Moral association graph: A cognitive model for moral inference

Aida Ramezani (armzn@cs.toronto.edu)
Department of Computer Science, University of Toronto

Yang Xu (yangxu@cs.toronto.edu)
Department of Computer Science, Cognitive Science Program, University of Toronto

Abstract

Moral inference is an emerging topic of critical importance in artificial intelligence. The contemporary approach often relies on language modelling to infer moral relevance or moral properties of a concept such as "smoking". This approach demands complex parameterisation and costly computation, and it tends to disconnect with psychological accounts of moralization. We present a simple cognitive model for moral inference grounded in theories of moralization. Our model builds on word association network known to capture human semantics and draws on rich psychological data. We demonstrate that our moral association graph model performs competitively to state-of-the-art language models, where we evaluate them against a comprehensive set of data for automated inference of moral norms and moral judgment of concepts, and in-context moral inference. Moreover, we show that our model discovers intuitive concepts underlying moral judgment and is applicable to informing short-term temporal changes in moral perception.

Keywords: moral inference; word association; moralization; language model; artificial intelligence

Introduction

Aligning artificial intelligence (AI) systems with human values is one of the most critical challenges we face today, and a prerequisite to tackling this issue is to understand how human values work. Morality plays a central role in societal and individual values, and recent development in AI has offered new tools for studying human moral values at a comprehensive scale. A core approach to this development is automatic moral inference, or machine prediction of human moral values and judgments, which typically draws on information from large text corpora through language modelling. Are language models the optimal and only way to do moral inference? Here we offer a simple, alternative approach that grounds moral inference in psychological theories of moralization and intuitive human semantics.

There has been a growing interest in connecting AI with human morality. Development over the past decade includes the collection of large-scale moral judgments (Forbes et al., 2020; Hoover et al., 2020; Sap et al., 2020; Hendrycks et al., 2021; Trager et al., 2022), machine inference of moral values from text (Garten et al., 2016; Jia & Krettenauer, 2017; Mooijman et al., 2018; Johnson & Goldwasser, 2018; Emelin et al., 2021; Roy et al., 2021; Xie et al., 2020; Liscio et al., 2022; Trager et al., 2022), machine prediction of moral norms (Jentzsch et al., 2019; Schramowski et al., 2019, 2022; Ramezani & Xu, 2023; Haemmerl et al., 2023), and alignment of AI systems with human moral judgments (Hendrycks et al., 2020; Jiang et al., 2021; Ammanabrolu et al., n.d.; R. Liu et al., 2022; Lourie et al., 2021). These recent advances in moral inference from large text corpora allow us to analyze human moral values at an unprecedented scale, but this line of research is often disengaged with psychological accounts describing how human morality works.

One area of significant relevance coming from moral psychology is moralization, the process in which something that previously had no moral relevance becomes associated with moral values (Rozin et al., 1997; Rozin, 1999). Moralization constantly shapes our moral values toward activities such as smoking cigarettes (Rozin & Singh, 1999) and consuming meat (Feinberg et al., 2019), concepts such as new technologies (e.g., GMOs as in Clifford (2019); Inbar et al. (2020)), as well as individuals (e.g., political leaders as in Brandt et al. (2015)). Existing work has identified potential key factors in people's moralization of concepts. These factors relate to one's rationalization of perceived harms and benefits (or "moral piggybacking"), e.g., someone moralizing "eating meat" might be due to that it involves killing animals (Feinberg et al., 2019). Moralization may also depend on emotion such as the feeling of disgust toward cigarettes (Rozin & Singh, 1999; Brandt et al., 2015; Skitka et al., 2018). These psychological studies offer valuable insights into the workings of morality, but they rely on case studies and an experimental setting. Our goal is to connect computational and psychological approaches to build scalable and interpretable models for moral inference.

We propose the Moral Association Graph model, a framework designed to support intuitive moral inference grounded in human word association network (see Figure 1). Word association is derived from a psychological game involving participants who are presented with cue words and prompted to respond with the first word(s) coming to their mind (e.g., *cigarette→nicotine*). Data collected from word association experiments reflect how words or concepts are mentally represented and connected to one another and serve as a proxy of human semantic network (Collins & Quillian, 1969) and mental representations of word meaning (Deese, 1965; Nelson et al., 2004; De Deyne et al., 2019; Van Rensbergen et al., 2015; C. Liu et al., 2022). It is shown that word association better captures the human semantics, e.g., via representing multi-modal properties of concepts, in comparison to distributional semantic (language) models trained on

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* ©2024 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

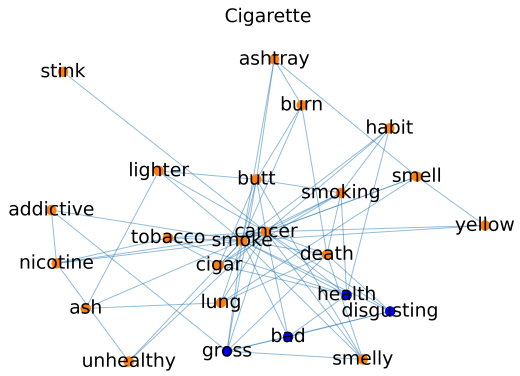

Figure 1: Word association network for the query concept *cigarette*. The edges indicate whether two concepts are responded by the same participant given the query concept, and the blue nodes indicate morally relevant concepts (e.g., *disgusting*, *gross*). Edge weights are omitted for visual clarity.

text (De Deyne et al., 2021, 2020).

We consider word association as an alternative to language models for moral inference because it captures people's immediate and intuitive reactions, some of which directly reflect moral perception (e.g., *cigarette→disgusting*). Furthermore, our approach is extremely simple: It requires no model training or fine-tuning, and it is parameter-free. Despite its simplicity, to our knowledge formal approach to moral inference with word association is an under-explored area. We show in a series of analyses that our moral association model either outperforms or achieves comparable performance to state-of-the-art generative language models in inferring people's moral norms and moral judgments at both conceptual and contextual levels. We also demonstrate how our model extracts meaningful key concepts that underlie people's moral judgment, and we apply our model to uncover changes in moral perception during the COVID-19 pandemic.

Computational framework

We describe our computational framework which is inspired by psychological work on moralization and built on the cognitive lexicon of word association. We also describe baseline models based on emotion and language modeling.

Moral association graph (MAG)

This model captures the strength of association between a query (or cue) word and target moral words based on graphs of word association. For example, a cue word like *cigarette* can be associated with concepts such as *bad, death, unhealthy*, which are moral words. We hypothesize that MAG should capture people's moral intuition, particularly how they moralize concepts (Rozin & Singh, 1999; Feinberg et al., 2019). To formulate MAG we use a dictionary of moral words denoted by M, and the adjacency matrix A. A is an asymmetric matrix of word association, where the rows represent the cue words and the columns represent the response or target words. Each entry, denoted by $A_{c,t}$, represents the number of participants who responded with the target word t given the cue word c ($A_{c,t}$ captures the association strength). MAG is formally specified as follows:

$$MAG(c) = \frac{\sum_{t \in T(c) \cap M} A_{c,t}}{\sum_{t \in T(c)} A_{c,t}} \quad (1)$$

Here $T(c)$ is the set of all target responses for the cue word c. This equation measures the proportion of target responses that falls within the moral lexicon, and we take the resulting score as an indicator of moral relevance. To operationalize this model we draw on the Small World of Words project (SWOW), a large-scale dataset of word association, which covers more than 12,000 words for English (SWOW-EN) (De Deyne et al., 2019), Dutch (SWOW-NL) (De Deyne et al., 2013), and Rioplatense Spanish (SWOW-RP) (Cabana et al., 2023). We gather first-order associations which offer information about the immediate mental connections among concepts. The responses in SWOW-EN are predominantly sourced from native English speakers in the United States. Participants in SWOW-NL are mainly located in Belgium and the Netherlands, and the SWOW-RP, focused on Rioplatense Spanish which is spoken in Latin America, includes responses from participants in Argentina and Uruguay. The SWOW-EN and SWOW-NL cover 100 annotations per cue word, while the SWOW-RP covers 70. The responses in all three datasets are normalized and spell-checked. Additionally, we use the `nltk` package for lemmatizing the words in SWOW-EN.[1]

We use Moral Foundations Dictionary (MFD) (Graham et al., 2009) as our base moral lexicon, which is one of the largest lexical resources developed for the Moral Foundations Theory (MFT) (Graham et al., 2013), and has been widely used to study people's moral values through linguistic data (Garten et al., 2016; Mooijman et al., 2018; Hoover et al., 2020; Mendelsohn et al., 2020; Xie et al., 2019). Using the up-to-date version of MFD (Frimer et al., 2017), we identify a total of $1,705$ words as our moral lexicon.

Alternative to moral association, we consider a baseline **Emotion Association Graph model (EAG)** given the role of emotion in moralization (Rozin, 1999). We do so by replacing the moral lexicon with an emotion lexicon in Equation 1. This model captures the strength of emotions associated with a concept. We use the emotion lexicon specified in a prior study ($n = 626$) (Xu et al., 2021), compiled from multiple studies on emotion classification (Johnson-Laird & Oatley, 1989; Ekman, 1999; Fehr & Russell, 1984; Shaver et al., 1987)[2].

Baselines using language models

We consider several baselines drawing on existing AI and NLP work on moral inference. One study uses the Moral

[1] Access datasets and the metadata at `https://smallworldofwords.org/en/project/research`.

[2] We further use the Google Translate API (`googletrans`) to translate the moral and the emotion lexicons to Dutch and Spanish.

Foundations Dictionary (Graham et al., 2009) and word embeddings to infer moral relevance of individual concepts (Xie et al., 2019). Here moral relevance for a query is estimated as a probability distribution based on the proximity of that query to the moral word clusters in semantic space. This model also estimates the probability distribution of a concept with respect to different moral foundations. We replicate this work using Word2Vec embeddings (Mikolov et al., 2013) from Google Ngrams, Dutch embeddings of WikiPedia (Tulkens et al., 2016), and Spanish Billion word Corpus and Embeddings (Cardellino, 2016) for English, Dutch and Spanish.

Other studies have used contextual language models for moral inference (Jentzsch et al., 2019; Schramowski et al., 2022; Alhassan et al., 2022). In one prominent line of work, BERT-based sentence representations (Reimers & Gurevych, 2019) of a set of morally relevant actions (e.g., killing, helping) are used to construct a morally imbued subspace that distinguishes right from wrong (Jentzsch et al., 2019; Schramowski et al., 2022). The moral score of a query is then determined by the similarity between the query and the moral subspace, where values around 0 indicate moral neutral, while values close to +1 or -1 signify high moral relevance. By using multilingual language models such as XLM-R (Conneau et al., 2020; Reimers & Gurevych, 2020), this model can be extended to identify moral values in different languages (Haemmerl et al., 2023). For our baseline, we employ the same setting in our experiments to explore the moral values of different concepts in English, Dutch, and Spanish.[3]

Finally, we consider generative large language models, such as GPT-3, that encode people's moral biases and preferences (Simmons, 2023; Fischer et al., 2023; Ramezani & Xu, 2023; Dillion et al., 2023). To compare against our MAG model, we probe GPT-3.5 and GPT-4 as strong baselines for identifying both the degree of moral relevance and the keywords for explaining people's moral judgments[4].

Interpreting moral association graph

Word association can offer meaningful insight into the processes underlying moralization. For example, health-related concerns and disgust-related feelings explain why some people regard smoking cigarettes as a moral issue (Rozin & Singh, 1999). Similarly in word association, if a participant's first, second, and third responses to the cue word *cigarette* are *wrong*, *smell*, and *nausea*, it suggests a negative evaluation, indicating a shared association between the smell of cigarettes and feeling nauseous. Similar patterns across many participants strengthens the indication that these factors contribute to negative moral views on *cigarettes*.

To uncover these relationships efficiently, we propose a formal procedure utilizing the co-occurrence relationships among the response words. For a given cue word c, we construct an undirected weighted graph (denoted by $G_c(V,E)$) with response words as the nodes. The edge weights represent the number of times two response words were mentioned by the same participant. In our previous example, the words *wrong*, *smell*, and *nausea* would all be connected, forming a triangle. Using the moral lexicon described, we then start a random walk that initiates from the moral words in this graph and continues the until convergence. Figure 1 visualizes this graph and the moral words for *cigarette*. Using a similar random walk process as the previous work in sentiment inference (Hamilton et al., 2016), we estimate the probability of arriving at a node during the walk based on its proximity to the words in the moral lexicon in Equation 2 below:

$$p^{(t+1)} = \beta \tilde{A} p^{(t)} + (1-\beta) m. \tag{2}$$

$$p_v^{(0)} = \frac{1/|V| + MAG(v)}{\sum_{u \in V} MAG(u) + 1}. \tag{3}$$

$$m_v = Z \begin{cases} degree(v) & v \text{ is a moral word} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Here, $\tilde{A}$ is the symmetric normalized adjacency matrix representing graph $G_c(V,E)$. As shown in Equation 3, $p^{(0)}$ is a vector of size $|V|$ where each entry corresponds to the MAG score of the respective word in the word association dataset, plus the smoothing factor of $\frac{1}{|V|}$. This vector controls the walk to propagate from words with the highest MAG scores (i.e., words with salient moral relevance). In Equation 4, we define m to be a vector of size $|V|$. For the moral words in V, their entries in m correspond to their weighted degree in $G_c(V,E)$, and for the rest of the words, it is set to be zero. This vector, further normalized by Z to sum up to 1, guides the random walk to remain close to the moral lexicon. Finally, β is a damping parameter that controls the divergence from moral words to longer paths. After convergence, we retrieve the top K words with the highest $p^{(t)}$ scores, which represent the underlying contexts wherein the cue words may be moralized.

Datasets for model evaluation

We compare and evaluate the models described in three main tasks of moral inference, drawing on data of human moral norms, and people's moral judgments of individual concepts and in natural context.

Data for moral norm inference

The World Values Survey (WVS) is a publicly-available global research survey investigating people's beliefs and values over the globe (Inglehart et al., 2014a,b; Haerpfer et al., 2021). Previous studies have used text-based methodologies to predict global ratings in the World Values Survey. Their findings suggest that the way people use language provides valuable insights into their beliefs and values (Arora et al., 2023; Ramezani & Xu, 2023). Following these studies, we use the participants' aggregate ratings in the ethical section of WVS as the ground truth for assessing peo-

[3]We use `xlm-r-100langs-bert-base-nli-mean-tokens` from the `sentence-transformers` package to embed the queries.

[4]We use the `gpt-3.5-turbo` and `gpt-4` engines and a temperature of 0.5.

ple's moral norms.[5] The ethical section of the World Values Survey explores moral and ethical values by asking people's stances on issues such as *abortion*. To align the population and time course of WVS with the participants and time course of SWOW projects, we use WVS waves 5, 6, and 7 (2005-2022) in countries including USA (for SWOW-EN), the Netherlands (for SWOW-NL, as Belgium is absent in WVS), Argentina, and Uruguay (for SWOW-RP). We normalize and use the absolute values of WVS responses, where a score of 0 corresponds to a non-moral issue and a score of 1 corresponds to a highly moral issue. The number of responses to each question varies from $1,000$ to $6,000$ participants.

Data for conceptual moral inference

To evaluate moral inference at the concept level, we use the extended Moral Foundations Dictionary (eMFD) (Hopp et al., 2021) which is a dictionary-based resource for extracting moral foundational content from text. By using human annotations on $2,995$ news articles, this dataset provides probability scores to a set of $3,270$ English words, indicating the likelihood of their mention in an article expressing a certain moral foundation.

Data for contextual moral inference

We use three datasets that evaluate moral inference in natural context. The first dataset is Moral Foundations Twitter Corpus (MFTC) (Hoover et al., 2020), which includes more than 30,000 tweets posted in different social discourses of Baltimore protests, US presidential election of 2016, hate speech language, etc. Each tweet is annotated by at least three annotators with the moral foundational labels. The annotators can annotate a tweet as "non-moral" indicating that the tweet discusses no moral issues. In our analysis, tweets that receive the "non-moral" label from all the annotators are considered as non-moral, and tweets that receive no "no-moral" label are considered as morally relevant. We retrieve $9,759$ moral tweets and $2,651$ non-moral tweets from MFTC.

The second dataset is Moral Foundations Reddit Corpus (MFRC) (Trager et al., 2022), which includes more than 16,000 Reddit posts from 12 difference subreddits. Similar to MFRC, each post is annotated by at least three human annotators. We use similar procedures to MFTC identifying $3,925$ moral posts and $3,208$ non-moral posts in MFRC.

The third dataset is SOCIAL-CHEM 101 (Forbes et al., 2020), which was used to train language models for moral norm analysis. The dataset contains 292,000 short text snippets, called the rules-of-thumb (RoT), such as "It's kind to sacrifice your well-being to take care of a sick person". We identify $75,615$ distinct RoTs labeled to be related to morality or ethics, and $190,658$ to be non-moral. We tokenize and lemmatize the text snippets all three datasets with the `nltk` package for our experiments.

[5] We follow the terms and conditions for using this resource, detailed at `https://www.worldvaluessurvey.org`.

Results

Evaluation of moral norm inference

We first evaluate models in inferring moral norms across cultures. We use WVS question keywords as queries for our models and intersect those with the word association data we used to construct our models MAG and EAG. We compare the performance of MAG with baseline models including EAG, and language models based on Word2Vec embedding (Xie et al., 2019), BERT embedding (Schramowski et al., 2022; Haemmerl et al., 2023), and GPT. The results in Figure 2 show that our MAG model consistently outperforms all models (except for GPT-4) in every case, indicating that the mental associations between different concepts and the moral lexicon capture people's intuitions about the morality of different concepts, without having to be trained on extensive textual data or be refined with human feedback through reinforcement learning (Ouyang et al., 2022). Consistent with previous studies probing morality in language models (Haemmerl et al., 2023), we find that performance of baselines decreases in Dutch and Spanish datasets, suggesting possible misrepresentation of moral values in non-English language models, and word embeddings. We also note that the MAG model has the lowest performance for Rioplatense Spanish. This dialect is not differentiated from Spanish in Google Translate, which could have hurt the performance of the MAG model, since it uses translated Spanish moral words.

Evaluation of conceptual moral inference

We next assess the models by comparing their predictive outputs for different concepts against the moral foundational probabilities collected in the extended Moral Foundations Dictionary (eMFD) (Hopp et al., 2021). In order to evaluate our framework in this fine-grained setting, we replace the overall MFD dictionary with moral foundational-specific words and adapt our MAG model to estimate the strength of association to each moral foundation. This modification yields five distinct moral-foundational association scores for each query concept. We compare these scores with the ground-truth data in eMFD and summarize the findings in Table 1. From the baselines, only the Word2Vec model developed by Xie et al. (2019) can distinguish between different moral foundations, but we also probe GPT-3.5 to provide moral foundational scores for different query words. The results show that across all moral foundations, MAG consistently outperforms the Word2Vec and GPT-3.5 models (Xie et al., 2019), suggesting again that the connections between words and their associative words with moral foundations offer reliable inference at the concept level.

Evaluation of contextual moral inference

To further assess our framework, we evaluate models on predicting moral relevance in natural sentences. We use the established moral datasets—MFTC (Hoover et al., 2020), MFRC (Trager et al., 2022), and SOCIAL-CHEM 101 (Forbes et al., 2020) as described. For each text snip-

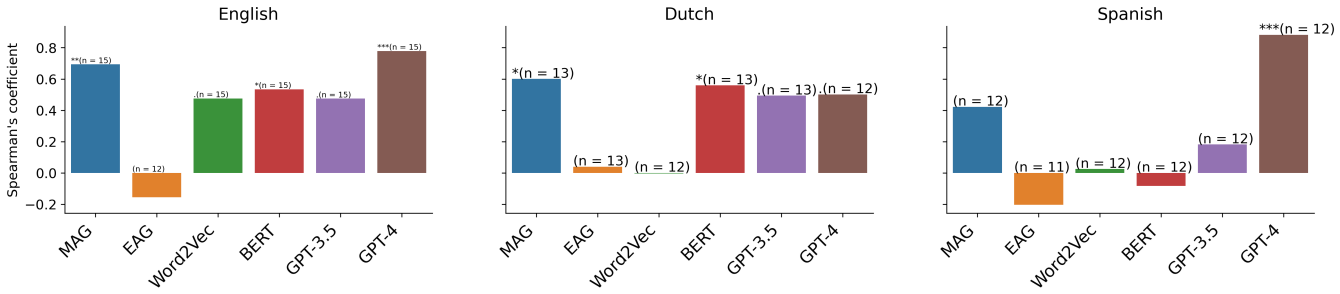

Figure 2: Results of moral norm inference based on correlation between empirical data from World Values Survey and inferred ratings from Moral Association Graph model (MAG), Emotion Association Graph model (EAG), and language model baselines (Word2Vec embedding (Xie et al., 2019), BERT embedding (Schramowski et al., 2022; Haemmerl et al., 2023), GPT-3.5, and GPT-4). The asterisks indicate the significance levels ("*", "**", "***" for $p < 0.05, 0.01, 0.001$ respectively) based on Spearman's rank correlation.

Moral Foundation	MAG	Word2Vec	GPT-3.5
Care/Harm ($n = 1895$)	**0.291**	0.28	0.28
Fairness/Cheating ($n = 1514$)	**0.232**	0.193	0.20
Authority/Subversion ($n = 1737$)	**0.301**	0.192	0.12
Loyalty/Betrayal ($n = 1714$)	**0.212**	−0.121	0.11
Sanctity/Degradation ($n = 1893$)	**0.246**	0.222	0.19
All moral foundations ($n = 8753$)	**0.199**	0.148	0.20

Table 1: Results of moral inference at the concept level. The first column shows the moral foundations and sample sizes. The second and third columns show correlations between empirical eMFD data and our MAG model prediction, Word2Vec and GPT-3.5 baselines. All values are statistically significant ($p \leq 10^{-4}$) shown with Spearman's rank correlation coefficients.

Model	Dataset 1 MFRC ($n = 7,125$)	Dataset 2 MFTC ($n = 11,910$)	Dataset 3 SOCIAL-CHEM 101 ($n = 265,898$)
MAG	0.475	**0.276**	**0.244**
EAG	0.256	0.105	0.067
Word2Vec	**0.522**	0.266	0.237
BERT	0.329	0.138	0.066
GPT-3.5	0.308	0.138	0.133
GPT-4	0.375	0.209	0.187

Table 2: Results of moral inference in natural context. Different models predict the moral relevance of articles in three datasets: MFRC (Trager et al., 2022), MFTC (Hoover et al., 2020), and SOCIAL-CHEM 101 (Forbes et al., 2020). All values are statistically significant ($p \leq 10^{-4}$) shown with coefficients from Spearman's rank correlation.

pet, we lemmatize its words and apply our models to assign a moral score to each word lemma. The overall moral relevance score for an article is calculated based on the average moral scores. Table 2 shows the predictive performance of different models using a correlation test between articles' aggregate moral scores and their moral relevance labels. Similar to the previous experiments, our MAG model reproduces the ground-truth moral relevance labels outperforming the majority of the baselines and performing on par with the Word2Vec based moral sentiment inference model (Xie et al., 2019).

Retrieving key concepts in moralization

Our MAG model also offers insight into the intuitive process of moralization. Our methodology for keyword retrieval leverages the relationship between first-level, second-level, and third-level association words to uncover the potential cognitive processes that give rise to moralization. Given that there are currently no gold standards for this task, we consider GPT-4 as a silver standard for measuring the success of our model. For a given query (e.g., *cigarette*), we prompt GPT-4 to provide a list of keywords explaining the reasons why some people consider this query as a moral issue. Using this silver standard, we assess our framework using the precision, recall, and F1-score in retrieving the top $K = 25$ keywords suggested by GPT-4. For the baseline, we use Word2Vec embedding model (Mikolov et al., 2013) trained on the Google Ngrams dataset and identify the top $K = 25$ semantic neighbors of a query term. Using a β value of 0.5, the word association model achieves the precision, recall and F1 values of 0.15, 0.10, and 0.12 while the Word2Vec model achieves 0.04, 0.04, and 0.04 respectively.

In Table 3, we show keywords identified by our association model, GPT-4, and WORD2VEC for *cigarette*. A quantitative analysis of our framework indicates its capability to identify meaningful connections between moralized concepts and their underlying contexts. For instance, aligned with the theoretical work on the moralization of smoking, our framework identifies keywords such as *cancer, unhealthy*, and *stink*, representing both the rationalization of harms associated with smoking cigarettes and the emotional responses to it (Rozin

Model	Keywords
MAG	smoke, health, cancer, unhealthy, tobacco dirty, cigar, stink, killer, stained, wrong together, smoking, disease, evil, waste hell, ignorant, death
GPT-4	health, addiction, secondhand smoke cancer , harmful, risk, death, pollution ethical, responsibility, choice, disease cost, danger, unhealthy, lungs, smoke damage, tobacco, habit
WORD2VEC	tobacco, smokes, cigs, cigarrette marlboros, smoking, ciggie unfiltered_camels, marlboro_lights newports, pall_malls, winstons smokeless_tobacco, smokers menthol_flavored, ciggies mccargar_bribed, cigarrettes marlboro_menthol

Table 3: Top keywords for the query concept *cigarette*, retrieved from different models for moral inference.

& Singh, 1999). Although GPT-4 serves as a competitive baseline for keyword retrieval, we observe that the keywords retrieved by our model exhibit conceptual similarities with those retrieved by GPT-4. For example, both models retrieve concepts related to *consumerism* and *waste*, offering meaningful insight into the moral reflection of *fashion*.

Applications to quantifying short-term moral change

We apply our framework to identify changes in people's moral perception in the COVID-19 pandemic. A section of the SWOW-RP dataset was gathered post-December 2020. Within this dataset, a subset of words were collected both before COVID-19 (December 2013 to March 2020) and during the pandemic (December 2020 to April 2022). These words were categorized into four groups of pandemic-related words (n = 107), emotion words (n = 119), routine words (n = 108), and control words (n = 150) (Laurino et al., 2023). Pandemic-related words refer to those that have gained new meanings or have been excessively used in relation to the pandemic (e.g., *protocol*). Emotion words correspond to feelings that could be affected by the pandemic (e.g., *anxiety*). Routine words describe daily activities impacted by the pandemic (e.g., *tourism*), and control words lack direct connection to the pandemic (e.g., *rain*). Large-scale experiments on this dataset reveal that pandemic-related words gained new senses and became more semantically associated with health and sanitary concepts during the pandemic (Laurino et al., 2023). Using the same dataset, we examine whether there are significant changes in the moral association of pandemic-related words. Specifically, we hypothesize that, among the words that have become more positively associated with moral values, pandemic-related words should exhibit the most substantial changes. Figure 3 (top) confirms our hypothesis: we identified 43 pandemic-related words that have acquired new moral associations. In comparison to the 64 control words with new moral associations, we observed that the degree of moral associations for pandemic-related words is significantly larger than that for control words (p-value from the Wilcoxon rank-sum test = 0.04). Additionally, pandemic-related words exhibit more substantial changes than routine words ($p \leq 0.01$). Compared to emotion words, the difference is marginally significant ($p = 0.07$). A permutation test comparing pandemic words with other groups confirms a more significant change for pandemic-related words ($p \leq 0.05$). Figure 3 (bottom) compares the precision in retrieving word groups among the top K words with the highest moral association change. As observed, pandemic terms are the most predominant in the top K words significantly exceeding chance for smaller values of K.

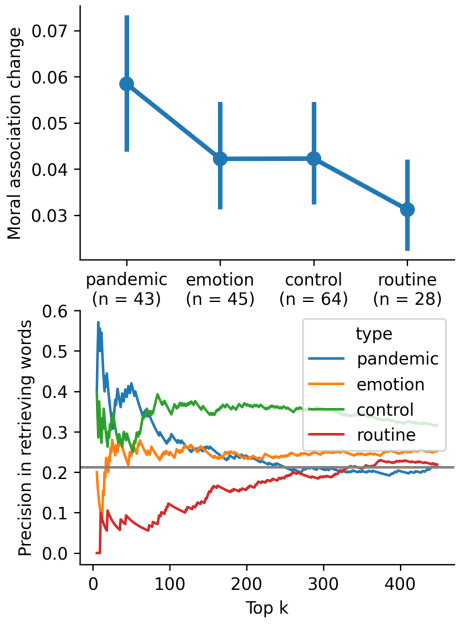

Figure 3: Top: Degrees of positive change in moral associations of different word groups. Bottom: Precision in retrieving top K words with the largest moral association change. The grey horizontal line shows chance level of retrieving the pandemic words.

Conclusion

We present a parameter-free, cognitive model of moral inference grounded in theories of moralization and human word association. Through rigorous evaluations across half a dozen moral datasets, we show that the mental associations between different concepts and moral words capture people's beliefs about the morality of these concepts, and provide an interpretable and promising computational framework for studying the underlying mechanisms of moralization. Future work may explore the relation of this framework to accounts of moral reasoning (e.g., Kleiman-Weiner et al. (2017); Barque-Duran & Pothos (2021); Jin et al. (2022)).

Acknowledgments

This research is supported by an Ontario Early Researcher Award #ER19-15-050 to YX.

References

Alhassan, A., Zhang, J., & Schlegel, V. (2022, June). 'Am I the Bad One'? predicting the moral judgement of the crowd using pre–trained language models. In N. Calzolari et al. (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 267–276). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2022.lrec-1.28

Ammanabrolu, P., Jiang, L., Sap, M., Hajishirzi, H., & Choi, Y. (n.d., July). Aligning to social norms and values in interactive narratives. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Arora, A., Kaffee, L.-a., & Augenstein, I. (2023, May). Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP] (C3NLP)* (pp. 114–130). Dubrovnik, Croatia: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.c3nlp-1.12 doi: 10.18653/v1/2023.c3nlp-1.12

Barque-Duran, A., & Pothos, E. M. (2021). Untangling decision routes in moral dilemmas: The refugee dilemma. *The American Journal of Psychology*, *134*(2), 143–166.

Brandt, M. J., Wisneski, D. C., & Skitka, L. J. (2015). Moralization and the 2012 us presidential election campaign. *Journal of Social and Political Psychology*, *3*(2), 211–237.

Cabana, Á., Zugarramurdi, C., Valle-Lisboa, J. C., & De Deyne, S. (2023). The "small world of words" free association norms for rioplatense spanish. *Behavior Research Methods*, 1–18.

Cardellino, C. (2016, August). *Spanish Billion Words Corpus and Embeddings.* Retrieved from https://crscardellino.github.io/SBWCE/

Clifford, S. (2019). How emotional frames moralize and polarize political attitudes. *Political Psychology*, *40*(1), 75–91.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 240–247.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... Stoyanov, V. (2020, July). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.747 doi: 10.18653/v1/2020.acl-main.747

De Deyne, S., Cabana, Á., Li, B., Cai, Q., & McKague, M. (2020). A cross-linguistic study into the contribution of affective connotation in the lexico-semantic representation of concrete and abstract concepts. In *Proceedings for the 42nd Annual Meeting of the Cognitive Science Society.*

De Deyne, S., Navarro, D. J., Collell, G., & Perfors, A. (2021). Visual and affective multimodal models of word meaning in language and mind. *Cognitive Science*, *45*(1), e12922.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "small world of words" english word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*, 987–1006.

De Deyne, S., Navarro, D. J., & Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods*, *45*, 480–498.

Deese, J. (1965). *The Structure of Associations in Language and Thought.* Baltimore: The Johns Hopkins Press.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*.

Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, *98*(45-60), 16.

Emelin, D., Le Bras, R., Hwang, J. D., Forbes, M., & Choi, Y. (2021, November). Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 698–718). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.54 doi: 10.18653/v1/2021.emnlp-main.54

Fehr, B., & Russell, J. A. (1984). Concept of emotion viewed from a prototype perspective. *Journal of Experimental Psychology: General*, *113*(3), 464.

Feinberg, M., Kovacheff, C., Teper, R., & Inbar, Y. (2019). Understanding the process of moralization: How eating meat becomes a moral issue. *Journal of Personality and Social Psychology*, *117*(1), 50.

Fischer, R., Luczak-Roesch, M., & Karl, J. (2023, 04). *What does ChatGPT return about human values? exploring value bias in ChatGPT using a descriptive value theory.*

Forbes, M., Hwang, J. D., Shwartz, V., Sap, M., & Choi, Y. (2020, November). Social chemistry 101: Learning to reason about social and moral norms. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 653–670). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-main.48 doi: 10.18653/v1/2020.emnlp-main.48

Frimer, J., Haidt, J., Graham, J., Dehghani, M., & Boghrati, R. (2017). Moral foundations dictionaries for linguistic analyses, 2.0. *Unpublished Manuscript*.

Garten, J., Boghrati, R., Hoover, J., Johnson, K. M., & Dehghani, M. (2016). Morality between the lines: Detecting moral sentiment in text. In *Proceedings of IJCAI 2016 workshop on Computational Modeling of Attitudes.*

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In *Advances in Experimental Social Psychology* (Vol. 47, pp. 55–130). Elsevier.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029.

Haemmerl, K., Deiseroth, B., Schramowski, P., Libovický, J., Rothkopf, C., Fraser, A., & Kersting, K. (2023, July). Speaking multiple languages affects the moral bias of language models. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 2137–2156). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.findings-acl.134 doi: 10.18653/v1/2023.findings-acl.134

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., … Puranen, B. (2021). World Values Survey: Round Seven – Country-Pooled Datafile. *Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat. Data File Version*, *2*(0).

Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016, November). Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 595–605). Austin, Texas: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D16-1057 doi: 10.18653/v1/D16-1057

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2020). Aligning AI with shared human values. In *9th International Conference on Learning Representations (ICLR).*

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2021). Aligning AI With Shared Human Values. In *International conference on learning representations.* Retrieved from https://openreview.net/forum?id=dNy_RKzJacY

Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A. M., Lin, Y., … Mendlen, M. (2020). Moral Foundations Twitter Corpus: A collection of 35k Tweets Annotated for Moral Sentiment. *Social Psychological and Personality Science*, *11*(8), 1057–1071.

Hopp, F. R., Fisher, J. T., Cornell, D., Huskey, R., & Weber, R. (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.

Inbar, Y., Phelps, J., & Rozin, P. (2020). Recency negativity: Newer food crops are evaluated less favorably. *Appetite*, *154*, 104754.

Inglehart, R., Haerpfer, C., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., … Puranen, B. (2014a). World Values Survey: all round. *Madrid: JD systems institute*, *12*.

Inglehart, R., Haerpfer, C., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., … Puranen, B. (2014b). World Values Survey: Round six-country-pooled. *Madrid: JD systems institute*, *12*.

Jentzsch, S., Schramowski, P., Rothkopf, C., & Kersting, K. (2019). The Moral Choice Machine: Semantics derived automatically from language corpora contain human-like moral choices. In *Proceedings of the 2nd AAAI/ACM Conference on AI, Ethics, and Society. Palo Alto (California): Association for the Advancement of Artificial Intelligence.*

Jia, F., & Krettenauer, T. (2017). Recognizing moral identity as a cultural construct. *Frontiers in Psychology*, *8*, 412.

Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. L., Forbes, M., Borchardt, J., … Choi, Y. (2021). Delphi: Towards Machine Ethics and Norms. *ArXiv*, *abs/2110.07574*.

Jin, Z., Levine, S., Gonzalez Adauto, F., Kamal, O., Sap, M., Sachan, M., … Schölkopf, B. (2022). When to make exceptions: Exploring language models as accounts of human moral judgment. *Advances in neural information processing systems*, *35*, 28458–28473.

Johnson, K., & Goldwasser, D. (2018, July). 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)* (pp. 720–730). Melbourne, Australia: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P18-1067 doi: 10.18653/v1/P18-1067

Johnson-Laird, P. N., & Oatley, K. (1989). The language of emotions: An analysis of a semantic field. *Cognition and Emotion*, *3*(2), 81–123.

Kleiman-Weiner, M., Saxe, R., & Tenenbaum, J. B. (2017). Learning a commonsense moral theory. *Cognition*, *167*, 107–123.

Laurino, J., De Deyne, S., Cabana, Á., & Kaczer, L. (2023). The pandemic in words: Tracking fast semantic changes via a large-scale word association task. *Open Mind*, 1–19.

Liscio, E., Dondera, A., Geadau, A., Jonker, C., & Murukannaiah, P. (2022, July). Cross-domain classification of moral values. In *Findings of the association for computational linguistics: Naacl 2022* (pp. 2727–2745). Seattle, United States: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022

.findings-naacl.209 doi: 10.18653/v1/2022.findings-naacl.209

Liu, C., Cohn, T., De Deyne, S., & Frermann, L. (2022). Wax: A new dataset for word association explanations. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing* (pp. 106–120).

Liu, R., Zhang, G., Feng, X., & Vosoughi, S. (2022). Aligning Generative Language Models with Human Values. In *Findings of the association for computational linguistics: Naacl 2022* (pp. 241–252).

Lourie, N., Le Bras, R., & Choi, Y. (2021). Scruples: A corpus of community ethical judgments on 32, 000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, pp. 13470–13479).

Mendelsohn, J., Tsvetkov, Y., & Jurafsky, D. (2020). A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence*, *3*, 55.

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space. *In International Conference on Learning Representations (ICLR), 2013a*.

Mooijman, M., Hoover, J., Lin, Y., Ji, H., & Dehghani, M. (2018). Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour*, *2*(6), 389–396.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., … Ray, A. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Ramezani, A., & Xu, Y. (2023, July). Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 428–446). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.acl-long.26 doi: 10.18653/v1/2023.acl-long.26

Reimers, N., & Gurevych, I. (2019, 11). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics. Retrieved from https://arxiv.org/abs/1908.10084

Reimers, N., & Gurevych, I. (2020, November). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4512–4525). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-main.365 doi: 10.18653/v1/2020.emnlp-main.365

Roy, S., Pacheco, M. L., & Goldwasser, D. (2021). Identifying morality frames in political tweets using relational learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.*

Rozin, P. (1999). The process of moralization. *Psychological Science*, *10*(3), 218–221.

Rozin, P., Markwith, M., & Stoess, C. (1997). Moralization and becoming a vegetarian: The transformation of preferences into values and the recruitment of disgust. *Psychological Science*, *8*(2), 67–73.

Rozin, P., & Singh, L. (1999). The moralization of cigarette smoking in the united states. *Journal of Consumer Psychology*, *8*(3), 321–337.

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., & Choi, Y. (2020, July). Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5477–5490). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.486 doi: 10.18653/v1/2020.acl-main.486

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., & Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, *4*(3), 258–268.

Schramowski, P., Turan, C., Jentzsch, S., Rothkopf, C., & Kersting, K. (2019). Bert has a moral compass: Improvements of ethical and moral values of machines. *arXiv preprint arXiv:1912.05238*.

Shaver, P., Schwartz, J., Kirson, D., & O'connor, C. (1987). Emotion knowledge: further exploration of a prototype approach. *Journal of Personality and Social Psychology*, *52*(6), 1061.

Simmons, G. (2023, July). Moral mimicry: Large language models produce moral rationalizations tailored to political identity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)* (pp. 282–297). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.acl-srw.40 doi: 10.18653/v1/2023.acl-srw.40

Skitka, L. J., Wisneski, D. C., & Brandt, M. J. (2018). Attitude moralization: Probably not intuitive or rooted in perceptions of harm. *Current Directions in Psychological Science*, *27*(1), 9–13.

Trager, J., Ziabari, A. S., Davani, A. M., Golazazian, P., Karimi-Malekabadi, F., Omrani, A., … Reyes, M. (2022). The Moral Foundations Reddit Corpus. *arXiv preprint arXiv:2208.05545*.

Tulkens, S., Emmery, C., & Daelemans, W. (2016, May). Evaluating unsupervised Dutch word embeddings as a linguistic resource. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 4130–4136). Portorož, Slovenia: European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L16-1652

Van Rensbergen, B., Storms, G., & De Deyne, S. (2015). Examining assortativity in the mental lexicon: Evidence from word associations. *Psychonomic Bulletin & Review*, *22*, 1717–1724.

Xie, J. Y., Ferreira Pinto Junior, R., Hirst, G., & Xu, Y. (2019, November). Text-based inference of moral sentiment change. 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)* (pp. 4654–4663). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D19-1472 doi: 10.18653/v1/D19-1472

Xie, J. Y., Hirst, G., & Xu, Y. (2020). Contextualized moral inference. *arXiv preprint arXiv:2008.10762*.

Xu, A., Stellar, J. E., & Xu, Y. (2021). Evolution of emotion semantics. *Cognition*, *217*, 104875.