MOKA: Moral Knowledge Augmentation for Moral Event Extraction

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

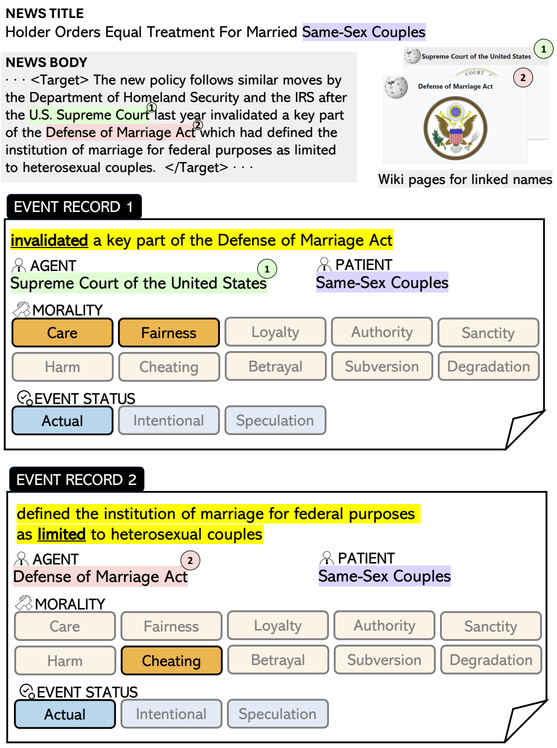
[1]Computer Science and Engineering, University of Michigan, Ann Arbor, MI
[2]Department of Computer Science, University of Hawaii at Hilo, Hilo, HI
[3]Department of Political Science, Northeastern University, Boston, MA
[1]{xlfzhang,wangluxy}@umich.edu
[2]wswu@hawaii.edu, [3]n.beauchamp@northeastern.edu

Abstract

News media often strive to minimize explicit moral language in news articles, yet most articles are dense with moral values as expressed through the reported events themselves. However, values that are reflected in the intricate dynamics among *participating entities* and *moral events* are far more challenging for most NLP systems to detect, including LLMs. To study this phenomenon, we annotate a new dataset, MORAL EVENTS[1], consisting of 5,494 structured event 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, which leverages knowledge derived from moral words and moral scenarios to produce structural representations of morality-bearing events. Experiments show that MOKA outperforms competitive baselines across three moral event understanding tasks. Further analysis shows even ostensibly nonpartisan media engage in the selective reporting of moral events.

1 Introduction

Many news media frame their stories to further a particular ideological viewpoint (Scheufele, 1999), often employing moral values rather than explicitly partisan language to subtly affect readers (Haidt and Graham, 2007; Haidt et al., 2009; Lakoff, 2010; Feinberg and Willer, 2015). However, existing NLP methods, including LLMs, face significant challenges in discerning moral values. Past work has shown that these limitations may be due to lack of context (Graham et al., 2009; Frimer et al., 2019), lack of moral reasoning capabilities (Jiang et al., 2021), and the complexity of moral stances (Zhou et al., 2023; Krügel et al., 2023). Detecting moral values is even harder for non-partisan news outlets which deliberately avoid explicit moral language in their reporting, but may express moral values indirectly by selecting which morally-laden events to report. Thus there remains an imperative need for NLP tools that can *decipher* moral values latent in narrated events and interactions among entities.



NEWS TITLE
Holder Orders Equal Treatment For Married Same-Sex Couples

NEWS BODY
· · · <Target> The new policy follows similar moves by the Department of Homeland Security and the IRS after the U.S. Supreme Court last year invalidated a key part of the Defense of Marriage Act which had defined the institution of marriage for federal purposes as limited to heterosexual couples. </Target> · · ·

Figure 1: Sample moral event extractions (MEEs) for a target sentence from our 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.

In particular, news articles tell complex stories that contain multiple people and events along with interactions among them. The participants in the events, the ordering of them, and the selection of

[1] Our data and codebase are available at https://github.com/launchnlp/MOKA.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4481–4502
June 16-21, 2024 ©2024 Association for Computational Linguistics

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 **event** level, 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 fine-grained analysis of 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.

Following prior work, we employ Moral Foundations Theory (MFT; Haidt and Graham, 2007; Graham et al., 2009, 2013), which posits five moral foundations, each containing two polarities of virtue and vice, e.g., `Care/Harm`. MFT has been widely used in analyzing both mainstream news (Hopp et al., 2021) and social media content (Lin et al., 2018; Hoover et al., 2020; Trager et al., 2022), but often in superficial ways based on explicit moral language, without utilizing the larger context and external knowledge to better understand interactions between participants in moral actions. To illustrate the challenges in moral event understanding, Figure 1 shows a sentence containing multiple events with moral values but little in the way of explicit moral language. The correct identification of Event Record 1 must take into account a longer context (e.g., the title) and 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 it usually carries a negative connotation and might imply `Harm` on a surface level, here it is actually beneficial to the patient and in fact embodies `Fairness` and `Care` towards *Same-Sex Couples*.

Our paper makes the following contributions. First, we define a new schema of **moral events**, grounded in MFT and linguistics. We then propose **moral event extraction (MEE)**: given unstructured text, detect morality-bearing event triggers, extract participants, and infer embodied moralities.

Second, to solve MEE, we curate a large dataset, **MORAL EVENTS**, consisting of moral event annotations of news articles from diverse US media outlets. This dataset is unique in that annotations are conducted on multiple news articles about the same story, allowing us to analyze differences in how news outlets of different ideologies report moral events. Moral participant annotations go beyond surface mentions and syntactic constraints, capturing *implicit participants* in moral actions.

We propose **MOKA**, a generative framework for MEE with **MO**ral **K**nwoledge **A**ugmentation. Capitalizing on the recent success of retrieval augmentation (Lewis et al., 2020; Févry et al., 2020; Izacard et al., 2022), MOKA integrates moral knowledge derived from varying granularities, moral words and moral scenarios. Additionally, to support MOKA pre-training, we crawl a bank of 344k morality-bearing examples, MORALITY BANK, leveraging validated morality lexicons (Graham et al., 2009; Frimer et al., 2019). Extensive experiments highlight the usefulness and robustness of MOKA over strong baselines, including SOTA event extraction models and ChatGPT (*gpt-3.5-turbo*). The results show that external moral knowledge is essential for LMs to excel at MEE and ethics-related moral reasoning in general. Further analysis of moral event reporting in news reveals substantive findings, including (1) left-right asymmetries where `Right-to-Left` moral events are more prevalent than the reverse regardless of underlying moral values or outlet ideology, and (2) a tendency of centrist media to focus primarily on moral events enacted by right-leaning entities.

2 Related Work

2.1 NLP Benchmarks for Morality

Recent NLP research has seen a surge in interest focusing on morality, including moral norms, ethical judgment, and social bias, Most work is based on MFT (Haidt and Graham, 2007; Graham et al., 2009), a social psychology theory that posits five moral foundations, each with two polarities: `Care/Harm`, `Loyalty/Betrayal`, `Fairness/Cheating`, `Authority/Subversion` and `Sanctity/Degradation`.

Many recently annotated morality **datasets** are limited to social media text, including Twitter (Johnson and Goldwasser, 2018; Hoover et al., 2020; Wang and Inbar, 2021) 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 everyday-situation judgments (Jiang et al., 2021). Only a few existing works study morality in news articles, mainly at the word-level (Mokhberian et al., 2020) or topic-level (Fulgoni et al., 2016; Shahid et al., 2020). By contrast, we collect a high-quality corpus of moral events from a wide range of news sources, to support 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 corpus (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, who are not mentioned in the text span. Meanwhile, eMFD only annotates text spans and their embodied moralities. In contrast, our work contains fine-grained structured event annotations including participants and linguistic features.

Moral foundation prediction is a task treated as categorical classification, accomplished by fine-tuning pre-trained language models (Lin et al., 2018; Alhassan et al., 2022). Recent works approach it with template-based natural language generation (Forbes et al., 2020; Jiang et al., 2021). 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 extract fine-grained structured moral events using both the context where events occur and the external moral knowledge. This allows us to capture nuances of moral actions involving different participants, and to better understand the role morality plays in shaping news narratives.

2.2 Event Extraction

Our work follows a long line of research in event extraction (EE), including two key stages: event detection (ED) and event argument extraction (EAE). ED is defined as identifying an 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). For the more challenging EAE task, three major approaches have been developed: sequence labeling (Chen et al., 2015; Nguyen et al., 2016; 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 incorporate ontology knowledge about argument roles; and generative models for structured extraction (Li et al., 2021; Lu et al., 2021; Du and Ji, 2022). More recently, LLMs have been used for EAE (Zhang et al., 2024), but with subpar performance compared with specialized systems (Li et al., 2023; Han et al., 2023).

Our work proposes a new understanding task, *moral event extraction*, a two-stage EE task with a special focus on morality-bearing events. Unlike conventional EE, where each event type has its own event schema, we define a universal schema for moral events grounded in MFT and linguistics. To the best of our knowledge, we are also the first to explicitly model *multi-granularity moral knowledge* for EE tasks, as well as moral reasoning and understanding in general.

3 Moral Events Curation

We define a new structured schema for a **moral event** which represents a moral action, visualized in Figure 1. A moral event consists of moral agents, moral patients, a morality-bearing event span and event trigger, embodied morality, and event status. We list major concepts below, and refer readers to Appendix A for full descriptions. A moral action is performed or enabled by **moral agents** and affects **moral patients**. An **event trigger** is usually a single word within an **event span** that can best characterize a moral action. This span embodies one or more **moralities** in MFT (Gray and Wegner, 2009). Note that moral patients may be *implicit*: they do not have to be mentioned in a *target sentence*. To study the linguistic phenomenon, we further annotate **event status** which describes the factuality of an event (Saurí and Pustejovsky, 2009; Lee et al., 2015), i.e., whether an event is *actual*, *intentional* or *speculative* (Mahany et al., 2022).

Annotation Process. We create our dataset, Moral Events, using the following process. We first sample 87 news stories from *SEESAW* (Zhang et al., 2022), where each story contains 3 articles on the same event but reported by media of different ideologies. To 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, e.g., 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, moral events are annotated by native English speakers (at least two for each article). Each annotator has access to all three articles in a story to maintain a non-biased view. we list the major steps of annotating a single article, with a detailed annotation protocol in Appendix B.

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 names, i.e., the names 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 appendix A.
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 **merges** and resolves annotation conflicts. 83% of articles are revised with approach (a) while 17% adopt approach (b).

We ensured the quality of the annotations at multiple steps in the collection process (Appendix C). Also see Table A1 for annotation agreements.

Statistics. MORAL EVENTS include 474 news articles from 158 stories, published by 63 different media outlets (26 left, 18 center, and 19 right). On average, each article contains 11.6 event annotations. The articles cover 38 salient topics reported from 2012 to 2022, and includes 1,952 distinct entities (see Table A2). We annotate diverse **entity types**: People (62.4%), Organization (20.4%), Geo-Political (9.6%), and Others (7.6%).

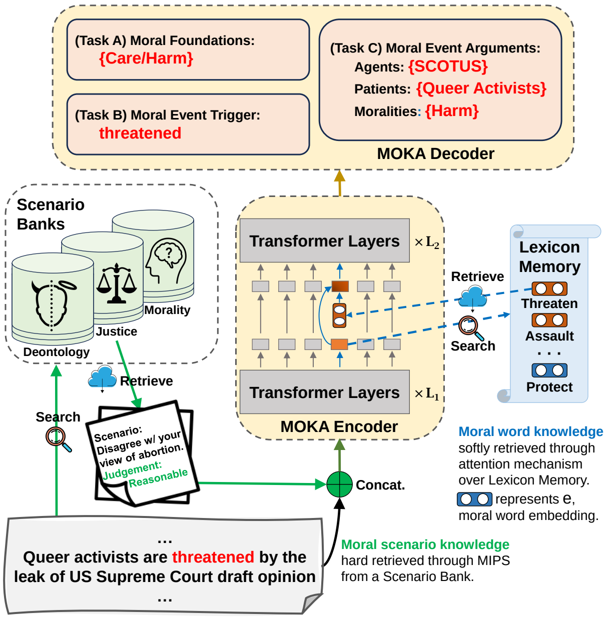

(Task A) Moral Foundations:
{Care/Harm}
(Task B) Moral Event Trigger:
threatened
(Task C) Moral Event Arguments:
Agents: {SCOTUS}
Patients: {Queer Activists}
Moralities: {Harm}
MOKA Decoder
Scenario Banks
Morality
Justice
Deontology
Retrieve
Search
Scenario: Disagree w/ your view of abortion. Judgement: Reasonable
Transformer Layers ×L2
Transformer Layers ×L1
MOKA Encoder
Retrieve
Search
Lexicon Memory
Threaten
Assault
. . .
Protect
Concat.
Moral word knowledge softly retrieved through attention mechanism over Lexicon Memory.
represents e, moral word embedding.
Moral scenario knowledge hard retrieved through MIPS from a Scenario Bank.
...
Queer activists are threatened by the leak of US Supreme Court draft opinion
...

Figure 2: Overview of MOKA for (downstream) moral event extraction. It highlights the process of retrieving and combining relevant scenarios, and the integration of moral word knowledge through attention-based retrieval. Embeddings in Lexicon are colored in red if moral words embody `Harm`, or blue if `Care`. "SCOTUS" is an acronym for "Supreme Court of the United States". {} indicates there can be multiple answers.

4 The MOKA Models

We now define the task of *moral event extraction* (MEE), which extracts structured moral events from unstructured texts. Similar to mainstream event extraction, we decompose MEE into the sub-tasks of event detection and event argument extraction, but with a focus on morality. To tackle these tasks, we develop a new 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). After two-stage pre-training, MOKA is fine-tuned on downstream tasks (§4.3). We instantiate MOKA with *Flan-T5-large*, though it is compatible with models of other architectures.

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, LLMs' access to moral facts is usually limited due to the lack

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

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

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; 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). So far, these retrieval mechanisms have been mostly used 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, no such KB exists in the realm of moralities. To start, we hypothesize that *an utterance embodies a morality if it contains a morality-bearing mention*, where a moral mention is an occurrence of a moral word.[4] We then combine two validated morality lexicons, MFD (Graham et al., 2009) and MFD2.0 (Frimer et al., 2019) into 891 *moral words*, and scrape example sentences that contain at least one moral mention from four authoritative online dictionaries.[5] We limit the sentence length to between 5 and 80 words, totaling $334k$ sentences. 95% of example sentences are used for pre-training, and the rest for validation. Samples are shown in Table A3.

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 one 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 MORALITY BANK.

When encoding a sentence, a moral mention is first tagged with a special token pair (`<Morality>`, `</Morality>`). For each mention, MOKA computes a query vector $\mathbf{h}_q$, which is the averaged representation of the special token pair and the enclosed moral mention. $\mathbf{h}_q$ 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 Attn(·,·) is the cross-attention mechanism in Vaswani et al. (2017). Finally, the sum of $\mathbf{h}_m$ and $\mathbf{h}_q$ is normalized, and 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 the 8th layer ($L_1 = 8$ and $L_2 = 16$ in Figure 2).

Moral Word Knowledge Pre-training. The pretraining objective is a combination of language modeling ($\mathcal{L}_{LM}$), morality prediction ($\mathcal{L}_{MV}$), moral word linking ($\mathcal{L}_{MWL}$), and moral label association ($\mathcal{L}_{MLA}$), each 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(s) embodied in a morality-bearing input sentence. To prevent MOKA from learning shortcuts, the *seed word* used to scrape the input sentence is always masked. Two new training objectives are also proposed to train the memory access mechanism effectively. For each moral mention, the **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* in Figure 2. The **moral label association** objective promotes MOKA's capability of associating a moral mention and its embodied morality(s). It is achieved by, for each morality embodied by a mention, maximizing the summation of attention scores over all moral words in $\mathbf{E}$ that share the same morality. To handle moral words that are associated with multiple moralities, we use *multi-label margin loss* (eq. (4)). Compared to cross entropy, this objective flattens scores over target moralities and mitigates saturated gradients. Detailed mathematical formulations of Lexicon Memory Access are in Appendix D.

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. While fundamental theoretical moral theories are prescriptive

[4] A moral word is a unique entry in the morality lexicon and the base form of moral mentions. Mentions like *threatening* and *threatened*, for example, both map to the *Threaten* entry.

[5] Cambridge (UK & US sites), Merriam-Webster, Dictionary.com, and YourDictionary.com.

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 can directly understand abstract moral principles without the need for interaction with concrete moral scenarios, those principles are too perplexing for machines. Thus, we guide MOKA to develop its moral sense by immersing itself 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, with statistics and examples in Table A4.

Scenario Retrieval. For each moral scenario bank, we convert (scenario, label) pairs into 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 ($K = 3$ in this study) relevant scenario pairs using dot product similarity between the query and keys, i.e., scenarios. The 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 (scenario, label) format, e.g., ("enjoying your life with your family", "morally good"), MOKA should generate a desired output.

To help further digest the 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 retrieved example and apply MLM objective to recover this label. By explicitly training the encoder to discern the associated moral label, it helps MOKA from collapsing to simply memorizing retrieved labels and making trivial inferences.

This approach is in line with retrieval augmented generation, where existing work mainly empowers a language model with a retriever to fetch text-form knowledge items from an external knowledge bank (Lewis et al., 2020; Fan et al., 2021; Shi et al., 2022; Izacard et al., 2022). Most existing work is limited to the use of a single knowledge source, with the exception of Pan et al. (2023); Zhang 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 of MOKA with **dual knowledge augmentation** for downstream moral event extraction (MEE) tasks. Concretely, the input passage is first used to retrieve K-scenarios pairs (K=3) from the moral scenario bank on which MOKA is pre-trained. Retrieved scenarios are combined with the original input to form a moral knowledge-enriched input. Next, we tag moral mentions on the fly, and follow the Lexcion Memory Access steps outlined in §4.1 to integrate moral word knowledge. MOKA is then 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. Meanwhile, as presented in Table 1 and 2, we disable $\mathcal{L}_{MLA}$ if we want to examine MOKA's efficacy when not explicitly informed of the specific morality(s) embodied by each moral mention.

For MOKA variants with single knowledge augmentation, we remove the corresponding module on which the variant is not pre-trained. For example, for MOKA augmented with moral scenario knowledge only, moral mention tagging and moral word knowledge integration (§4.1) are not applied.

5 Experiments

5.1 Tasks and Datasets

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

Task A: Moral foundation prediction. Conditioned on a document and *one* moral event span, make a 5-way judgment on the moral foundation for the given moral event.

Task B: Moral event trigger detection. Given a document, detect moral event triggers from the target sentence.

Task C: Moral event argument extraction. Given a document and *one* moral event span, produce

Model	MORAL EVENTS		eMFD Corpus	
	F1	Acc.	F1	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 ($\mathcal{L}_{MWL}$)	63.9	83.9	27.8	69.0
+ moral label association ($\mathcal{L}_{MLA}$)	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
+ RLM	62.3	83.8	30.1	70.3
Deontology (Hendrycks et al.)	62.5	83.6	30.5	70.5
+ RLM	62.2	83.5	30.4	70.4
Social chem (foundation; Forbes et al.)	62.2	83.7	32.4	70.6
+ RLM	64.1	84.0	32.5	70.7
MOKA with dual moral knowledge augmentation				
Delphi (moral judgement; Jiang et al.)	63.3	83.6	32.9	70.7
- $\mathcal{L}_{MLA}$	63.9	84.1	32.1	70.6
Deontology (Hendrycks et al.)	64.0	84.0	32.9	70.8
- $\mathcal{L}_{MLA}$	64.2	84.0	**34.3**	**71.1**
Social chem (foundation; Forbes et al.)	**65.3**	**84.3**	33.7	71.0
- $\mathcal{L}_{MLA}$	64.1	84.0	33.4	71.0
Improvements over best baseline	2.7%	0.8%	3.9%	2.9%

Table 1: Weighted F1 and accuracy on MORAL EVENTS and eMFD (Hopp et al., 2021) for Task A (average of 5 runs). Best results are in **bold**. MOKAs that outperform all baselines are highlighted on a scale of 5 red shades. "+" and "-" indicate the inclusion or exclusion of a particular training objective. MOKA augmented with dual moral knowledge (RLM enabled) achieve better performances across the board by notable margins. Full results and color scheme explanations are in Table A8.

triplets in the form of moral agents, patients, and a 10-way morality inference. This demands profound moral reasoning skills to correctly understand the interplay between participants and moralities.

As eMFD (Hopp et al., 2021) only annotates moral foundations but not event attributes, it is 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.

We split MORAL EVENTS by chronological order, and use the 90 news articles published in the 2nd half of 2022 as the test set. We sample a subset of articles from eMFD, and partition them randomly on the article level. Table A5 shows the detailed statistics of splits on both datasets.

5.2 Baselines and MOKA Variants

For Task A, we follow Alhassan et al. (2022) and compare with encoder-only models: RoBERTa (Liu et al., 2019) and its variant continually trained on news, POLITICS (Liu et al., 2022). We include a dictionary approach (Brady et al., 2017), where the moral foundation is determined by the presence of moral words defined in morality lexicons (Graham et al., 2009; Buttrick et al., 2020). [6] For Task B and C, since they are newly introduced in this work to study different aspects of moral events, we follow the EE literature and compare with a SOTA baseline, DEGREE (Hsu et al., 2022). For all tasks, we also compare with Flan-T5 (Chung et al., 2022) with downstream fine-tuning only, and ChatGPT (*gpt-3.5-turbo*).[7]

We consider three **MOKA variants**. First, with **moral word knowledge augmentation only**, we experiment with pretraining on `Morality Bank` only with $\mathcal{L}_{LM}$ and $\mathcal{L}_{MV}$ objectives. We then incrementally add the new $\mathcal{L}_{MWL}$ and $\mathcal{L}_{MLA}$ objectives. For **moral scenario knowledge augmentation only**, we connect MOKA with one Moral Scenario Bank at a time,[8] and test the effectiveness of the RLM objective. Putting all together, we obtain the full model with the **dual moral knowledge augmentation**. We further examine MOKA's efficacy with and without $\mathcal{L}_{MLA}$ in the *moral word knowledge pre-training stage*.

5.3 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 criterion as in prior work (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 1 shows the results for Task A. Performances on MORAL EVENTS and eMFD exhibit distinct trends. MORAL EVENTS follows a natural moral foundation distribution (Table A7), whereas eMFD has a roughly even distribution. Encoder-only models show strong performances on both datasets, where RoBERTa-large achieves the best F1 scores on MORAL EVENTS. ChatGPT, despite its stunning capability, struggles to understand

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

[7] Prompts, adapted from Li et al. (2023), are shown in Table A10.

[8] We also experimented with conflating all Moral Scenario Banks together. However, this did not improve performance.

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 ($\mathcal{L}_{MWL}$)	45.6	55.9	47.6	59.8	**31.5**	41.7
+ moral label association ($\mathcal{L}_{MLA}$)	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
+ 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
+ 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
+ RLM	47.5	56.0	48.8	60.5	31.0	41.7
MOKA with dual moral knowledge augmentation						
Delphi (moral judgement; Jiang et al.)	47.4	56.8	48.1	60.3	30.2	40.5
- $\mathcal{L}_{MLA}$	46.7	57.2	47.6	60.0	30.2	40.5
Deontology (Hendrycks et al.)	46.8	**58.2**	47.9	60.3	30.9	41.1
- $\mathcal{L}_{MLA}$	**48.1**	57.3	48.2	61.0	30.7	41.1
Social chem (foundation; Forbes et al.)	46.5	58.1	48.4	61.0	30.5	40.8
- $\mathcal{L}_{MLA}$	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 2: Results on MORAL EVENTS for Tasks B and C (average of 5 runs). Best results are in **bold**. MOKAs that outperform all baselines are highlighted on a scale of 5 red shades. "+" and "-" indicate the inclusion or exclusion of a particular training objective. MOKA augmented with dual knowledge achieve consistently better performances on trigger detection and morality inference, while the best results on participant extractions (i.e., *agent* and *patient*) are reached by single-knowledge variants. Full results and color scheme explanations are in Table A9.

and discern moral foundations. Flan-T5-large, the backbone model in MOKA, yields unsatisfying results, especially on eMFD, due to a lack of ethics-related documents in its pertaining stage (Jiang et al., 2021). In contrast, moral knowledge augmentation in MOKA improves Flan-T5's moral reasoning capabilities by 35% (F1 of 34.3 vs. 25.4).

Table 2 presents model results on Task B and C. Similar to Task A, ChatGPT performs worse than specialized EE systems. While DEGREE is a SOTA model in the general domain, it does not outperform fine-tuning a Flan-T5 model, highlighting the unique challenges posed by moral event understanding. On the other hand, when equipped with dual moral knowledge, MOKA yields the best results for trigger detection and morality inference. Particularly, the 7% performance gain on morality inference can be attributed to MOKA effectively assimilating moral knowledge at different granularities after two stages of moral knowledge-centric pre-training. Note, however, that single-knowledge variants reach the best participant extraction results, and our hypothesis is that the injected moral knowledge does not make participants available in moral reasoning. Further, as shown in Table A6, we study moral reasoning abilities across different event statuses: 1) Triggers of *Actual* events are easier to detect than others; 2) Morality of *speculative* events can be better identified, due to a higher usage of explicit moral language.

6 Further Analyses

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**), and go beyond that to show how the selective reporting of moral *narratives* reveals more subtle and asymmetrical forms of bias (**RQ2**).

RQ1: Does moral language usage correlate with media ideology? Figure 3 shows that more extreme outlets unsurprisingly tend to use more moral language overall, whereas the centrist media use it least. The most frequent moral foundation is `Care/Harm`, in line with findings on social media text (Figure 4; Hoover et al., 2020; Trager et al., 2022). Both news media and social media use relatively little `Sanctity/Degradation`, which measures religious purity and disgust and is rarely reported in the news. However, news media 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, so-

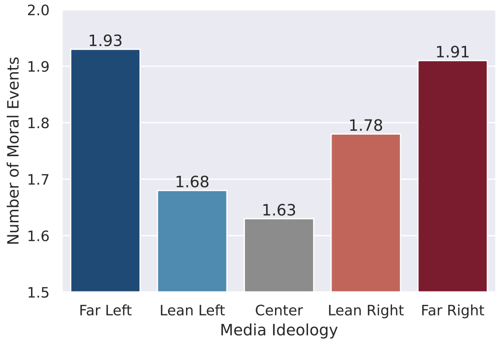


Figure 3: Number of moral events in each 100-word segment. Highly partisan media outlets tend to include more moral language than non-partisan ones.

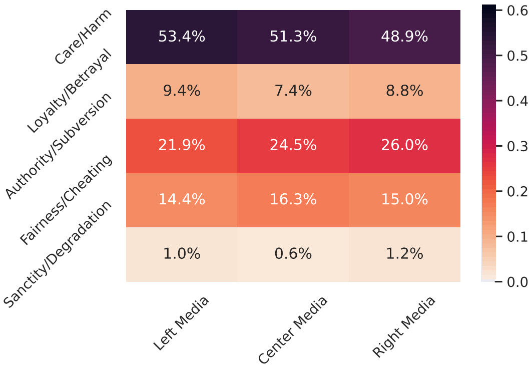


Figure 4: Employed moral foundation distribution by media outlets of different ideologies.

cial media covers more Fairness/Cheating, due to its greater focus on explicit morality as seen in AITA Reddit forum which has a special focus on personal ethical violations (Alhassan et al., 2022). Finally, within these general tendencies, we also find that left-leaning media focus more on Care/Harm, while right-leaning media focus more on Authority/Subversion, in line with MFT (Graham et al., 2009).

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 appear in at least two news articles were coded by a domain expert for their partisan leaning on a binary left/right scale, yielding 197 coded entities and $1,253$ associated events. Figures A1 and A2 show the correlations between agent-patient relationship and outlet ideologies for the two most prevalent foundations, Care/Harm and Authority/Subversion. This reveals 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, 2010; Graham et al., 2009).

Within Care/Harm (Figure A1), the left media report relatively more Right-harm-Left events than the right media do, and vice versa. Interestingly, an asymmetry is observed that media across different ideologies all report more Right-harm-Left events than the reverse (i.e. Left-harm-Right). That applies to centrist outlets as well, which show a pronounced tendency of reporting more Care/Harm where the Right entity is the agent.

For Authority/Subversion (Figure A2), we find both left and right outlets report more Authority from Right-to-Left, while centrist media are once again more focused on Right-agent events overall. 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, while centrist media also report more Right-subverts-Left events.

To summarize, mainstream media strive for balance in ideological language, entities, and even expressed values, but when we examine agent-value-patient triplets, ideological differences become evident. We found both important **left-right asymmetries**, and the **distinctive behavior of centrist media**, which overwhelmingly focuses on Right agents. These illustrate the importance of event-level morality analysis in political news.

7 Conclusion

We studied the task of moral event extraction—a novel reasoning task with the objective of, given unstructured text, producing structural representations for morality-bearing events including their triggers, participating entities, and embodied morality. To support this study, we curate a new dataset, MORAL EVENTS, including $5,494$ structured annotations. 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, derived from moral words and moral scenarios. Further analyses reveal the effectiveness of using moral events to discern ideological biases even when outlets report seemingly objective events.

Acknowledgments

This work is supported in part through National Science Foundation under grant IIS-2127747, Air Force Office of Scientific Research under grant FA9550-22-1-0099, and computational resources and services provided by Advanced Research Computing (ARC), a division of Information and Technology Services (ITS) at the University of Michigan, Ann Arbor. We appreciate ACL ARR reviewers for their helpful feedback. We thank Jahnavi Enaganti, Yunwoo Chang, Miya Brado, Syafawani Abdul Rahim and Natalie Wilson for their efforts in MORAL EVENTS construction.

Limitations

GPU resources. The framework proposed in this work is an encoder-decoder based generative model. It is thus more time-consuming than standard discriminative models for training and evaluation, which in turn results in a higher carbon footprint. Specifically, we train each model ($\sim$ 770 million parameters) on 1 single NVIDIA RTX A40 with significant CPU and memory resources. The training time for each model ranges from 1 to 3 days, depending on the configurations.

MORAL EVENTS and Annotations. While we offer comprehensive training guidelines and implement necessary quality control processes, users of our MORAL EVENTS might not fully agree with our annotated structural annotations of moral events. We deeply respect different views, especially those from underrepresented groups, and are eager to explore variations in how individuals from different geographical backgrounds interpret these events in future work.

Due to budget constraints, we were only able to annotate 474 articles and $5,494$ moral events. However, it is worth noting that, the annotated events in MORAL EVENTS already outnumber one of the most prevalent event extraction dataset – ACE 2005 (Doddington et al., 2004). Future endeavors might leverage AI systems (e.g., ChatGPT) to scale up moral event annotations with minimal human efforts.

Moral Foundation Theory. In this study, we build our approach MOKA on top of a prominent social psychology theory – Moral Foundation Theory (MFT). MFT, however, has its own *cultural bias*. That is, the theory is largely based on research conducted in Western cultures, particularly in the United States. As such, the concluded five moral dimensions might not be universally applicable. Furthermore, we assume a *static nature* in MFT, i.e. there is a stable set of moral foundations. However, recent work embarked on splitting the dimension of `Fairness` into `Equality` and `Proportionality` (Atari et al., 2023), and extending the original MFT to include `Liberty` (Iyer et al., 2012) and `Honor` (Atari et al., 2020), which need to be taken into account in future modeling as well.

Ethical Consideration

MORAL EVENTS collection. All news articles were collected in a manner consistent with the terms of use of the original sources as well as the intellectual property and the privacy rights of the original authors of the texts, i.e., source owners. During data collection, the authors honored privacy rights of content creators, thus did not collect any sensitive information that can reveal their identities. All participants involved in the process have completed human subjects research training at their affiliated institutions. We also consulted Section 107[9] of the U.S. Copyright Act and ensured that our collection action fell under the fair use category.

MORAL EVENTS annotation. In this study, manual work is involved. All the participants are college students, and they are compensated fairly (15 USD/hr per school policy). We hold weekly meetings to give them timely feedbacks and grade them quite leniently to express our appreciation for their consistent efforts. Lastly, they consent that their annotated data can be further repurposed and distributed for research purposes.

[9] https://www.copyright.gov/title17/92chap1.html#107.

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.

Mohammad Atari, Jesse Graham, and Morteza Dehghani. 2020. Foundations of morality in iran. *Evolution and Human behavior*, 41(5):367–384.

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani. 2023. Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*.

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.

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 pre-trained 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.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *ArXiv*.

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.

Ravi Iyer, Spassena Koleva, Jesse Graham, Peter Ditto, and Jonathan Haidt. 2012. Understanding libertarian morality: The psychological dispositions of self-identified libertarians.

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.

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.

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.

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.

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.

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*.

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.

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.

Xinliang Frederick Zhang, Carter Blum, Temma Choji, Shalin Shah, and Alakananda Vempala. 2024. ULTRA: Unleash LLMs' potential for event argument extraction through hierarchical modeling and pair-wise refinement. *ArXiv*, abs/2401.13218.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Merging generated and retrieved knowledge for open-domain QA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4710–4728, Singapore. Association for Computational Linguistics.

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

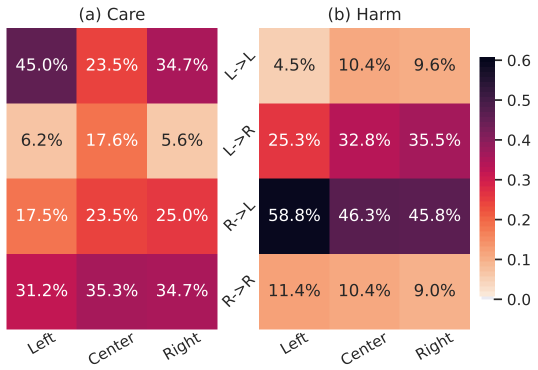


Figure A1: Correlation among agent-patient relationships, media outlet ideologies, and `Care`-/`Harm`-bearing moral events. Each percentage indicates the proportion of reporting a certain agent-patient interaction, and each column sums up to 100%. For example, 6.2% means that, among all `Care`-bearing events reported by left-leaning media, 6.2% of them are enabled by a `Left`-leaning entity and affecting a `Right`-leaning entity.

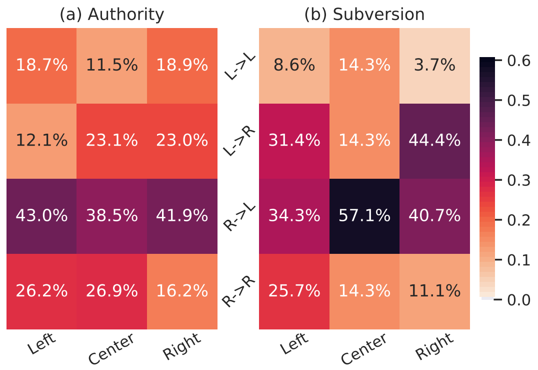


Figure A2: Correlation among agent-patient relationships, media outlet ideologies, and `Authority`-/`Subversion`-bearing moral events. Each percentage indicates the proportion of reporting a certain agent-patient interaction, and each column sums up to 100%.

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 A1: Krippendorff's alpha on various event attributes for revised (approach a) and merged (approach b) event annotations.

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 A2: Top-30 frequent entities in MORAL EVENTS sorted by their frequencies, i.e., the number of articles in which an entity appears.

A 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 action 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 that concisely depicts the event/action 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 *target sentence*. In line with ACE 2005 (Doddington et al., 2004) and the LDC annotation guideline[10], the moral event also includes an **event trigger** that can best characterize the moral action.

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*

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

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 A3: Sample examples from our constructed MORALITY BANK. The seed words used to crawl sentences are highlighted in red. The morality of each sentence is determined by the morality of the corresponding seed word. For each moral mention in text, it is tagged with a special symbol pair, <Morality> and </Morality>, and its embodied morality is visually represented using $_{\text{subscript}}$. These morality-bearing example sentences are employed to train MOKA during *moral word knowledge pre-training stage*.

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 A4: 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 full set of plausible labels for each scenario bank. These (Scenario, Label) pairs are employed to train MOKA during *moral scenario knowledge pre-training stage*. Note, we do not use all data points from Delphi for the sake of training efficiency, but downsample them to the numbers indicated in the the last column *# of examples*.

(Saurí and Pustejovsky, 2009; Lee et al., 2015). We further divide *non-actual* into *intentional* and *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).

B MORAL EVENTS 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.

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 names, 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

	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 A5: 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.

	Trigger Detection				Morality Prediction			
	Actual	Intentional	Speculative	Overall	Actual	Intentional	Speculative	Overall
Flan-T5 (large)	47.9	44.3	**43.1**	46.2	53.2	51.8	55.6	53.8
MOKA	50.2	**44.4**	42.1	46.9	**56.8**	**52.4**	**60.4**	**57.7**
w/o moral label association	**50.3**	43.9	**43.1**	**47.2**	56.6	**52.4**	60	57.4

Table A6: F1 results of select models by event status for trigger detection (task B) and morality identification (part of Task C). The performances of the dual-knowledge augmented MOKA and its variant are based on the aggregated results of MOKA trained on the three moral scenario banks reported in Table 2. Here, we have observed interesting, substantive findings: 1) *Actual* events are always easier to detect than *non-actual* ones (including both *intentional* and *speculative* events); and 2) In terms of morality prediction, it's generally easier to predict the morality for *speculative* events. This is attributed to the fact that speculative events are usually presented in the form of "entity A speculating entity B doing something good/bad to entity C", which tends to have a higher usage of explicit moral languages.

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.

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.

C 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. 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. Agreement levels are included in table A1. 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. 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.

	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 A7: Distribution of moralities in moral event annotations in MORAL EVENTS. Numbers in *Virtue* and *Vice* columns are raw counts of annotated moralities.

Likewise, a similar quality control study is conducted on annotated articles undergoing **merging**, approach (b) as described in §3. Agreement levels are included in table A1. 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** match 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 designed two-pass annotation process.

D 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>`) 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.

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

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

where $\mathcal{M}$ denotes the morality 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 9th layer in MOKA encoder.

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 $\{(t_i, m_i, V_i)\}$, where $t_i = x_j$ for some $j \in [1, T]$. Here, t_i is a moral mention in $\mathbf{x}$, m_i is the corresponding moral word, and $V_i = \{v_{i,1}, v_{i,2}, ...\}$ is the set of associated moralities.

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

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

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

where α_{m_p} is computed using eq. (2). We then denote $\mathbf{A}_i$ as the set of embodied moralities' aggregated attention scores, i.e., $\mathbf{A}_i = \{\mathrm{A}_{i,1}, \mathrm{A}_{i,2}, ...\}$ where $|\mathbf{A}_i| = |V_i|$. For simplicity, we use $\mathbf{A}_i^C$ 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 **minimize** *multi-label margin loss* as shown in eq. (4):

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

$$:= \frac{\sum_{y \in \mathbf{A}_i} \sum_{z \in \mathbf{A}_i^C} (1 + z - y)}{10} \tag{5}$$

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

Model	Moral Events		eMFD Corpus	
	F1	Acc.	F1	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 ($\mathcal{L}_{MWL}$)	63.9	83.9	27.8	69.0
+ moral label association ($\mathcal{L}_{MLA}$)	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
+ RLM	63.2	84.2	30.3	70.3
Delphi (moral judgement; Jiang et al.)	63.7	84.1	30.4	70.4
+ RLM	62.3	83.8	30.1	70.3
Deontology (Hendrycks et al.)	62.5	83.6	30.5	70.5
+ RLM	62.2	83.5	30.4	70.4
Justice (Hendrycks et al.)	62.5	83.7	30.4	70.3
+ RLM	62.4	83.6	31.8	70.6
Social chem (judgement; Forbes et al.)	63.6	84.2	30.0	70.2
+ RLM	62.9	83.6	30.7	70.4
Social chem (foundation; Forbes et al.)	62.2	83.7	32.4	70.6
+ RLM	64.1	84.0	32.5	70.7
Social chem (morality; Forbes et al.)	62.7	83.8	30.0	70.3
+ RLM	63.3	84.1	32.5	70.6
MOKA with dual moral knowledge augmentation				
Delphi (moral agreement; Jiang et al.)	64.4	84.3	34.0	71.0
- $\mathcal{L}_{MLA}$	63.3	84.0	33.2	70.8
Delphi (moral judgement; Jiang et al.)	63.3	83.6	32.9	70.7
- $\mathcal{L}_{MLA}$	63.9	84.1	32.1	70.6
Deontology (Hendrycks et al.)	64.0	84.0	32.9	70.8
- $\mathcal{L}_{MLA}$	64.2	84.0	34.3	71.1
Justice (Hendrycks et al.)	64.0	84.0	32.9	71.0
- $\mathcal{L}_{MLA}$	63.7	84.1	33.3	71.0
Social chem (judgement; Forbes et al.)	64.3	84.2	32.7	70.9
- $\mathcal{L}_{MLA}$	64.2	84.3	33.4	71.1
Social chem (foundation; Forbes et al.)	**65.3**	**84.3**	33.7	71.0
- $\mathcal{L}_{MLA}$	64.1	84.0	33.4	71.0
Social chem (morality; Forbes et al.)	64.5	83.9	**34.6**	**71.3**
- $\mathcal{L}_{MLA}$	63.8	84.0	33.3	70.9
Improvements over best baseline	2.7%	0.8%	4.8%	3.2%

Table A8: Full 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**. "+" and "-" indicate the inclusion or exclusion of a particular training objective. *Color scheme*: MOKA and its single-knowledge-augmentation 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.

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 ($\mathcal{L}_{MWL}$)	45.6	55.9	47.6	59.8	**31.5**	41.7
+ moral label association ($\mathcal{L}_{MLA}$)	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
+ 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
+ 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
+ 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
+ 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
+ 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
+ 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
+ RLM	47.2	55.5	48.7	60.7	31.0	41.5
MOKA with dual moral knowledge augmentation						
Delphi (moral agreement; Jiang et al.)	46.5	56.9	48.4	60.5	30.5	41.0
- $\mathcal{L}_{MLA}$	47.3	57.2	47.5	60.6	30.7	40.9
Delphi (moral judgement; Jiang et al.)	47.4	56.8	48.1	60.3	30.2	40.5
- $\mathcal{L}_{MLA}$	46.7	57.2	47.6	60.0	30.2	40.5
Deontology (Hendrycks et al.)	46.8	58.2	47.9	60.3	30.9	41.1
- $\mathcal{L}_{MLA}$	**48.1**	57.3	48.2	61.0	30.7	41.1
Justice (Hendrycks et al.)	46.9	57.4	48.6	61.1	31.0	41.2
- $\mathcal{L}_{MLA}$	47.4	56.9	48.0	60.9	31.1	41.1
Social chem (judgement; Forbes et al.)	46.5	57.7	47.9	60.6	29.7	40.6
- $\mathcal{L}_{MLA}$	46.8	57.7	47.5	60.9	30.1	40.3
Social chem (foundation; Forbes et al.)	46.5	58.1	48.4	61.0	30.5	40.8
- $\mathcal{L}_{MLA}$	46.7	57.7	48.2	60.5	30.0	40.1
Social chem (morality; Forbes et al.)	47.0	58.2	48.0	60.6	30.5	40.5
- $\mathcal{L}_{MLA}$	46.7	**58.5**	47.9	60.8	30.2	40.9
Improvements over best baseline	4.1%	7.9%	0.6%	1.8%	2.6%	2.2%

Table A9: 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**. "+" and "-" indicate the inclusion or exclusion of a particular training objective. *Color scheme*: MOKA and its single-knowledge-augmentation 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 A10: Prompts used to test ChatGPT’s moral reasoning capability, adapted from Li et al. (2023). For task A, although we prompt ChatGPT to predict *start* and *end* indexes in its structural output, we only use its predicted value of the *trigger* field, due to ChatGPT’s insufficient numerical reasoning capabilities.