022) communities, and adopt span-level Exact Match (EM) and token-level F1 as two evaluation metrics.

Table 2 shows Task A evaluation results. Due to different annotation schemes, performances on MORAL EVENTS and eMFD corpus exhibit distinct trends. Specifically, our MORAL EVENTS aligns with a natural moral foundation distribution as seen in Table 1, while an artifact is evident in the eMFD corpus, where all foundations are roughly equally represented. Encoder-only models show strong performances on both datasets, where RoBERTa-large achieves the best F1 scores on our MORAL EVENTS. ChatGPT, despite its stunning capability in various tasks, also struggles to understand and discern moral foundations. Flan-T5-large, the backbone model in MOKA, yields unsatisfying results, especially on eMFD corpus, due to a lack of ethics-related documents in its pertaining stage (Jiang et al., 2021). Thanks to our designed moral knowledge augmentation, MOKA effectively improves Flan-T5's moral reasoning capabilities by 35% (F1 of 34.3 vs. 25.4). Besides, MOKA achieves the best performances across all metrics and datasets when augmented with both moral word and scenario knowledge. In contrast, the inclusion of single knowledge leads to suboptimal performances, which are generally falling behind by 1 F1 point on our MORAL EVENTS.

Table 3 presents model performances on Task B and C. Similar to Task A, ChatGPT is performing worse than specialized EE systems. Although DEGREE stands as a SOTA model in the general domain, typically surpassing 70% F1 points on event extraction datasets like ACE 2005 (Doddington et al., 2004), its performance is not superior to fine-tuning a Flan-T5 model. This highlights the unique challenges posed by moral event understanding. On the contrary, when equipped with dual moral knowledge, MOKA yields the best performance in terms of trigger detection and morality inference. Particularly, the 7% performance

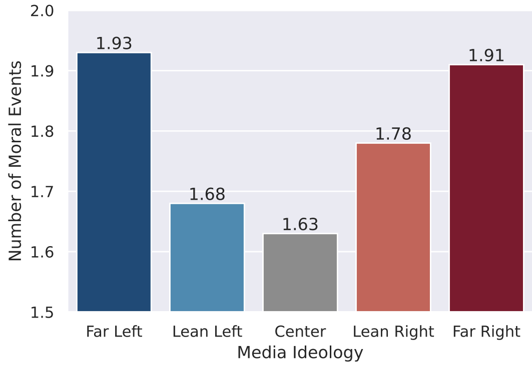

Figure 3: Number of moral events in each 100-word segment. Highly partisan media outlets tend to include more moral language than non-partisan ones.

gain on morality inference over the best model in the literature can be attributed to MOKA effectively assimilating moral knowledge at different granularities after undergoing two stages of moral-knowledge-centric pre-training. That being said, it is worth noting that the best participant extraction results are reached by single-knowledge variants. We posit that this phenomenon might be ascribed to the injected moral knowledge not directly availing the moral reasoning of participants.

Owing to space limitation, we display key results in Table 2 and 3, presenting the performance on one scenario bank from each ethics-related dataset, and full results can be found in Table A4 and A7.

6 Further Analysis

In this section, we further investigate the use of moral language in news media through the lens of selective reporting of moral events. We validate past work showing how different ideologies focus on different moralities (**RQ1**), but then go beyond that to show how the selective reporting of moral *narratives*, in the form of ideological agents and patients plus moral events, reveals far more subtle and asymmetrical forms of bias and selective reporting (**RQ2**).

RQ1: Does moral language usage correlate with media ideology? As seen in Figure 3, more extreme outlets unsurprisingly tend to use more moral language overall, and the least use is by the Centrist media. As seen in Figure 4, the most frequent moral foundation is Care/Harm, in line with findings on social media text (Hoover et al., 2020; Trager et al., 2022). Similarly, both domains use relatively little Sanctity/Degradation, which measures religious, social, and physical purity and disgust, which are rarely reported in the news. However, news media

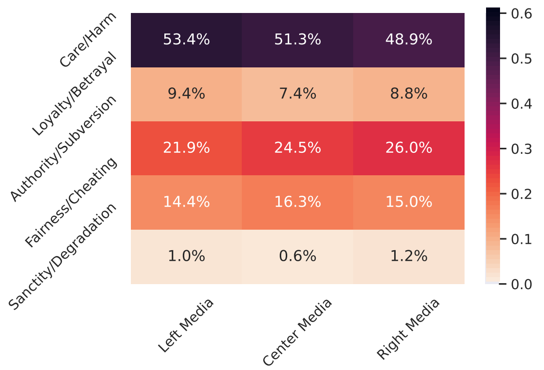

Figure 4: Employed moral foundation distribution by media outlets of different ideologies.

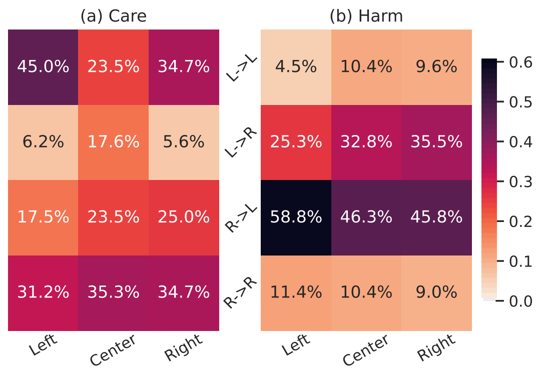

Figure 5: Correlation among agent-patient relationships, media outlet ideologies, and **Care-**/**Harm**-bearing moral events.

use a far higher proportion of Authority/Subversion than social media because much of the news focuses on politicians and other ruling figures. In contrast, social media covers more Fairness/Cheating, due to its greater focus on explicit morality as seen in movements like #BLM and #MeToo, and also due to AITA Reddit forum with its special focus on personal ethical violations (Alhassan et al., 2022).

Within these general tendencies, we find that the left media focuses more on Care/Harm, while the right media focuses more on Authority/Subversion, in line with MFT (Graham et al., 2009). The differences are relatively small compared to the shared tendency to report more Care/Harm by both sides.

RQ2: How is media bias revealed by the selective reporting of agent-morality-patient narratives? Moral narratives are fundamentally constructed out of three elements: an agent, a patient, and an action with an associated morality. To understand how ideology and morality shape the news, we must examine these three elements jointly.

To measure agent and patient ideologies, all entities that appeared in at least two news articles were coded by a domain expert for their partisan leaning on a binary left/right scale,[13] yielding 197 coded entities and $1,253$ associated events. Figures 5 and 6 show the correlations between agent-patient relationship and outlet ideologies for the two most prevalent foundations, Care/Harm and Authority/Subversion. These correlations reveal rich differences between left, right, and center media that do not fall into the simple partisan symmetries that have been posited previously (Gentzkow and Shapiro, 2005; Graham et al., 2009; Gentzkow and Shapiro, 2010).

Within *Care/Harm* (Figure 5), left media report relatively more `Right-harm-Left` events, and vice versa for the right media, which fits well into standard ideological typologies. However, an interesting asymmetry is that all media report more `Right-harm-Left` events than the reverse, including Centrist outlets. Additionally, we are surprised to see that the Centrist outlets are not balanced but instead show a pronounced tendency to report more *Care/Harm* where the `Right` entity is the agent.

For *Authority/Subversion* (Figure 6), we find outlets of both left and right ideologies report far more on `Right-to-Left` *Authority* overall, and the Centrist media are once again far more focused on `Right`-agent events. These asymmetries are even more notable with *Subversion*, where we see right media reporting (disapprovingly) on `Left` entities subverting the `Right` but also (approvingly) on `Right` subverting the `Left`; and the Centrist media are radically skewed and prefer stories of the `Right` subverting the `Left`.

To summarize, mainstream media strive for balance in ideological language, entities, and even expressed values, but when we example agent-value-patient triplets, ideologically-driven biases are immediately evident. For instance, right media prefer to depict the `Left` harming or subverting the `Right`, and vice versa. Meanwhile, we also discern both important *left-right asymmetries*, and also the *distinctive behavior of Centrist media*, which overwhelmingly focuses on `Right` agents. These results illustrate the importance of entity-event-level morality analysis in political news.

7 Conclusion

Motivated by Moral Foundation Theory (MFT), we have embarked on the study of moral event extraction. This is a novel moral reasoning task with the objective of, given unstructured text, produc-

[13]We use `Left` and and `Right` to refer to **entities** with different ideologies along US political spectrum.

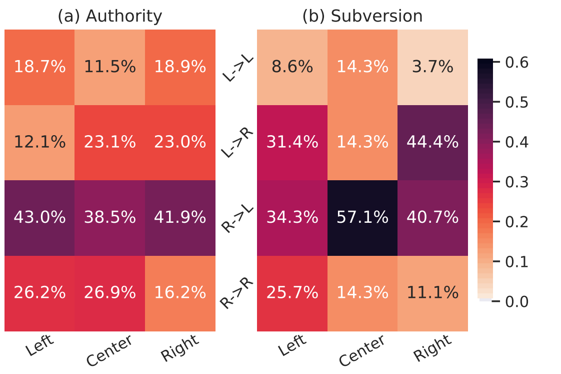

Figure 6: Correlation among agent-patient relationships, media outlet ideologies, and **Authority-/Subversion**-bearing moral events.

ing structural representations for morality-bearing events including detecting triggers, identifying participating entities, and inferring embodied morality. To support this study, we curate a new dataset, MORAL EVENTS, consisting of $5,494$ structured annotations. To tackle this challenging task, we propose MOKA, a moral reasoning-enhanced event extraction framework with moral knowledge augmentation. Specifically, we employ retrieval augmentation by integrating moral knowledge at varying granularities, including knowledge derived from both moral words and moral scenarios. Our further analyses reveal the effectiveness of utilizing moral events to discern ideologically-driven biases even when the outlets report events that seem objectively presented.

References

Areej Alhassan, Jinkai Zhang, and Viktor Schlegel. 2022. ‘Am I the bad one’? predicting the moral judgement of the crowd using pre–trained language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 267–276, Marseille, France. European Language Resources Association.

Avnika B Amin, Robert A Bednarczyk, Cara E Ray, Kala J Melchiori, Jesse Graham, Jeffrey R Huntsinger, and Saad B Omer. 2017. Association of moral values with vaccine hesitancy. *Nature Human Behaviour*, 1(12):873–880.

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature*, 563(7729):59–64.

Yochanan E Bigman and Kurt Gray. 2018. People are averse to machines making moral decisions. *Cognition*, 181:21–34.

Dylan Bourgeois, Jérémie Rappaz, and Karl Aberer. 2018. Selection bias in news coverage: learning it, fighting it. In *Companion Proceedings of the The Web Conference 2018*, pages 535–543.

William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318.

Nicholas R. Buttrick, Robert G. Moulder, and Shigehiro Oishi. 2020. Historical change in the moral foundations of political persuasion. *Personality and Social Psychology Bulletin*, 46:1523 – 1537.

Yuyang Chai, Chong Teng, Hao Fei, Shengqiong Wu, Jingye Li, Ming Cheng, Dong-Hong Ji, and Fei Li. 2022. Prompt-based generative multi-label emotion prediction with label contrastive learning. In *Natural Language Processing and Chinese Computing*.

Wenhu Chen, Pat Verga, Michiel de Jong, John Wieting, and William W. Cohen. 2023. Augmenting pretrained language models with QA-memory for open-domain question answering. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1597–1610, Dubrovnik, Croatia. Association for Computational Linguistics.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176, Beijing, China. Association for Computational Linguistics.

Yubo Chen, Hang Yang, Kang Liu, Jun Zhao, and Yantao Jia. 2018. Collective event detection via a hierarchical and bias tagging networks with gated multi-level attention mechanisms. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1267–1276, Brussels, Belgium. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Dina Demner-Fushman, Sophia Ananiadou, Kevin Bretonnel Cohen, John P. Pestian, Junichi Tsujii, and Bonnie Lynn Webber. 2008. Themes in biomedical natural language processing: Bionlp08. *BMC Bioinformatics*, 9:S1 – S1.

Shumin Deng, Ningyu Zhang, Luoqiu Li, Chen Hui, Tou Huaixiao, Mosha Chen, Fei Huang, and Huajun Chen. 2021. OntoED: Low-resource event detection with ontology embedding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2828–2839, Online. Association for Computational Linguistics.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Xinya Du and Claire Cardie. 2020a. Document-level event role filler extraction using multi-granularity contextualized encoding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8010–8020, Online. Association for Computational Linguistics.

Xinya Du and Claire Cardie. 2020b. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Xinya Du and Heng Ji. 2022. Retrieval-augmented generative question answering for event argument extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4649–4666, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2021. Augmenting transformers with KNN-based composite memory for dialog. *Transactions of the Association for Computational Linguistics*.

Matthew Feinberg and Robb Willer. 2015. From gulf to bridge: When do moral arguments facilitate political influence? *Personality and Social Psychology Bulletin*, 41(12):1665–1681.

Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. 2020. Entities as experts: Sparse memory access with entity supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4937–4951, Online. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Jeremy A Frimer, Reihane Boghrati, Jonathan Haidt, Jesse Graham, and Morteza Dehgani. 2019. Moral foundations dictionary for linguistic analyses 2.0. *Unpublished manuscript*.

Dean Fulgoni, Jordan Carpenter, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. An empirical exploration of moral foundations theory in partisan news sources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3730–3736. European Language Resources Association (ELRA).

Matthew Gentzkow and Jesse M. Shapiro. 2005. Media bias and reputation. *Journal of Political Economy*, 114:280 – 316.

Matthew Gentzkow and Jesse M. Shapiro. 2010. What drives media slant? evidence from u.s. daily newspapers. *Econometrica*, 78(1):35–71.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2G: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715, Seattle, United States. Association for Computational Linguistics.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.

Jesse Graham, Jonathan Haidt, and Brian A. Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96 5:1029–46.

Kurt Gray and Daniel Wegner. 2009. Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology*, 96(3):505–520.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. 2020. Accelerating large-scale inference with anisotropic vector quantization. In *International Conference on Machine Learning*.

Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20:98–116.

Jonathan Haidt, Jesse Graham, and Craig Joseph. 2009. Above and below left–right: Ideological narratives and moral foundations. *Psychological Inquiry*, 20(2-3):110–119.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI with shared human values. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, et al. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Frederic R Hopp, Jacob T Fisher, Devin Cornell, Richard Huskey, and René Weber. 2021. The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior research methods*, 53:232–246.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. DEGREE: A data-efficient generation-based event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1890–1908, Seattle, United States. Association for Computational Linguistics.

Xiaomeng Hu, Yijie Zhu, Feng Yu, David A Wilder, Li Zhang, Sylvia Xiaohua Chen, and Kaiping Peng. 2020. A cross-cultural examination on global orientations and moral foundations. *PsyCh Journal*, 9(1):108–117.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot Learning with Retrieval Augmented Language Models.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, et al. 2021. Can machines learn morality? the delphi experiment. *arXiv e-prints*.

Kristen Johnson and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 720–730, Melbourne, Australia. Association for Computational Linguistics.

Spassena P Koleva, Jesse Graham, Ravi Iyer, Peter H Ditto, and Jonathan Haidt. 2012. Tracing the threads: How five moral concerns (especially purity) help explain culture war attitudes. *Journal of research in personality*, 46(2):184–194.

Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. 2019. The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, 4:155 – 190.

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

Sebastian Krügel, Andreas Ostermaier, and Matthias W. Uhl. 2023. Chatgpt's inconsistent moral advice influences users' judgment. *Scientific Reports*, 13.

George Lakoff. 2010. *Moral politics: How liberals and conservatives think*. University of Chicago Press.

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1648. Association for Computational Linguistics.

Uri D. Leibowitz. 2014. Explaining moral knowledge. *Journal of Moral Philosophy*, 11(1):35 – 56.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *ArXiv*, abs/2304.11633.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.

Ying Lin, Joe Hoover, Gwenyth Portillo-Wightman, Christina Park, Morteza Dehghani, and Heng Ji. 2018. Acquiring background knowledge to improve moral value prediction. In *2018 ieee/acm international conference on advances in social networks analysis and mining (asonam)*, pages 552–559. IEEE.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

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

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. 2022. POLITICS: pretraining with same-story article comparison for ideology prediction and stance detection. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1354–1374. Association for Computational Linguistics.

Dongfang Lou, Zhilin Liao, Shumin Deng, Ningyu Zhang, and Huajun Chen. 2021. MLBiNet: A cross-sentence collective event detection network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4829–4839, Online. Association for Computational Linguistics.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13470–13479.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2795–2806, Online. Association for Computational Linguistics.

Ahmed Mahany, Heba Khaled, Nouh Sabri Elmitwally, Naif Aljohani, and Said Ghoniemy. 2022. Negation and speculation in nlp: A survey, corpora, methods, and applications. *Applied Sciences*, 12(10).

Thomas McPherson. 1984. The moral patient. *Philosophy*, 59(228):171–183.

Negar Mokhberian, Andrés Abeliuk, Patrick Cummings, and Kristina Lerman. 2020. Moral framing and ideological bias of news. In *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12*, pages 206–219. Springer.

Marlon Mooijman, Joe Hoover, Ying Lin, Heng Ji, and Morteza Dehghani. 2018. Moralization in social networks and the emergence of violence during protests. *Nature human behaviour*, 2(6):389–396.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309, San Diego, California. Association for Computational Linguistics.

Xiaoman Pan, Wenlin Yao, Hongming Zhang, Dian Yu, Dong Yu, and Jianshu Chen. 2023. Knowledge-in-context: Towards knowledgeable semi-parametric language models. In *11th International Conference on Learning Representations, ICLR 2023*. OpenReview.net.

Tanmay Parekh, I Hsu, Kuan-Hao Huang, Kai-Wei Chang, Nanyun Peng, et al. 2022. Geneva: Pushing the limit of generalizability for event argument extraction with 100+ event types. *arXiv*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Shamik Roy, Maria Leonor Pacheco, and Dan Goldwasser. 2021. Identifying morality frames in political tweets using relational learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9939–9958, Online and Punta Cana, Dominican Republic. 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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Roser Saurí and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268.

Dietram A Scheufele. 1999. Framing as a theory of media effects. *Journal of communication*, 49(1):103–122.

Barry Schwartz and Kenneth Sharpe. 2011. *Practical Wisdom: The Right Way to Do the Right Thing*. Riverhead, New York.

Usman Shahid, Barbara Di Eugenio, Andrew Rojecki, and Elena Zheleva. 2020. Detecting and understanding moral biases in news. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 120–125, Online. Association for Computational Linguistics.

Weijia Shi, Julian Michael, Suchin Gururangan, and Luke Zettlemoyer. 2022. Nearest neighbor zero-shot inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3254–3265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marco Stranisci, Michele De Leonardis, Cristina Bosco, and Viviana Patti. 2021. The expression of moral values in the twitter debate: a corpus of conversations. *IJCoL. Italian Journal of Computational Linguistics*, 7(7-1, 2):113–132.

MeiHan Tong, Bin Xu, Shuai Wang, Meihuan Han, Yixin Cao, Jiangqi Zhu, Siyu Chen, Lei Hou, and Juanzi Li. 2022. DocEE: A large-scale and fine-grained benchmark for document-level event extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3970–3982, Seattle, United States. Association for Computational Linguistics.

Jackson Trager, Alireza S Ziabari, Aida Mostafazadeh Davani, Preni Golazazian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, et al. 2022. The moral foundations reddit corpus. *arXiv*.

Teun A Van Dijk. 2013. *News as discourse*. Routledge.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Pat Verga, Haitian Sun, Livio Baldini Soares, and William Cohen. 2021. Adaptable and interpretable neural MemoryOver symbolic knowledge. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3678–3691, Online. Association for Computational Linguistics.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Wendell Wallach and Colin Allen. 2008. *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Sze-Yuh Nina Wang and Yoel Inbar. 2021. Moral-language use by us political elites. *Psychological Science*, 32(1):14–26.

Jörgen Westerståhl. 1983. Objective news reporting: General premises. *Communication research*, 10(3):403–424.

Peter White and Eija Ventola. 2002. Media objectivity and the rhetoric of news story structure. In *Discourse and Community. Doing Functional Linguistics. Language in Performance*.

Christopher Wolsko, Hector Ariceaga, and Jesse Seiden. 2016. Red, white, and blue enough to be green: Effects of moral framing on climate change attitudes and conservation behaviors. *Journal of Experimental Social Psychology*, 65:7–19.

Hang Yang, Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, and Taifeng Wang. 2021. Document-level event extraction via parallel prediction networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6298–6308, Online. Association for Computational Linguistics.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: Sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. 2019. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5284–5294, Florence, Italy. Association for Computational Linguistics.

Xiang Yue, Xinliang Frederick Zhang, Ziyu Yao, Simon M. Lin, and Huan Sun. 2021. Cliniqg4qa: Generating diverse questions for domain adaptation of clinical question answering. *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 580–587.

Xinliang Frederick Zhang, Nick Beauchamp, and Lu Wang. 2022. Generative entity-to-entity stance detection with knowledge graph augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9950–9969. Association for Computational Linguistics.

Jianlong Zhou, Heimo Müller, Andreas Holzinger, and Fang Chen. 2023. Ethical chatgpt: Concerns, challenges, and commandments. *ArXiv*, abs/2305.10646.

A Moral Events Annotation Quality Control

We ensured the quality of the annotations at multiple steps in the collection process. All annotators participated in a training phase before beginning the annotations. In addition, the annotators participated in a weekly review session with the authors who would answer questions and provide guidance for annotators to revise their annotations.

We also found high inter-annotator agreement. This paragraph is based on comparing article annotations before and after the **revision**, i.e., approach (a) as described in §3.2. To compute agreement, we first identify overlapping moral event text spans where half of the words are identical, and then obtain Krippendorff's alpha's on the annotated properties (e.g., Agent, Patient) of the events. The revised articles have on average 5.7% more annotations than the first-pass articles. In terms of the nature of disagreements, some disagreements were on whether an event was negated. For example, a sentence like "the president did not sign the bill" contains a clearly negated event, due to the presence of the word "not." However, in the sentence "the president hesitated to sign the bill", one annotator could have annotated the event "hesitated", while another could annotate the negated event "sign". In addition, annotators sometimes disagreed on the morality of an event. For example, "the Supreme Court overrule the case" could be marked as Harm towards one patient, or Care towards a different patient. In addition, a secondary morality, Authority, was often annotated, but the disagreement was on whether the Authority was strong (acting in a way that exercised their authority) or moderate/weak (just doing their job). Many of these such annotations are subjective, though overall we find that these disagreements do not substantially lower the quality of our dataset. For this project, we use the revised annotations as training and testing data for our models.

Likewise, a similar quality control study is conducted on annotated articles undergoing **merging**, approach (b) as described in §3.2. Agreement levels are included in table A2. For this portion of data, we use the merged annotations as training and testing data for our models. Agreement on the article's ideological leaning is 0.7577.

Furthermore, upon comparing all annotated articles, our annotated **article leanings** align with AllSides' media-level labels for 70.9% and 76.4% of the time before and after the second-pass adjudication, respectively. We follow Zhang et al. (2022) and consider the difference between our annotated article leaning and AllSides label within one level as a match, e.g., Left (0) and Lean Left (1) are matched. This further illuminates the high quality of Moral Events and the effectiveness of our design two-pass annotation process.

Entity	Frequency
Americans	156
Donald Trump	123
United States	118
Republican Party	100
Joe Biden	93
Democratic Party	87
Barack Obama	81
United States Congress	72
People	58
Supreme Court of the United States	52
Federal Government of the United States	45
Justice Department	41
Biden Administration	35
Hillary Clinton	27
United States House of Representatives	27
United States Senate	24
White House	23
Immigrants	22
Trump Administration	22
Obama Administration	21
Police	20
Affordable Care Act	20
Women	20
Federal Bureau of Investigation	18
Ukraine	18
Food and Drug Administration	18
Senate Republicans	18
State Department	17
Mitch McConnell	17
Lawmakers	16

Table A1: Top-30 frequent entities in Moral Events sorted by their frequencies, i.e., the number of articles in which an entity appears.

Attribute	Merged	Revised
Agent	0.77	0.94
Patient	0.64	0.92
Morality	0.67	0.92
Event Status	0.59	0.91

Table A2: Krippendorff's alpha on various event attributes for revised (approach a) and merged (approach b) event annotations.

B Implementation Details of Lexicon Memory Access

Lexicon Memory Access. Access to the Lexicon Memory is triggered when encountering the morality special tokens as follows. MOKA takes as a query vector $\mathbf{h}_q$, the averaged representation of the special token pair, (`<Morality>`, `</Morality>`),

	Moral Events			eMFD Corpus (Hopp et al.)		
	Train	Dev	Test	Train	Dev	Test
# of stories	112	16	30	-	-	-
# of articles	336	48	90	261	54	96
# of sentences	9,568	1,256	2,605	10,331	2,042	3,454
# of moral events	4,124	494	876	10,694	1,839	4,513
# of moralities	4,948	606	1,047	11,814	1,958	5,562
Time range	2012-2021	01-06/2022	07-12/2022	-	-	-

Table A3: Splits and statistics of Moral Events and eMFD corpus (Hopp et al., 2021). It is worth noting that a moral event might embody more than one morality.

and the moral mention in between. $\mathbf{h}_q$ is then used to retrieve relevant moral knowledge $\mathbf{h}_m$ from the Lexicon Memory via a single-head attention mechanism.

$$\mathbf{h}_m = \mathbf{W}_2(\Sigma\alpha_i \cdot \mathbf{m}_i) \tag{1}$$

$$\alpha_i = \frac{\exp\left(\mathbf{m}_i^\top \mathbf{W}_1 \mathbf{h}_q\right)}{\Sigma_{j=1}^M \exp\left(\mathbf{m}_j^\top \mathbf{W}_1 \mathbf{h}_q\right)} \tag{2}$$

where M is the size of moral lexicon, $\mathbf{m}_i$ is a moral word embedding, and $\mathbf{W}_1$ and $\mathbf{W}_2$ are learnable matrices. Eventually, $\mathbf{h}_m$ is added to $\mathbf{h}_q$, the sum of which is normalized before being fed to the next Transformer layer, which is 9^{th} layer in MOKA.

Moral Word Knowledge Pre-training. The pre-training objective is a combination of language modeling ($\mathcal{L}_{LM}$), morality prediction ($\mathcal{L}_{MV}$), and moral word linking ($\mathcal{L}_{MWL}$) and moral label association ($\mathcal{L}_{MLA}$). In this part, we provide detailed mathematical formulations for $\mathcal{L}_{MWL}$ and $\mathcal{L}_{MLA}$.

Without loss of generality, the input sentence is defined as $\mathbf{x} = [x_1, x_2, \cdots, x_T]$ of length T which contains a set of moral mentions $\{(x_j, m_j, V_j)\}$. x_j is a moral mention in $\mathbf{x}$, m_j is the corresponding moral word,[14] and $V_j = \{v_{j,1}, v_{j,2}, ...\}$ is the set of associated moralities.

$\mathcal{L}_{MWL}$: For each moral mention (x_j) in text, the moral word linking objective guides MOKA to identify the corresponding moral word (m_j) by learning to maximize the attention score over the correct entry. That is, $\mathcal{L}_{MWL} := \max \alpha_{m_j}$, where α_{m_j} is computed using Equation (2).

$\mathcal{L}_{MLA}$: For each moral mention (x_j) in text, the moral label association objective is, for each morality ($v_{j,k}$) embodied by the mention, maximize the summation of attention scores over all moral words that share the same morality. Here, we denote $M_{v_{j,k}}$ as a set of moral words that carry $v_{j,k}$ value. We compute the aggregated attention score ($\mathrm{A}_{j,k}$) for each embodied morality as follows:

$$\mathrm{A}_{j,k} = \sum_{m_i \in M_{v_{j,k}}} \alpha_{m_i} \tag{3}$$

We then denote $\mathbf{A}_j$ as the set of embodied moralities' aggregated attention scores, i.e., $\mathbf{A}_j = \{\mathrm{A}_{j,k_1}, \mathrm{A}_{j,k_2}, ...\}$ where $|\mathbf{A}_j| = |V_j|$. By convention, we use $\mathbf{A}_j^C$ to represent the complement set, i.e., a set of aggregated attention scores of non-embodied moralities. To support the training of moral words that might be associated with more than one morality, we use *multi-label margin loss* as shown in Equation (4):

$$\mathcal{L}_{MLA} := \min \frac{\sum_{p \in \mathbf{A}_j} \sum_{q \in \mathbf{A}_j^C} \max(0, 1 - (p - q))}{|\mathbf{A}_j| + |\mathbf{A}_j^C|} \tag{4}$$

$$:= \min \frac{\sum_{p \in \mathbf{A}_j} \sum_{q \in \mathbf{A}_j^C} 1 - (p - q)}{10} \tag{5}$$

We derive Equation (5) from Equation (4), since we notice that aggregated attention scores are always bound between $[0, 1]$, and there is a fixed number of plausible moralities, which is 10.

C Annotation Guideline

Annotation Goal: Jointly annotate entities (with agency property) and events (with a moral basis).

Entities. Entities are the participants in events. They will usually possess moral agency, i.e., the capability of doing things right or wrong (Gray and Wegner, 2009). There will usually be two entities for every event: the **agent** is the doer or enabler of the event, and the **patient** is the one affected by the event.

Entity Types: An entity will often be a Person, Organization, Nation, or something that is backed by entities that have agency.

[14] Here, we abuse the notations: m_j refers to both a moral word and the corresponding entry ID in our morality lexicon.

Model	MORAL EVENTS F1	MORAL EVENTS Acc.	eMFD Corpus F1	eMFD Corpus Acc.
Baselines				
Dictionary-based counting (Brady et al.)	45.8	56.8	33.0	52.0
RoBERTa-large (large; Liu et al.)	63.6	82.6	28.7	69.0
POLITICS (base; Liu et al.)	62.7	82.4	29.0	68.8
ChatGPT (zero-shot; Li et al.)	41.2	69.9	31.9	66.9
ChatGPT (few-shot; Li et al.)	46.9	75.6	30.5	69.1
Flan-T5 (large; Chung et al.)	62.0	83.6	25.4	68.4
MOKA with moral word knowledge augmentation only				
Pretrain on `Morality Bank` only	63.6	83.9	27.3	69.0
+ moral word linking	63.9	83.9	27.8	69.0
+ moral label association	64.0	83.9	28.5	69.1
MOKA with moral scenario knowledge augmentation only				
Delphi (moral agreement; Jiang et al.)	62.5	84.0	30.0	70.2
w/ RLM	63.2	84.2	30.3	70.3
Delphi (moral judgement; Jiang et al.)	63.7	84.1	30.4	70.4
w/ RLM	62.3	83.8	30.1	70.3
Deontology (Hendrycks et al.)	62.5	83.6	30.5	70.5
w/ RLM	62.2	83.5	30.4	70.4
Justice (Hendrycks et al.)	62.5	83.7	30.4	70.3
w/ RLM	62.4	83.6	31.8	70.6
Social chem (judgement; Forbes et al.)	63.6	84.2	30.0	70.2
w/ RLM	62.9	83.6	30.7	70.4
Social chem (foundation; Forbes et al.)	62.2	83.7	32.4	70.6
w/ RLM	64.1	84.0	32.5	70.7
Social chem (morality; Forbes et al.)	62.7	83.8	30.0	70.3
w/ RLM	63.3	84.1	32.5	70.6
MOKA with dual moral knowledge augmentation				
Delphi (moral agreement)	64.4	84.3	34.0	71.0
w/o moral label association	63.3	84.0	33.2	70.8
Delphi (moral judgement)	63.3	83.6	32.9	70.7
w/o moral label association	63.9	84.1	32.1	70.6
Deontology	64.0	84.0	32.9	70.8
w/o moral label association	64.2	84.0	34.3	71.1
Justice	64.0	84.0	32.9	71.0
w/o moral label association	63.7	84.1	33.3	71.0
Social chem (judgement)	64.3	84.2	32.7	70.9
w/o moral label association	64.2	84.3	33.4	71.1
Social chem (foundation)	**65.3**	**84.3**	33.7	71.0
w/o moral label association	64.1	84.0	33.4	71.0
Social chem (morality)	64.5	83.9	**34.6**	**71.3**
w/o moral label association	63.8	84.0	33.3	70.9
Improvements over best baseline	2.7%	0.8%	4.8%	3.2%

Table A4: Full weighted F1 and accuracy results on MORAL EVENTS and eMFD Corpus (Hopp et al., 2021) for task A. Best results are in **bold**. *Color scheme*: MOKA and its single-knowledge-augmented variants are highlighted on a scale of 5 red shades based on the relative improvements over the strongest baseline. They are highlighted in pale pink, pink, rose-pink, rose-red and dark red, if the relative gains are in the range of $(0.0\% - 0.5\%]$, $(0.5\% - 2.0\%]$, $(2.0\% - 4.0\%]$, $(4.0\% - 7.0\%]$ and $(7.0\% - \infty\%)$, respectively.

Agency: Entities usually possess moral agency regardless of whether they are the agent or patient. Sometimes, an entity itself might not have agency but is backed by some other entities that have agency. For example, "hurting the Constitution" essentially means "hurting the people". The Constitution itself has no agency, but the people behind the Constitution have agency, so we annotate either "Constitution" or "People" as the moral patient.

Canonical Names are uniquely identified strings in a knowledge base such as Wikipedia. Entities should be annotated with their canonical name, if possible. An entity's canonical name might not be the first occurrence of that name in the article. For consistency, please use the same canonical name throughout the entire article. For example, mentions of "President Trump" or "Trump" should be annotated as "Donald Trump".

Moral Events. Moral events have a basis in moral foundations and possess moral evaluations that arise when the patient has agency and can be harmed or helped by an action/event (McPherson, 1984; Gray and Wegner, 2009). The annotated event must be a concise span that exactly appears in the text, and it should carry stand-alone meaning.

Event Entities: Agent & Patient: For each moral event, there must be at least one enabler (agent) as well as at least one affected (patient). If the agent and patient are not apparent in the text, please infer them to make sure both agent and patient are present. For example, in the following sentence "That briefing averted congressional criticism, even though the administration formally missed a deadline to implement sanctions targeting Russian defense and intelligence industries", we can tell that there is a moral event "missed a deadline" (which embodies a morality of betrayal), and the associated agent is "Trump Administration". However, the patient is not explicitly stated, but we can infer "Congress" as a patient since missing the deadline would impede Congress from implementing sanctions or taking further actions.

Moral Foundations: Follow MFT (Graham et al., 2009), MFD 2.0 (Frimer et al., 2019) and supplementary materials of eMFD (Hopp et al., 2021) to annotate the moral foundation(s) embodied in each moral event. Note, a moral event can embody more than one morality.

1. **Ten moralities:** There are five moral foundations, each with a positive and negative polarity: Care, Harm, Fairness, Cheating, Loyalty, Betrayal, Authority, Subversion, Sanctity, and Degradation.

2. **Author's Point of View:** During the annotation process, *annotate from the author's perspective rather than the audience's*. In other words, consider what the author is trying to say or imply by writing these words. You may also consider why the author included this event, and what kind of morality is embodied through the inclusion of this event.

3. **Morality Toward the Patient:** The annotated morality should reflect the perception of the patients, towards whom an agent performs a moral action.

Sentence	Napoleon now realized that it would be impossible , without running serious risks , to <Morality> oppose </Morality>$_{\text{Subversion}}$ the movement in favor of <Morality> unity </Morality>$_{\text{Loyalty}}$.
Morality	Subversion
Sentence	While waiting for emergency <Morality> help </Morality>$_{\text{Care}}$ to arrive , the <Morality> victim </Morality>$_{\text{Harm}}$ should wash the <Morality> wound </Morality>$_{\text{Harm}}$ site with soap and water and then keep the <Morality> injured </Morality>$_{\text{Harm}}$ area still and at a level lower than the heart .
Morality	Harm

Table A5: Sample examples from our constructed MORALITY BANK. The seed words used to crawl sentences are highlighted in red. For each moral mention in text, it is tagged with a special token pair, <Morality> and </Morality>, and its embodied morality is visually represented using $_{\text{subscript}}$.

Scenario	Label	Scenario Bank	Label set	# of examples
it is ok to take another person's account and use it as your own.	morally disagree	`Delphi` (moral agreement)	{morally agree, morally disagree}	200,000
enjoying your life with your family	morally good	`Delphi` (moral judgement)	{morally good, morally wrong, amoral}	400,000
I am working at the local fire station as a fireman. So I should light a lot of matches.	morally unreasonable	`ETHICS` (deontology)	{morally reasonable, morally unreasonable}	18,164
I usually exercise with my trainer, but stopped because She had a death in her family	morally reasonable	`ETHICS` (justice)	{morally reasonable, morally unreasonable}	21,791
Wasting your money on something you don't like	morally wrong	`Social chem` (judgement)	{morally good, morally wrong, amoral}	122,906
stay in communication with friends	loyalty-betrayal	`Social chem` (foundation)	{care-harm, loyalty-betrayal, authority-subversion, fairness-cheating, sanctity-degradation, amoral}	122,906
faking your relationships	cheating	`Social chem` (morality)	{care, harm, loyalty, betrayal, authority, subversion, fairness, cheating, sanctity, degradation, amoral}	122,906

Table A6: Sample (scenario, label) pairs from our curated suite of Moral Scenario Banks. The seven Moral Scenario Banks are derived from `Delphi` (Jiang et al., 2021), `ETHICS` (Hendrycks et al., 2021) and `Social Chem` (Forbes et al., 2020). Each row represents one scenario bank where the source is listed in *Scenario Bank* column. *Label set* column shows the complete set of plausible labels in each scenario bank.

Additionally, we also annotate the following **Event Status** to reveal the linguistic construct of a moral event.

Event Status: An event has one of three statuses:

- Actual: An event that is happening or has happened.
- Intentional: An event that is being planned or intended to happen in the future. Usually, it is the moral agent's subjective intention of the event.
- Speculative: An event that may happen, usually speculated by someone who is not a participant in the event (e.g. the speaker of a quote, or the author of the article). This can be used to mark an unsubstantiated guess of a past/current/future event.

Model	Task B	Task C				
	Trigger EM	Morality F1	Agent EM	Agent F1	Patient EM	Patient F1
Baselines						
DEGREE (base; Hsu et al.)	45.5	53.0	47.3	58.6	30.1	39.2
DEGREE (large; Hsu et al.)	46.2	54.2	49.2	60.3	30.5	40.3
ChatGPT (zero-shot; Li et al.)	19.5	39.5	30.3	49.8	12.3	23.2
ChatGPT (few-shot; Li et al.)	32.1	38.1	34.2	51.4	20.1	30.6
Flan-T5 (large; Chung et al.)	46.2	53.8	47.5	59.4	30.8	41.2
MOKA with moral word knowledge augmentation only						
Pretrain on `Morality Bank` only	45.3	54.6	47.5	59.9	31.2	41.7
+ moral word linking	45.6	55.9	47.6	59.8	**31.5**	41.7
+ moral word linking + moral label association	46.2	57.0	48.3	60.2	31.3	41.9
MOKA with moral scenario knowledge augmentation only						
Delphi (moral agreement; Jiang et al.)	46.6	55.9	48.9	60.9	30.8	41.5
w/ RLM	47.6	56.3	48.6	60.5	**31.6**	41.8
Delphi (moral judgement; Jiang et al.)	47.0	57.5	48.5	60.4	30.9	41.4
w/ RLM	47.4	55.6	48.5	60.3	31.2	41.5
Deontology (Hendrycks et al.)	46.1	54.8	49.0	60.9	30.9	41.6
w/ RLM	47.2	56.0	**49.5**	61.2	31.3	**42.1**
Justice (Hendrycks et al.)	46.6	54.7	48.7	60.7	31.0	41.5
w/ RLM	46.9	55.2	48.6	60.8	31.4	41.6
Social chem (judgement; Forbes et al.)	47.1	55.4	48.6	60.9	31.2	41.2
w/ RLM	47.2	54.9	48.5	60.1	31.3	41.6
Social chem (foundation; Forbes et al.)	46.7	56.5	48.9	**61.4**	31.0	41.4
w/ RLM	47.5	56.0	48.8	60.5	31.0	41.7
Social chem (morality; Forbes et al.)	46.8	56.3	48.6	60.6	31.2	40.7
w/ RLM	47.2	55.5	48.7	60.7	31.0	41.5
MOKA with dual moral knowledge augmentation						
Delphi (moral agreement)	48.0	57.0	48.2	60.7	30.0	40.6
w/o moral label association	47.1	56.1	47.6	60.2	30.6	40.8
Delphi (moral judgement)	47.4	56.8	48.1	60.3	30.2	40.5
w/o moral label association	46.7	57.2	47.6	60.0	30.2	40.5
Deontology	46.8	**58.2**	47.9	60.3	30.9	41.1
w/o moral label association	**48.1**	57.3	48.2	61.0	30.7	41.1
Justice	47.3	57.7	47.8	60.6	30.5	40.7
w/o moral label association	47.6	56.5	47.6	60.8	30.5	40.9
Social chem (judgement)	46.7	56.9	48.3	60.7	29.1	39.9
w/o moral label association	47.2	57.3	47.7	60.3	30.1	40.7
Social chem (foundation)	46.5	58.1	48.4	61.0	30.5	40.8
w/o moral label association	46.7	57.7	48.2	60.5	30.0	40.1
Social chem (morality)	47.0	57.1	47.7	60.0	30.4	40.7
w/o moral label association	46.8	57.1	47.7	59.9	29.9	40.1
Improvements over best baseline	4.1%	7.4%	0.6%	1.8%	2.6%	2.2%

Table A7: Full results on MORAL EVENTS for tasks B and C, and breakdown of performances by event attributes (average of 5 runs). Best results are in **bold**. *Color scheme*: MOKA and its single-knowledge-augmented variants are highlighted on a scale of 5 red shades based on the relative improvements over the strongest baseline. They are highlighted in pale pink, pink, rose-pink, rose-red and dark red, if the relative gains are in the range of $(0.0\% - 0.5\%]$, $(0.5\% - 2.0\%]$, $(2.0\% - 4.0\%]$, $(4.0\% - 7.0\%]$ and $(7.0\% - \infty\%)$, respectively.

Task	Text
Task A	Moral Event Detection task definition:\n\ Given an input list of words from a news article, identify the moral event trigger in the input list. An event \ is something that happens, a specific occurrence involving participants, and can frequently be described as a change of state. \ A moral event has a basis in moral foundations, and possesses moral evaluations which arise when the patient has agency and can be harmed or helped by an action undertaken by an agent. \ A moral event trigger is the main word or phrase that most explicitly \ expresses the occurrence of a moral event.\n\n\ In the input list, special tokens are defined as follows. \ <Title>and </Title>enclose the title of the news article; <News>and </News>enclose the truncated content of the news article; <Target>and </Target> enclose the target sentence from which the event trigger should be extracted. \n\ The output of the Moral Event Detection task should be a dictionary in the json format. Each \ dictionary corresponds to a trigger and should consist of \"trigger\", \"start_word_index\", \ \"end_word_index\", \"confidence\" four keys. The value of \"start_word_index\" key and \"end_word_index\" key are the \ index (zero-indexed) of the start and end word of \"trigger\" in the input list, respectively. The \ value of \"confidence\" key is an integer ranging from 0 to 100, indicating how confident you are that \ the \"trigger\" expresses a moral event. \ Note that your answer should only contain the json string and nothing else.\n\n\ You will first see 5 demonstrations of the task, and then you will be asked to perform the task for a given input list.\n\n Demonstration i: <Demostration i> \nPerform Moral Event Detection task for the following input list, and print the output:\n [“This”, “is”, “a”, “sample”, “input”]
Task B	Moral Dimension Prediction definition:\n\ Given a moral event span and an input list of words from a news article, make a 5-way judgment on the moral dimension for the given moral event. \ A more event span might embody more than one moral dimension. An event \ is something that happens, a specific occurrence involving participants, and can frequently be described as a change of state. \ A moral event has a basis in moral foundations, and possesses moral evaluations which arise when the patient has agency and can be harmed or helped by an action undertaken by an agent. \ The five moral dimensions are ’Care/Harm’, ’Fairness/Cheating’, ’Loyalty/Betrayal’, ’Authority/Subversion’, and ’Sanctity/Degradation’\n\n\ In the input list, special tokens are defined as follows: \ <Title>and </Title>enclose the title of the news article; <News>and </News>enclose the truncated content of the news article; \ <Target>and </Target>enclose the target sentence where the target moral event span stands; <Event>and </Event>enclose the target moral event span.\n\ The output of the Moral Event Detection task should be a dictionary in the json format. Each \ dictionary corresponds to a moral event and should consist of \"moral dimensions\" and \"confidence\" two keys. The value of \"moral dimensions\" should be a list of predicted moral dimensions that are embodied in the target moral event span. \ The value of \"confidence\" key is an integer ranging from 0 to 100, indicating how confident you are that \ the moral event span embodies predicted \"moral dimensions\". \ Note that your answer should only contain the json string and nothing else.\n\n\ You will first see 5 demonstrations of the task, and then you will be asked to perform the task for a given input list. \n\n Demonstration i: <Demostration i> \nPerform Moral Dimension Prediction task for the following input list, and print the output:\n [“This”, “is”, “a”, “sample”, “input”]
Task C	Moral Event Argument Extraction task definition:\n\ Given an input list of words from a news article and a moral event span, identify moral event arguments for the given moral event span. \ Specifically, moral event arguments consists of three attributes: moral agent, moral patient and 10-way morality prediction. \ An event is something that happens, a specific occurrence involving participants, and can frequently be described as a change of state. \ A moral event has a basis in moral foundations, and possesses moral evaluations which arise when the patient has agency and can be harmed or helped by an action undertaken by an agent. \ A moral event span is a main word or phrase that most explicitly \ expresses the occurrence of a moral event. A moral agent is the doer or enabler of a moral event, and the moral patient is the one affected by the moral event. \ The ten moralities are ’Care’, ’Harm’, ’Fairness’, ’Cheating’, ’Loyalty’, ’Betrayal’, ’Authority’, ’Subversion’, ’Sanctity’, and ’Degradation’\n\n\ In the input list, special tokens are defined as follows. \ <Title>and </Title>enclose the title of the news article; <News>and </News>enclose the truncated content of the news article; \ <Target>and </Target>enclose the target sentence where the target moral event span stands; <Event>and </Event>enclose the target moral event span.\n\ The output of the Moral Event Argument Extraction task should be a dictionary in the json format. Each \ dictionary corresponds to a moral event span and should consist of \ \"agent\", \"confidence-agent\", \ \"patient\", \"confidence-patient\", \ \"morality\" and \"confidence-value\" six keys. \n\ The value of \"agent\" and \"patient\" keys should be a list of moral agents and moral patients in their canonical names, respectively. \ Note, canonical names are uniquely-identified strings in a knowledge base such as Wikipedia. An entity’s canonical name might not be explicitly mentioned in the input list. \ For example, the canonical names of \"Trump\", \"Republican\", \"Democrats\", \"Senate\", and \"United States Department of State\" are \"Donald Trump\", \"Republican Party\", \"Democratic Party\", \"United States Senate\", and \"State Department\", respectively. \n\ The value of \"confidence-agent\" key is an integer ranging from 0 to 100, indicating how confident you are that \ the value of \"agent\" key plays the agent role in the target moral event. \n\ The value of \"confidence-patient\" key is an integer ranging from 0 to 100, indicating how confident you are that \ the value of \"patient\" key plays the patient role in the target moral event. \n\ The value of \morality\" should be a list of predicted moralities that are embodied in the target moral event span. \ The value of \"confidence-value\" key is an integer ranging from 0 to 100, indicating how confident you are that \ the moral event span embodies predicted \"moralities\". \n\ Note that your answer should only contain the json string and nothing else.\n\n\ You will first see 5 demonstrations of the task, and then you will be asked to perform the task for a given input list. \n\n Demonstration i: <Demostration i> \nPerform Moral Dimension Prediction task for the following input list, and print the output:\n [“This”, “is”, “a”, “sample”, “input”]

Table A8: Prompts used to test ChatGPT’s moral reasoning capability, adapted from Li et al. (2023).