This is the author's copy of the paper to appear in Nature Communications.

Historical reconstruction of human moralization with word association and text corpora

Aida Ramezani[1*], Jennifer E. Stellar[2], Matthew Feinberg[3], Yang Xu[1,4,5]

[1]Department of Computer Science, University of Toronto.
[2]Department of Psychological & Brain Sciences, University of Toronto.
[3]Rotman School of Management, University of Toronto.
[4]Cognitive Science Program, University of Toronto.
[5]Vector Institute for Artificial Intelligence.

*Corresponding author(s). E-mail(s): armzn@cs.toronto.edu;

Abstract

Moralization, the process by which concepts and practices gain moral attributes, plays a pivotal role in shaping individual and social behaviour. However, research on how moralization unfolds over time remains limited. We present HistMoral, an open-access computational framework based on human word association, historical text corpora, and graph neural networks that enables scalable, retrospective analysis of moral trajectories of many different concepts. We apply our framework to analyze the moral time courses of over 20,000 concepts within the Corpus of Historical American English over the past 150 years, as well as within the New York Times annotated corpus from 1987 to 2007. Our findings provide robust evidence of moralization across diverse categories, from diseases to world leaders, and identify moralization around economic-political shifts of recent decades.

Moralization is the process by which concepts, practices, and entities gain moral attributes over time [1, 2]. How this process unfolds is one of the oldest mysteries of human nature [3, 4]. Many studies have examined moralization within controlled lab settings at a single timepoint or over short time periods, typically focusing on individual topics such as *vegetarianism*, *abortion*, and *genetic engineering* [1, 5–9]. These empirical investigations have shown that moralization arises from psychologically grounded mechanisms, such as moral emotions [10, 11], and reasoning about harms and benefits [1, 6, 12]. While these studies offer valuable insights into the psychological mechanisms driving moral change, they are inherently constrained by limited sample sizes, short timeframes, and a narrow range of moralized concepts. For instance, moral perceptions of *cigarette smoking* shifted gradually over decades from morally acceptable to morally inappropriate in many contexts [10]. Additionally, many issues were moralized decades or even centuries ago, such as opposition to *slavery* or *dueling* [13], meaning that studying them is beyond the scope of current methods. These limiting factors raise natural questions regarding the generalizability of findings, leaving open important issues about how moralization unfolds over longer timescales and across a broader set of subjects. To address this important question, we introduce an open-access framework that automatically reconstructs the historical moral trajectories of lexicalized concepts (see Figure 1a for illustrations), enabling efficient and systematic discoveries about when and why moralization occurs.

To study human morality at a broader scale, recent research in artificial intelligence (AI), particularly natural language processing, has utilized big data analysis and language models to automatically infer human moral values from textual data. This emerging field of automated moral inference in computer science employs computational tools for the automated analysis of morality. It includes investigating the rise and fall of moral concepts over time through large-scale word frequency analysis [14] or word co-occurrence statistics [15], using word-embedding and neural networks to predict moral sentiments from text [16, 17], and aligning artificial intelligence systems with human values and moral decisions [18].

Most relevant to our study, previous work has shown that language models pretrained on large text corpora embed information that reflects people's moral perception toward concepts such as 'becoming a good parent' or 'harming people' [19]. Other studies drawing on Moral Foundations Theory [20] have used word co-occurrence statistics to study changes in moral relevance of words [15], or used diachronic word embeddings derived from historical texts [21]. However, a main limitation of these computational approaches is that they are typically detached from psychological theories of moralization, and lack the semantic richness necessary for robust, reliable predictions across long term horizons. These issues raise critical questions regarding how applicable these models are in service of exploring human moralization.

To overcome the challenges above, we introduce a computational framework that automatically reconstructs the moral trajectories of different concepts within the English lexicon over time. Our model builds on Moral Foundations Theory, which holds that moral judgments often arise immediately from intuitive, evolutionarily shaped foundations [20, 22], while also incorporating insights from psychological research on moralization, which examines how and why issues acquire or lose moral significance over time. Specifically, we draw on large-scale psychological datasets of word association—semantic networks that relate concepts with differing strengths based on their associations in people's minds—which forms the basis of moral network graphs that indicates a concept's moral properties [23, 24]. Importantly, recent work has shown this graph-based model of moral association based on psychological accounts of moralization can predict human moral perception with precision [25]. For instance, the term *smoking* is strongly associated with words like *unhealthy, addiction*, and *bad* in English word association data [26], indicating that when people see the word *smoking*, they tend to immediately think about its harmful health effects (see Figure 1b for an illustration).

Although this moral word association approach provides an innovative way of uncovering the moral properties of a concept, one main challenge is how to extend this approach in a historical setting where word association data is unavailable. This limits our ability to understand how concepts may have shifted in their moral relevance over time, to identify what psychological and historical factors contribute to the moralizaton (or unmoralization) of a concept, or to build models that can predict future moralization of concepts. Addressing these limitations requires moving beyond models that capture only a single point in time, toward an approach that characterizes moralization across historical periods.

Here, we present HistMoral: a computational framework that addresses these challenges and reconstructs historical moral trajectories across time. Our framework extends the word association approach to historical moral inference. We do so by "approximating" word association networks using available large-scale linguistic data across historical periods. This approximation process works by assigning new associations to words based on how often words co-occur in a historical corpus. However, since co-occurrence statistics are inherently noisy predictors of conceptual relations [27], we enhance this approximation by extracting contextual semantic representations (i.e., embedding vectors) using contextual language models such as BERT [28]. These embedding vectors capture word meanings based on usage across textual data.

To derive morally relevant information from these contextual representations and co-occurrence patterns, we train a graph neural network (GNN) [29] to assign moral association scores to words—scores that reflect human moral perception. Figure 1b illustrates HistMoral framework and its psychological grounding using a representative example. Historical text corpora reveal how frequently words appear in proximity to one another, while word association networks, derived from large-scale human experiments, capture conceptual relations between concepts. Strong associations with moral terms (e.g., associating *smoking* with *bad*) captures perceived moral relevance. Our reconstruction algorithm leverages textual corpora to approximate word associations and infer moral association scores for a wide range of concepts. We describe the details of our methodology in *Methods* Section.

Our approach offers interpretable and historical estimates of moral association scores that are theoretically aligned with human moral cognition. Table 1 presents a comparison of our framework with alternative computational methods. While lexical approaches, such as word co-occurrence frequency statistics, can capture temporal dynamics in diachronic corpora, they lack the semantic richness required for reliable moral inference. AI-driven methods, including early word embedding models [21] and more recent large language models [19], encode such contextual semantic information but primarily rely on pattern recognition from large-scale data, without a theoretical grounding to justify and explain their moral estimates. In contrast, our framework, which is grounded in psychological moral theories and word association datasets, leverages contextual semantics from text corpora to enable diachronic inference. Our framework can also be applied across diverse diachronic datasets, which is infeasible for pre-trained language models. As shown in the *Supplementary Table 1*, our method also achieves more accurate moral reconstructions than alternative computational approaches.

The output of our framework is the reconstructed moral trajectories of different concepts. Specifically, at each time point, we estimate two types of moral association score defined according to existing work [21, 25] as the predicted outputs from the model: moral relevance and moral polarity. Moral relevance approximates the proportion of moral terms (e.g., *bad, addiction*) that human participants associate with a given query concept (e.g., *smoking*). This score reflects the relative frequency at which a query concept is associated with morality in people's minds. A higher moral relevance score indicates a stronger tendency for people to associate the query concept with moral terms. For example, the concept *killing* has a high moral relevance score because most people associate it with moral terms like *wrong, crime*, and *evil*. Moral polarity is defined as the ratio (or relative frequency) of a query concept's association with morally positive terms (e.g., *caring*) compared to negative terms (e.g., *harming*). A higher score indicates a stronger tendency for people to associate the concept with morally positive terms compared to negative ones. For example, the concept *killing* has a low moral polarity score, because most of its word associates are negative words (e.g., *crime*, *evil*, and *wrong*). The concept *helping* on the other hand has a high moral polarity score since most of its word associates are positive words (e.g., *good*, *kind*, and *nice)*. In *Methods*, we provide a detailed description of how both moral relevance and moral polarity are calculated and the set of moral terms used for our analysis. Figure 1a illustrates how our framework identifies dynamically changing moral relevance scores throughout the past 150 years for three different concepts.

In summary, this work presents a computational framework that reconstructs historical moral trajectories across time. By combining the word association approach with a machine learning framework trained on historical corpora, we generate moral time courses that capture how different concepts have gained or lost moral relevance throughout history, and whether that relevance reflected primarily positive regard or negative judgment toward the concept. We validate our framework by demonstrating its ability to accurately predict known cases of moralization, and we apply it to systematically analyze moral change across thousands of concepts over the past 150 years.

Results

We focus on two large-scale longitudinal text corpora: the Corpus of Historical American English (COHA), a historical corpus containing approximately 400 million words and covering from the 1850s to the 2000s split by decades [30], and the New York

Times Corpus (NYT), which includes 1.8 million news articles published between 1987 and 2007 [31]. We include the details of these corpora in *Methods*. We chose to analyze COHA and NYT for their comprehensive, large-scale diachronic textual data that span long periods of time and are available to researchers. In practice, however, our framework is highly flexible and can be applied to any large-scale longitudinal text corpora.

Below, we first evaluate our framework trained using COHA and NYT against empirical moral association scores from contemporary human data. We then demonstrate how our framework can be applied to the analysis of moral time courses of different concepts and conceptual categories. Furthermore, we show that our framework discovers systematic patterns in historical moralization and identifies moral change associated with economic and political shifts.

Framework validation across empirical and historical trends

To validate our framework, we compare our model's results with empirical human word association data. Then, we move on to examining whether our framework can effectively predict high moral relevance and polarity scores for concepts well-known to have high levels of each in historical contexts.

We first evaluate the model-predicted scores against the empirical scores computed from human word association data collected between 2011 and 2018 [26] based on procedures in prior work [25]. Figures 2a and 2b show that, for both moral relevance and polarity, the model prediction based on reconstruction from the COHA corpus correlates strongly with the empirical measurements of moral relevance ($r^2 = 0.445$, Pearson's $r = 0.669$, $P < 0.0001$, 95% CI = $[0.632, 0.703]$, $n = 937$) and moral polarity ($r^2 = 0.447$, Pearson's $r = 0.675$, $P < 0.0001$, 95% CI = $[0.637, 0.710]$, $n = 843$). Similar results observed with reconstruction from the NYT corpus (moral relevance: $r^2 = 0.414$, Pearson's $r = 0.645$, $P < 0.0001$, 95% CI = $[0.604, 0.682]$, $n = 883$; moral polarity: $r^2 = 0.439$, Pearson's $r = 0.665$, $P < 0.0001$, 95% CI = $[0.624, 0.702]$, $n = 789$). Of note, our model predicted concepts including *sinner, unfair*, and *divorce*—which are strongly associated with morality—as more morally relevant than morally neutral concepts such as *exist*. Our model also predicted concepts like *embrace*, *candy*, and *violence* to fall within a gradient of moral polarity, ranging from positive to negative. In *Supplementary Table 1*, we provide additional evidence for the effectiveness of our model reconstruction in comparison to alternative model architectures and baseline models.

Along with reconstructing moral scores, we used sentiment analysis to extract word valence (positiveness and negativeness) at each historical time point to account for the potential confound of valence in our experiments. When examining all concepts with scores averaged over time we found that valence was weakly negatively related to moral relevance (Pearson's $r = -0.140$, $P < 0.0001$, 95% CI = $[-0.153, -0.126]$, $n = 20,788$). Valence was moderately positively related to moral polarity (Pearson's $r = 0.445$, $P < 0.0001$, 95% CI = $[0.434, 0.456]$, $n = 20,788$). However, the association between valence and moral polarity is not so high as to consider the two constructs redundant.

Next, to further test the effectiveness of our model, we assess whether our model reconstructs historical moral association scores by identifying certain concepts to be more likely to be moralized than others over the course of history. In particular, we focus on analyzing the concepts pertaining to "disease (health)" and "political figure (politics)," which have been consistently reported in the literature as being of high moral relevance [8, 32–40]. The purpose of focusing on these topics is to offer confirmatory and qualitative validation of our computational framework. Specifically, we would expect our model to give such well-established moralized terms a very high moral relevance score. To this end, we identify words from the conceptual categories "disease" and "political figure," and compare their degree of moral relevance to randomly selected word sets of equivalent size from the COHA corpus. We use the normalized moral scores with z-score standardization for our analyses (see *Methods* for details).

Figure 2c compares the model-estimated moral relevance of disease-related concepts (e.g., *cholera* in the 1990s) with a random set of concepts matched in set size. The results indicate that disease terms exhibit higher moral relevance than the control set of random concepts (Wilcoxon signed-rank test: $W = 8927.0$, $P < 0.0001$, one-tailed; Cohen's $d = 2.24$, sample size = 135). A non-parametric test was used as the paired differences violated the normality assumption. To validate these findings, we repeated the analysis across 1000 randomly constructed control sets. The distribution of mean moral relevance scores from these random samples was approximately normal (Shapiro-Wilk: $W = 0.998$, $P = 0.316$, Mean = -0.020, $SD = 0.081$, $n = 1000$). A one-tailed one-sample t-test confirmed that the disease terms showed significantly higher moral relevance than the average of random control sets ($t(999) = -684.400$, $P < 0.0001$).

These results help validate our model by showing that it provides high moral relevance scores to concepts known to be highly moralized based on previous research on the moralization of diseases [32]. Additionally, for moral polarity, disease-related concepts were perceived to be more morally negative compared to the randomly constructed sets of concepts (one-sample t-test: $t(999) = 929.996$, $P < 0.0001$, one-tailed). Although the Shapiro-Wilk test indicated deviation from normality for these random sets ($W = 0.995$, $P = 0.003$), the effect size and sample sizes are both large (Cohen' $d = 29.41$, sample size = 1000). This finding shows that the model-reconstructed moral polarity reflects a negative moral perception toward diseases. We repeated this analysis by mapping each disease term to one of the top 5 (non-disease) terms with the most similar valence scores. Our results remained robust for both moral relevance (one-sample t-test: $t(999) = -711.96$, $P < 0.0001$, one-tailed), and moral polarity (one-sample t-test: $t(999) = 428.60$, $P < 0.0001$, one-tailed), indicating that our findings cannot be explained by valence alone.

We observe similar trends for political figures (i.e., world leaders such as *Winston Churchill*). Figure 2d compares the moral relevance estimates for various political figures with a random sample of first names of the same size, showing that political figures generally have higher moral relevance (Wilcoxon signed-rank test: $W = 1681.0$, $P < 0.0001$, one-tailed; Cohen's $d = 0.75$, sample size = 65). Repeating this analysis across 1000 randomly constructed sets of names further confirmed that political figures have stronger moral relevance (one-sample t-test: $t(999) = -223.940$, $P < 0.0001$, one-tailed).

To further evaluate the effectiveness of our framework in capturing historical shifts in moral relevance, we analyzed how major world events influence the moralization of political entities. Using the Historical Conflict Event Dataset [41], we find that entities involved in conflicts exhibit higher moral relevance and more negative moral polarity during wartime compared to peacetime (moral relevance: Wilcoxon signed-rank test $W = 13876.0$, $P < 0.0001$, one-tailed; Cohen's $d = 0.15$; moral polarity: Wilcoxon signed-rank test $W = 2712.0$, $P < 0.0001$, one-tailed; Cohen's $d = 0.17$, sample size = 183). Figure 2e illustrates these changes, with the y-axis indicating the difference in moral relevance between wartime and peacetime. For instance, the term Axis—referring to the Axis Powers in World War II—shows a substantial spike in moral relevance during the 1940s, as shown in Figure 2f. A second peak appears in the 2000s, possibly due to the usage of the phrase “Axis of evil” by then-president George W. Bush.

We also observe that the losing side in international conflicts is associated with both higher moral relevance and more negative moral polarity than the winning side (moral relevance: Wilcoxon signed-rank test $W = 9508.5$, $P = 0.038$, one-tailed; Cohen's $d = 0.20$; Wilcoxon signed-rank test $W = 13790.5$, $P = 0.001$, one-tailed; Cohen's $d = 0.27$, sample size = 210). These results indicate that our model assessed the moral relevance and polarity of these concepts in line with what would be expected given our understanding of the moralization of political entities during wartime. To account for the potential confounding effect of valence on moralization during conflicts, we trained a binary logistic regression model to predict conflict outcomes (winner vs. loser) using differences in sentiment, moral polarity, and moral relevance scores between the two sides as predictors. The model achieved an accuracy of 0.59 (random baseline

= 0.5, sample size = 207). Notably, only moral polarity emerged as a statistically significant predictor ($\beta = 0.34$, $P = 0.036$, 95% CI = $[0.022, 0.653]$), while sentiment and moral relevance were not statistically significant predictors ($P = 0.41$ and $P = 0.82$, respectively). These results suggest that conflict outcomes are more likely to be framed in terms of moral righteousness, even when controlling for sentiment. Further details of these analyses are provided in the *Methods* Section.

Since our framework draws on word co-occurrence patterns to reconstruct moral association, we further investigate whether our reconstructed moral estimates are artifacts of contextual co-occurrence with moral words (indirect or surface-level association), or directly capture semantic-level moral association. In other words, we examine whether our model is simply capturing that certain terms often appear near moral words, or if our model is actually capturing conceptual relationships with moral values and concepts. To address this, we performed an analysis that distinguishes between direct and indirect types of moralization. Specifically, we consider direct moralization when a term becomes moralized conceptually as a result of conceptual association with moralized ideas. We consider indirect moralization when a term's moral relevance increases due to frequent co-occurrence with moral terms in our textual data, but does not demonstrate a deeper association with moralized ideas. For example, the word *cigarettes* may co-occur with words like *law*, and *regulation*, but not yet be conceptually moralized in the society. We find that our moral relevance scores (proxy of direct moralization) and indirect moralization scores are moderately positively correlated across time (Pearson's $r = 0.53$, $P < 0.0001$, 95% CI = $[0.526, 0.532]$, $n = 168,767$). This outcome aligns with the intuition that moralized terms tend to appear in moral contexts. However, focusing specifically on diseases and named political figures, we find no statistically significant correlation between the two (diseases: Pearson's $r = 0.083$, $P = 0.337$, 95% CI = $[-0.087, 0.249]$, $n = 135$, political figures: Pearson's $r = 0.036$, $P = 0.794$, 95% CI = $[-0.234, 0.301]$, $n = 54$). This result suggests that our moral relevance estimates and indirect moralization scores do not always move in the same direction, and that our model is capturing direct moralization of the target concept.

This set of results show that our model is a valid predictor of moral associations. As we showed in various ways in these experiments, our model-reconstructed moral relevance and moral polarity not only predict empirical moral association scores, but are also in line with what is expected to be a subject of moral judgment across time.

Systematic moralization of conceptual categories

Our findings thus far show high moral association scores for categories such as "disease" terms. This suggests there may be shared trends in moral trajectories of concepts within the same category. To explore this pattern more thoroughly in the lexicon, we examine the moral relevance scores of a comprehensive set of related concepts or conceptual categories. This analysis is aimed at characterizing the similarities and variations across a diverse range of concepts and evaluating the utility of our framework for making discoveries about systematic patterns of moralization.

To define conceptual categories, we used an existing comprehensive database that contains 117 distinct groups of concepts, where groups/categories are selected from previous categorization literature [42–48], and normed concepts within each category are empirically collected from human participants [49].

We first assessed the effectiveness of the model in reconstructing human moral perception at the category level. We did so by using our model to compute the averaged category-wise moral relevance scores and validated them against human data [26]. A Spearman's correlation test indicates that our model reconstruction correlates strongly with the moral relevance scores obtained from procedures used in existing work [25, 26] with $\rho = 0.744$ ($P < 0.0001$, 95% CI = $[0.65, 0.82]$). For comparison, we also used a state-of-the-art large language model, GPT-4o, to reconstruct the same empirical data and found a substantially weaker, but statistically significant correlation with $\rho = 0.565$ ($P < 0.0001$, 95% CI = $[0.43, 0.68]$). See Figure 3a for a summary of these results and *Supplementary Notes Section 1.2* for details of this analysis.

Figure 3a visualizes the moral relevance scores estimated by the model for all of these categories with respect to their rates of change in moral relevance between the 1850s and 2000s. The rates are calculated from regressing the moral relevance scores against time, and therefore show the average growth in moral relevance in one unit of time (i.e., a decade). As shown, the overall moralization patterns vary across conceptual categories. For example, the conceptual category "disease" exhibits both high moral relevance and a high rate of change, suggesting it has become increasingly morally relevant over time. In contrast, the category "supernatural being," which includes concepts like *god, devil, ghost* and *alien*, is morally relevant but shows a decline in moral relevance over time. Similar to our analysis, previous work has shown that the frequency of English moral terms tend to fluctuate with respect to major world events in the 20th century [14]. Therefore, we further investigate whether our estimated moralization rates might simply reflect increases/decreases in frequency of certain moral terms associated with these categories. Using the frequencies (drawn from COHA) of concepts in the conceptual category database across time, we find a weak positive correlation between word log frequency and moral relevance score (Pearson's $r = 0.084$, $P < 0.0001$, 95% CI = $[0.068, 0.100]$, $n = 14,659$). Our examples in *Supplementary Figure 1* further show that frequency levels do not reliably predict moral association scores of conceptual categories such as "disease" and "supernatural being." These results support the interpretation that observed declines (or rises) in moral relevance scores reflect changes in the moral framing and perception of concepts over time, rather than just shifts in word usage frequency.

To better understand whether there is systematic moralization in the conceptual categories, we performed a predictive analysis using moral relevance and moral polarity scores of a concept to predict category membership based on its proximity to other concepts in the moral space. Our results show that category membership can be predicted with an accuracy of 46.7% based solely on moral association profiles, outperforming the majority vote baseline, which achieves 9.1% accuracy by always predicting the most common category in the data (see *Supplementary Figure 2*). Figure 3b further shows that similar concepts exhibit synergistic movement in moral trajectories. The heatmap, derived from pairwise correlations in the moral trajectories of concepts across different conceptual categories, suggests that certain conceptual categories share correlated moral trajectories over time. For example, concepts in the "family relationship" category have similar moralization trajectories to each other, but are also similar to those of concepts in the "social relationship category." A comparison of average pairwise correlations within these conceptual categories versus randomly generated categories of equivalent sample size reveals that concepts within the same category exhibit more similar moral trajectories than concepts across different categories (one-sample t-test: $t(999) = -203.94$, $P < 0.0001$, Cohen's $d = 4.45$, two-tailed; see *Methods* for additional details).

Moral change under economic and political shifts

We next demonstrate the utility of our framework in identifying moral changes over recent decades, focusing on the NYT corpus from 1987 to 2007. Specifically, we study the relationships between economic and political shifts with moral trajectories, since both factors are known to correlate with people's moral views [50–55].

To explore the relationship between moral change and the economy, we focus on consumer price fluctuations of retail products in the United States. Our experiment is motivated by past research findings showing that as products become morally negative, the amount people are willing to pay for them decreases [56], meanwhile as products become more morally positive people are willing to pay more for them [57, 58].

Using our framework to study such effects, Figure 4a shows that annual changes in retail prices has a clear negative relationship with changes in moral polarity (Pearson's $r = -0.144$, $P = 0.0001$, 95% CI = $[-0.216, -0.071]$, $n = 704$). Figure 4b further illustrates this relationship for common product categories including "Gasoline," "Bread," and "Beef."

Training an ordinary least squares (OLS) model with product ID, current year, price of the product in the previous year, and sentiment valence scores as control variables shows that year-by-year changes in moral polarity predict changes in product pricing ($\beta = -0.147$, $P = 0.016$, 95% CI = $[-0.266, -0.027]$, $t(335) = -2.419$). The negative coefficient indicates that changes in moral polarity of a product are inversely related to changes in its retail price. Specifically, when a product becomes associated with more morally positive concepts relative to the previous year (i.e., positive change in moral polarity), its price tends to decrease. Conversely, when a product becomes associated with more morally negative concepts relative to the previous year (i.e., negative change in moral polarity), its price tends to increase. See *Methods* and *Supplementary Equation 1* for more details on the dataset and analysis.

Although these findings offer insights into moral-economic dynamics, they should be interpreted with caution. While prior work has shown that people are less willing to pay more for morally negative items [56–58], we find the opposite pattern, that moral negativity predicts higher retail prices. However, our analysis does not establish causal directionality, in part due to the limited long-term price data across the products. The observed relationship could reflect different possible mechanisms, such as price increase leading people to view products more negatively, vice taxes inflating the price of morally negative products, or common external factors affecting both morality and retail prices. Moreover, our analysis focuses on actual retail prices, which may not fully capture consumers' underlying willingness to pay. To probe this finding further, we conducted a qualitative investigation of the NYT corpus and found instances where price fluctuations were explicitly discussed in moral terms. For instance, price increases were linked to environmental concerns and political rivalry (e.g., energy prices in the 2000s) and public health issues (e.g., red meat contamination in 2005), reflecting the potential influence of common external factors on both price and moral perception.

Next, we explore the relationship between moral change and political discourse in the United States. Building on our analysis of historical world events, we hypothesize that impactful modern political events—such as Congressional debates and presidential elections—should influence changes in public moral perception (and vice versa). To test this hypothesis, we analyzed the changes in moral association over time in two key political contexts: (1) Congressional speeches and (2) presidential election cycles in the United States.

We draw on the United States Congressional speech data to identify common political topics and phrases [59]. Figure 4c illustrates 22 politically relevant topics recorded in this dataset. For example, *economical challenge* is a phrase from the topic "economy." First, we hypothesize that concepts gaining moral relevance would also gain prominence in political discourse. To test this hypothesis, we performed an OLS analysis predicting annual changes in the frequency of words in Congressional speeches based on shifts in their moral association scores, while controlling for sentiment scores. Our findings reveal a statistically significant positive relationship between annual changes in words' moral relevance scores and their frequency in Congressional speeches ($\beta = 0.499$, $P < 0.0001$, 95% CI = $[0.378, 0.621]$, $t(3,257) = 8.079$). We also find a statistically significant negative relationship between annual changes in moral polarity scores and word frequencies based on Congressional speeches ($\beta = -0.141$, $P = 0.022$, 95% CI = $[-0.261, -0.021]$, $t(3,257) = -2.300$). These results show that words experiencing increases in moral relevance (i.e., becoming more moralized) and declines in positive moral sentiments are also more frequently mentioned in Congressional speeches and debates, which suggests an interplay between moralization and political discourse.

We also examine patterns of moralization in relation to U.S. presidential election cycles. Figure 4c shows changes in moral relevance across 22 politically salient topics, comparing years when Democratic candidates won the presidency (1992, 1996) to those when Republican candidates prevailed (1988, 2000, 2004). We observe that topics such as the "economy," "federalism," "environment," and "education" exhibit higher moral relevance in NYT in years corresponding to Democratic victories. Since the presidential elections are held in November, the estimated moral relevance scores mostly measure moral association in months leading to the election, and therefore likely predict the upcoming election results. To disentangle these effects from broader temporal trends, we conducted a controlled mixed-effects OLS analysis to estimate the impact of election outcomes on moral relevance scores in the upcoming year (see *Supplementary Equation 3* and *Supplementary Table 11* for details). This analysis reveals two key findings: (1) moral relevance scores are higher in the year of the election, and become lower after the election, and (2) election outcomes have differential effects across political topics. For instance, environmental issues have lower moral relevance scores in months leading to Republican victories, but their moral relevance scores increase after the election, potentially reflecting intensified moral framing by the losing Democratic party.

Figure 4d illustrates phrases with the largest differences in moral relevance under Democratic wins (top) versus Republican wins (bottom) in the year of the elections. In Democratic victory years, concepts like *drug* show an increase in moral relevance compared to Republican victory years, whereas concepts like *HIV* are more morally relevant in Republican victory years (see *Methods* for details). We believe these results may reflect broader political dynamics and unobserved temporal effects. For instance, past work finds that among the congressional representatives in the United States, the use of moral language increased after one's political party lost an election [60]. Therefore it is likely that the opposition party that most fears it will lose the upcoming election (and does so), is more likely to push a moralized agenda. Note that these patterns reflect the moral framing of these concepts as represented in The New York Times—a newspaper generally associated liberal-leaning stances—which might be different from the party leaders themselves.

In additional analyses described in the *Supplementary Notes Section 1.6*, we examined the temporal relationship between moral change and social concerns about relevant moral issues. Our findings show a correspondence between increases in moral association and growing social concerns as recorded in empirical surveys. Furthermore, a bottom-up analysis described in *Supplementary Notes Section 1.7* based on both NYT and COHA corpora shows that concepts with the sharpest increases in moral relevance are mostly relevant to major socio-political events and movements (see *Supplementary Tables 7* and *8*).

Discussion

Smoking cigarettes, gambling, and nuclear weapons are wildly different concepts, but they have one thing in common: moralizing these concepts has fundamentally changed how we think about them, make judgments about other people, and evaluate policy initiatives (cf.,[61]). We presented a general framework that supports historical reconstruction of moralization trajectories for more than 20,000 concepts. Our approach led to system-level discoveries about moralization, including the detection of moralization in cohorts of concepts such as diseases and world leaders, quantification of synergistic moralizing trajectories across conceptual categories, and identification of moral change associated with economic and political shifts.

Our work offers a scalable approach to understanding the process of moralization over long periods of time and beyond the psychological analyses of moralization based on isolated individual concepts (e.g.,[6, 10]). By reconstructing moral associations over time, HistMoral not only identifies historical instances of moralization but also uses moral polarity estimates to determine whether these shifts occurred in a morally positive or negative directions—an aspect that has remained largely under-explored in prior moralization research. In all, HistMoral is a highly versatile tool for the computational analysis of moral diachronics and automatic discoveries about moralization under historical, cultural, and political shifts spanning extended time periods.

Early work defined moralization as "a process during which morally neutral concepts gain moral attributes" [1]. This definition, however, raises a deeper question about whether moral neutrality is ever possible. Therefore, more recent psychological theorizing defines moralization more broadly as "increases in the degree to which moral relevance is attached to issues," relaxing initial moral neutrality constraint [2]. Under this new definition, moralization includes two components. The first is moral recognition, which involves the psychological process of attributing moral significance to a concept. The second is moral amplification, which involves the process by which a concept becomes more morally relevant. Our computational framework can incorporate both of these processes that comprise moralization, and shows that many concepts consistently receive very low moral relevance scores suggesting their moral neutrality across historical periods.

Our finding that individuals' names can gain moral associations also raises important questions about how our model differentiates between ideas and practices and individuals. While the psychological mechanisms driving moral association may differ between ideas, actions and persons, the concept of moralization (defined as gaining moral attributes or the intensification of such qualities [2]) applies to both. People, like practices, can become moralized and, as a result, subject to moral judgment or condemnation. However, in such cases, our framework does not distinguish whether the name itself becomes moralized due to repeated associations with a moralized individual, or whether the effect is specific to the individual's actions and context. Nevertheless, as our framework is trained to predict changes in moral association scores (i.e., intuitive associations with moral foundations) it can capture moral framings that go beyond the current definitions of moralization, and track moral dynamics around people's names, global regions, practices, and events.

We use large-scale textual corpora to reconstruct historical patterns of moral perception. While textual data provides valuable insights into societal moral concerns of the time, it also encodes a variety of linguistic signals that may correlate with, but are not specific to, morality. For example, prior research has demonstrated that the frequency of moral foundation terms across historical periods exhibits non-linear trends that often correspond to major global events, such as the world wars [14]. Although our framework incorporates these lexical signals, it moves beyond frequency-based metrics to estimate the moral semantics of words, offering deeper explanatory power than raw lexical statistics.

Valence is another variable closely linked to morality, particularly to moral polarity. Moral condemnation is frequently accompanied by strong affective responses such as anger, disgust, or shame, which are typically high in emotional valence [5]. While it is straightforward to detect high-valence but morally neutral concepts (e.g., *desserts*), the boundary between valence and morality becomes more ambiguous for morally relevant concepts. For instance, smoking may be perceived as both immoral and unpleasant, making it difficult to disentangle moral judgment from affective evaluation [22]. To ensure that our model captures moral perception rather than general sentiment, we incorporated sentiment analysis throughout our experiments as a control.

Although our approach offers scalability and broad historical coverage, it remains primarily descriptive in nature. As such, we cannot determine whether moralization is driven by internal processes or is shaped by external influences such as socio-political conditions. Furthermore, while our findings show that entities like geographic locations and personal names can gain moral associations, it remains an open question whether the mechanisms of moralization operate consistently across different types of entities. For instance, smoking may become moralized as public awareness of its health risks increases. In contrast, in the case of personal names, such as Adolf that may evoke associations with Adolf Hitler, it is unclear whether the name itself becomes morally charged or simply reflects moralization of the individual. This distinction also echoes the conceptual difference between direct and indirect moralization introduced earlier. Specifically, does frequent exposure to morally relevant discourse about an individual (i.e., indirect moralization) strengthen moral associations to the point of producing direct moralization of the individual's name or identity? Addressing such questions will require future research to disentangle the cognitive, linguistic, and social driving forces of moralization across diverse conceptual domains.

By providing high-resolution estimates of moral association across a wide array of concepts, our framework offers a foundation for future studies to uncover the potential causal factors of moralization. However, conducting such investigations at scale remains a significant challenge. Computational approaches are limited in their ability to simulate controlled manipulations since we have limited access to the oft-hidden causes that might have triggered moralization, and even more restricted access to people's minds in a historical setting.

We acknowledge that text, alone, cannot capture societal moral perception comprehensively, as language use recorded in available textual data is affected by external factors such as editorial policies or genres. Moreover, textual data does not include multi-modal signals such as tone, gestures, and visual stimuli. For example, previous work has shown that photographs can improve people's ethical reasoning that text descriptions cannot [62]. Future work can investigate multi-modal approaches that enrich text-based moral association with vision and speech information [63].

Although our framework offers a method for studying moralization over time, it is limited to English corpora and thus reflects the English-speaking population in the United States. In reality, moral views vary across cultures [18]. Our framework also reflects population-level or societal views of concepts, but individuals can also vary in their moralization of various concepts. However, we believe that our framework can be flexibly adapted to cross-cultural or cross-linguistic analyses of moralization due to our theoretical grounding in Moral Foundations Theory [20, 64]. Future work can consider including linguistic data from a more diverse set of cultures and examining corpora from individuals to investigate the variability in moralization that arises due to culture or individual factors.

We have developed a computational framework for the historical reconstruction of moral time courses. Our framework offers a paradigm for moral inference backward in time by combining methods from artificial intelligence with large-scale human data. This approach lends credence to the idea that language use can be a powerful medium to inform the evolutionary trajectories of moral perception in society (e.g.,[21, 65]), and it also creates an opportunity for exploring new tools to perform language-informed moral forecasting into the future (cf.,[39, 66]). Understanding the process of moralization is a long-term endeavor that requires a close integration of theory, formal methodology, and empirical validation. Our work lays a foundation for characterizing moralization through the lens of a historical analysis of the human conceptual system more comprehensively than was previously possible.

Methods

Longitudinal text corpora

We used two longitudinal corpora—the Corpus of Historical American English (COHA) [30] and the New York Times Corpus (NYT) [31]—to reconstruct historical moral time courses. COHA, one of the largest structured corpora of historical English, contains over 475 million words across 100,000 documents and it is organized by decade from the 1810s to the 2000s. Due to document sparsity in the earliest decades, our analysis starts in the 1850s. An alternative to COHA is the Google Books Corpus, which spans a broader period (1500s to 2000s). However, it is limited in that it provides sentence fragments of up to five words (5-grams) as opposed to complete sentences. The NYT corpus serves as a modern longitudinal dataset, spanning from 1987 to June 2007, with over 1.8 million articles. For consistency, we balanced the NYT documents by year. Both of these corpora are tokenized by the publishers. We further extracted the sentences, removed all non-alphanumeric characters, and lemmatized all the tokens using `NLTK` toolkit [67].

Computation of moral scores based on word association data

Our framework builds on previous research in moral association graphs which is limited to estimating moral scores based on word association at a single point in time [25]. Here, we describe how the moral scores are computed. For a given query concept, the *moral relevance score* represents the proportion of morally significant response words (e.g., *bad*) associated with the query word (e.g., *smoking*). The *moral polarity score* reflects the balance of morally positive and negative words linked to the query word. For instance, for a query c associated with T response terms in a word association network, if M of those responses are moral concepts, the moral relevance score is then

$\frac{M}{T}$. From these M responses, if P are morally positive and N are morally negative, the moral polarity score for c is then defined as $\frac{P-N}{2T} + \frac{1}{2}$. Under this formulation, words associated with an equal number of positive and negative moral words receive a moral polarity score of 0.5, while words associated solely with positive or negative moral responses have scores of 1 or 0, respectively.

We used the Moral Foundations Dictionary (MFD) to identify moral terms in word association networks [50]. The MFD has been widely used in computational research on morality [16, 65, 68–70] and was developed for moral analysis based on Moral Foundations Theory [20]. This theory proposes five core foundations to explain universal and culturally specific moral views. These foundations, each with virtue and vice polarities, are Care/Harm, Fairness/Cheating, Authority/Subversion, Loyalty/-Betrayal, and Sanctity/Degradation. We used these moral categories to detect morally positive and negative terms within our dataset. However, since words in the MFD can be classified under moral foundations that may not align with their overall sentiment (e.g., *war* is classified under the Loyalty foundation, a virtue), we used an external dataset of valence ratings [71] for refinement. This dataset provides empirical valence scores for over 13,000 English words, ranging from 1 (most unpleasant) to 9 (most pleasant), with a mean of 5.06 and a median of 5.2. For our analysis, MFD virtue words with valence ratings below 4.5 (e.g., *war*) were reclassified as negative, and vice words with valence ratings above 6 were reclassified as positive. Additionally, we excluded specific MFD terms (*family, mother, father, child, character, protection, cry, clean, cleaning, cleaner, order, dirt,* and *community*) due to their frequent use in non-moral contexts, which could potentially distort the reliability of our moral association metrics. By incorporating both MFD version 1 (which includes a category for general moral terms such as *morality*) and version 2, which contains an expanded lexicon, we compiled a total of $1,718$ moral terms, comprising 703 positive and 994 negative terms (general moral terms are not polarity annotated).

For estimating moral scores, we used the Small World of Words (SWOW) dataset [26], which contains association data for over 12,000 English words, with each word linked to 100 associative responses. To ensure alignment between our textual corpora and the word association dataset, we lemmatized words in SWOW using the `NLTK` package.

Reconstruction model with graph neural networks

To extend the moral association graphs for reconstructing moral scores through historical times, we developed a new graph-based model using graph convolutional network (GCN) [29]. The GCN takes textual representations and word co-occurrence data as input, with target values set to empirical moral scores (moral relevance and moral polarity) from human word association data. After parameter training, this model learns a mapping between moral association and corpus-based co-occurrence data, and it then generalizes this learned mapping to a historical context and therefore estimates moral scores across different time points in a longitudinal corpus (beyond the period in which it was trained on where word association data is available).

During training, the GCN model takes an adjacency matrix $\mathbf{A}$ of a weighted input graph and a feature vector $\mathbf{X}$. The model then generates new representations or target values for each node through a layer-wise propagation rule, expressed as: $\mathbf{H}^{(l+1)} = \sigma(\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}\mathbf{H}^{(l)}\mathbf{W}^{(l)})$. Here, $\tilde{\mathbf{A}}$ represents the normalized adjacency matrix $\mathbf{A}$ with added self-connections, and $\tilde{\mathbf{D}}$ is a diagonal matrix where $\tilde{\mathbf{D}}_{i,i} = \sum_j \tilde{\mathbf{A}}_{i,j}$, and σ stands for a non-linear activation function. The hidden representations of nodes at each layer (l) are represented by $\mathbf{H}^{(l)}$, with the $\mathbf{H}^{(0)}$ being the input feature matrix $\mathbf{X}$. The GCN's parameters are the layer-wise weight matrices $\mathbf{W}^{(l)}$, and these are optimized during training. We use a GCN with two layers and a hidden representation dimension of 128.

In our framework, the adjacency matrix $\mathbf{A}$ is constructed based on word co-occurrences (excluding stop words) within a given snapshot of time in a text corpus. For each word, we identify its top 100 neighbors with the highest degrees of co-occurrence, using a context window of size 1 (bigram co-occurrence). The edges are weighted according to the positive point-wise mutual information (PPMI) metric. Specifically, for words w_i and w_j, the adjacency weight is defined as $\mathbf{A}_{i,j} = \max(0, \log \frac{p(w_i,w_j)}{p(w_i)p(w_j)})$ if w_i and w_j are within each other's top 100 neighbors; otherwise, $\mathbf{A}_{i,j} = 0$. The degree matrix $\mathbf{D}$ is a diagonal matrix defined by $\mathbf{D}_{i,i} = \sum_j \mathbf{A}_{i,j}$. We construct $\tilde{\mathbf{A}}$, the normalized adjacency matrix with self-connection loops, by normalizing the matrix $\mathbf{A} + \mathbf{D}$ so that each row sums to 1.

The input feature matrix $\mathbf{X}$ is constructed based on the contextual representations of words within each time point in the text corpus. These representations are distributed semantic vectors generated by pre-trained large language models, which capture the meanings of words based on their surrounding context. In particular, we use masked language models like BERT [28], which take sentences as input and generate contextualized representations for each token. In our framework, we feed sentences from the text corpus into a BERT model and retrieve word representations from the model's final layer. Here, we report results using the `bert-base-uncased` model (accessed through HuggingFace's `transformers`), but we find our results are robust across alternative models, as shown in *Supplementary Table 1*. We then take an average of each word's contextual representations from all the sentences it appears in, and use these average vectors as rows of the matrix $\mathbf{X}$.

The final representations of words are mapped to their respective moral scores through a linear transformation. We explore three alternative architectures for this mapping. The first architecture, denoted as the 'baseline' model, maps the input feature matrix $\mathbf{X}$ directly to word association scores without employing the GCN. In the second architecture, referred to as the 'GCN' model, the representations from the GCN's final layer ($\mathbf{H}^{(2)}$) are mapped to moral scores. The third approach, the 'Residual GCN' model, combines the outputs from both the GCN and the input features by mapping $\mathbf{H}^{(2)} + \mathrm{ReLU}(\mathbf{XW})$, where $\mathbf{W}$ denotes a linear transformation that aligns the input feature space with the GCN's output space. Evaluation results provided in the *Supplementary Table 1* show that the 'Residual GCN' model outperforms the 'GCN' and the baseline models for both moral relevance and moral polarity estimations. Therefore we adopt 'Residual GCN' model in our experiments.

For training, we use the most contemporary time point in each text corpus (the 2000s in COHA and 2007 in NYT) where word association data are available and can be aligned with textual data. Training targets are derived from empirical moral relevance and polarity scores based on human word association networks in SWOW. We identified $5,864$ cue words in SWOW with a moral relevance score above zero, and supplemented this set with 200 additional words with zero moral relevance scores for negative sampling. From this combined set, $4,681$ words were present in COHA, and $4,821$ were present in NYT. Similarly, for moral polarity, we identified $5,099$ cue words in SWOW with scores greater or less than 0.5, indicating varying degrees of moral positivity and negativity. This set was also supplemented with 200 additional words with moral polarity scores of 0.5, resulting in $4,213$ words available in COHA and $4,821$ in NYT.

For each cross-section of data (COHA/NYT $\times$ moral relevance/moral polarity), we reserved 20% of examples for testing, with results reported in the main text. The remaining 80% was used to train five models, each trained on a different 75/25 split of the data for training and validation, sampled randomly. To improve convergence, we log-transformed the moral relevance and polarity scores using the formula $\log(100p + 1)$, where p is the score for a given word. We trained using a Gaussian negative log-likelihood loss function, treating the moral scores as samples from Gaussian distributions with predicted expectations and fixed variances of 1. We used a base learning rate of 0.01 with a decaying factor of 0.1 at steps 100 and 500. We trained the models for 10000 steps, and selected the checkpoints with the best performance on the validation set.

After training, we applied these models to historical corpora at different time points, with each time point comprising over $10,000$ lexical items, including unigrams and a curated selection of bigrams. Specifically, at each time point in both the COHA and NYT corpora, we extracted the $10,000$ most frequent unigrams. Additionally, we supplemented these items with specific concepts drawn from the Small World of Words dataset and a curated set of bigrams representing major global events, technologies, diseases, and notable figures (e.g., *genetic engineering, civil war, yellow fever, President Obama, Joseph Stalin*). These bigrams were identified using relevant Wikipedia entries, and we have included the full list in our repository to facilitate reproducibility. The inclusion of bigrams alongside unigrams was intended to capture a broader range of concepts that single-word terms alone could not adequately express. In the *Supplementary Table 9*, we provide the total number of extracted concepts (unigrams and bigrams) from COHA and NYT at each time point.

For each cross-section of data, we standardized the outputs from all five models using z-score standardization, reporting the final moral scores as the mean of these standardized predictions. Under this scheme, words with negative moral relevance scores are more morally irrelevant or neutral than average, while those with positive moral relevance scores are more morally relevant. Similarly, words with negative moral polarity scores are linked to more negative moral concepts, while positive moral polarity scores indicate words associated with more positive moral concepts relative to the global average.

Indirect moralization

We compare our reconstruction performance with a baseline that only uses contextual co-occurrence information. Specifically, we define a new metric called indirect moralization (denoted as Ind). To estimate indirect moralization at each time point, we use the normalized adjacency matrix $\tilde{\mathbf{A}}$, and gather the degree of co-occurrence with moral terms (drawn from MFD) for each concept c: $\text{Ind}(c) = \sum_t \tilde{\mathbf{A}}_{c,t} \mathbb{1}(t \in \text{MFD})$. By definition, indirect moralization lies in the range of 0 to 1, where 0 means none of the word's nearest neighbors are moral terms, and 1 means all the word's nearest neighbors are moral terms.

Sentiment scores

We use the `SentimentIntensityAnalyzer` module from `NLTK` library as a standard sentiment analysis toolkit. We assign the `compound` score from this module to the sentences in our textual corpora. The `compound` score is normalized within the range of -1 (most negative) to 1 (most positive). To estimate words' sentiment scores at each time point, we take an average of the `compound` scores of all sentences the words appear in.

Model evaluation

To evaluate the effectiveness of our framework across historical and empirical datasets, we excluded terms that occur fewer than 50 times at a given time point in COHA to ensure reliable model estimation.

We identified disease terms using independent sources: 1) Wikipedia lists of diseases and epidemics, and 2) prototypical disease terms generated empirically by human participants [49]. The intersection of terms from these independent sources with COHA resulted in a set of 22 disease terms, which are *anthrax, cancer, cholera, diabetes, diphtheria, flu, hepatitis, hiv, hivaids, influenza, leukemia, malaria, measles, plague, polio, salmonella, scarlet fever, smallpox, tuberculosis, typhoid, typhus*, and *yellow fever*. In selecting the randomized control sets, each (disease, time point) pair was matched with a randomly selected non-disease term from the same time point, and this sampling procedure was repeated 1,000 times.

Using Wikipedia, we identified a list of notable political leaders from the past century. Additionally, we included the presidents of the United States in this list. In the *Supplementary Table 10* we present these leaders along with the query terms used to locate them in COHA. If a leader could be identified using multiple query terms (e.g., *lenin* and *vladimir lenin*), we calculated the average moral scores for these terms. To compare political leaders with other names, we use the SSA dataset of popular baby names in the United States. Similar to the disease analysis, each (political leader, time point) pair was matched to a randomly selected name from the same time point, where the time point was selected such that it corresponded to the duration of the political leader's term in office. This sampling procedure was repeated for 1,000 times.

To analyze the effect of political conflicts on moralization, we used the Historical Conflict Event Dataset [41]. This dataset offers a comprehensive description of conflicts over the world, in a timescale from 1468 BC to the invasion of Iraq in 2003. The dataset further provides geographic coordinates and the year(s) of the conflict, along with its participants and winning and losing sides (which are mostly country names). Conflicts are also annotated with the Lehmann Zhukov scale, which specifies the size of the battle [72]. Using this scale, we selected conflicts with Lehmann Zhukov scale of at least 3, which corresponds to battles with 20,000 to 100,000 men on either sides to ensure sufficient representation of the battle in COHA. For example, the dataset identifies the term *vietnamese* as a participant in the Vietnam War during the decade of 1970s. In our experiment, for each participant term we identified all the decades where the participant was involved in a conflict (wartime) and was not involved in any conflicts (peacetime). We then compared the average moral scores between wartime and peacetime using paired t-tests.

Analysis of conceptual categories

We used a category norm production database to examine how moralization occurs across related concepts. This database includes 117 conceptual categories (67 concrete and 50 abstract) populated with norms generated by 64 human participants, who were asked to recall as many members of each category as possible within 60 seconds [49]. The categories in this database were selected from previous categorization literature [42–48], and spanned different taxonomic levels (e.g., "crime," "non-violent crime," and "violent crime"). This procedure resulted in a dataset of cognitively related concept clusters. We chose this database as opposed to the previous ones because it covered a larger set of categories balanced across concrete and abstract concepts, and was collected more recently, which aligns better with the period in which the word association data we used for training our models (SWOW [26]) was collected. To ensure the reliability of this crowd-sourced database, we filtered out any concepts that were nominated by only a single participant within each category. To assess each category's moral relevance, we averaged the moral relevance scores of all its member concepts over time points in the dataset. This approach provided an estimate of the overall moral relevance of each category and allowed us to compare categories in moral space. To evaluate the rate of moralization within each category, we regressed each category's moral relevance scores against standardized historical time points from COHA, using the year coefficient as a measure of the category's rate of moralization.

We concentrated our analysis on categories with an overall positive moral relevance score, to examine whether concepts within these (morally relevant) categories exhibited similar moral trajectories. For this purpose, we calculated pairwise Pearson's correlation coefficients for the moral relevance scores between concepts within the same category and across different categories. Correlations were restricted to concept pairs that overlapped in at least ten decades (equivalent to one century) within COHA. The results of these pairwise correlations are presented in Figure 3b, where intra-category (diagonal) and inter-category comparisons are both shown. Formally, the correlation values between categories of C_1 and C_2 (which can be the same) are calculated using the average correlation values between category members $\mathrm{corr}(C_1, C_2) = \frac{1}{N} \sum_{w_1 \in C_1} \sum_{\substack{w_2 \in C_2 \\ w_2 \neq w_1}} \mathrm{corr}(w_1, w_2)$, where $N = \sum_{w_1 \in C_1} \sum_{\substack{w_2 \in C_2 \\ w_2 \neq w_1}} 1$.

To assess whether concepts within the same category share more aligned moral relevance trajectories, we generated random categories with matched membership counts and compared the average intra-category correlation (i.e., weighted average of the diagonal entries in Figure 3b). This permutation test was repeated 1,000 times, with results reported in the main text.

Analysis of moral change

To investigate the relationship between moral change and fluctuations in consumer product price, we used average retail prices reported by the U.S. Bureau of Labor Statistics. This database offers monthly price statistics for various categories of food items, utility gas, fuel oil, electricity, and automotive fuels. To quantify the total annual change in product prices, we compared monthly prices to those from the same month in the previous year, calculating the average price change (relative change) for each product over the 12-month period, over the years overlapping with the NYT corpus. We matched each product item to a corresponding token that best represents it; for instance, the product "Apples, Red Delicious, per lb. (453.6 gm)" is mapped to the token *apple*. In the *Supplementary Table 5*, we show the list of products in this database along with their simplified names used in our analyses. Using these matched names, we estimated the annual changes in moral scores, which had been previously normalized through z-score standardization. This process yielded tuples structured as (product item, moral association change, relative price change, year of change). We only included example where moral relevance of the product increased over the span of a year, in order to focus our investigation on items with sufficient moral relevance signals. This filtering resulted in a total of 704 examples.

To examine the effect of political shifts on moral change, we use the Congressional Record for the 43rd-114th Congresses database [59]. This database contains transcripts of speeches delivered on the floors of the United States House of Representatives and Senate from the 43rd to the 114th Congress. Additionally, it provides key unigrams and bigrams manually classified into 22 substantive political topics (topics shown on the y-axis of Figure 4c). Since our framework primarily uses unigram concepts, with a small subset of manually selected bigrams, many bigram phrases in the Congressional Record database (e.g., *veteran bill*) are not represented in our lexicon (these concepts also tend to be sparsely represented in the corpora we analyzed). Furthermore, some bigrams are presented in wildcard formats (e.g., *econom* challeng**), allowing for matches to multiple items (e.g., *economical challenge* and *economy challenge*). To estimate the moral scores for these bigrams, we calculated the average moral scores of their constituent words. For example, the moral relevance score of *econom* challeng** was derived from the average of the moral relevance scores of *economy*, *economical* (first constituents), and the moral relevance score of *challenge* (second constituents).

Data availability

The historical moral association time courses generated in this study have been deposited in the OSF repository with DOI https://doi.org/10.17605/OSF.IO/KJYNQ under Files section. The Corpus of Historical American English (COHA) [30] is available with purchase at https://www.english-corpora.org/coha/. We accessed the New York Times Corpus [31] through the University of Toronto's institutional membership in the Linguistic Data Consortium (LDC; LDC2008T19). The dataset is also publicly available at https://huggingface.co/datasets/irds/nyt. The Moral Foundations Dictionary [73] is available at https://doi.org/10.17605/OSF.IO/EZN37. The English word association network (SWOW [26]) can be accessed through https://smallworldofwords.org/en/project/research, and the moral association scores are available at https://osf.io/pe6qt/wiki/home/?view_only=6781f237174a4eb7ae2b0e826fb2fb8c. Disease and political terms can be found on the following Wikipedia pages: https://en.wikipedia.org/wiki/List_of_infectious_diseases, https://en.wikipedia.org/wiki/List_of_epidemics_and_pandemics, https://en.wikipedia.org/wiki/List_of_human_disease_case_fatality_rates, https://en.wikipedia.org/wiki/Lists_of_state_leaders_by_century. The list of popular names in the United States is available on the Social Security Administration website https://www.ssa.gov/oact/babynames/limits.html. The Historical Conflict Event Dataset [41] can be accessed through https://journals.sagepub.com/doi/10.1177/00220027221119085. Conceptual category norm dataset is available at https://osf.io/jgcu6/. Average price dataset is available at the U.S. Bureau of Labor Statistics website https://www.bls.gov/cpi/factsheets/average-prices.htm. Congressional speech record dataset [59] is available at https://data.stanford.edu/congress_text. The Gallup survey series is available under restricted access at https://www.gallup.com/analytics/214565/universities-colleges-using-gallup-analytics.aspx, which we obtained through the University of Toronto's library.

Code availability

Code for replicating our analyses is deposited at https://doi.org/10.5281/zenodo.17692655 [74]. Additionally, an open-access tool for visualizing the moral time courses can be found at https://warz.shinyapps.io/MoralityVisualizer/.

References

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

[2] Rhee, J. J., Schein, C. & Bastian, B. The what, how, and why of moralization: A review of current definitions, methods, and evidence in moralization research. *Social and Personality Psychology Compass* **13**, e12511 (2019).

[3] Hobbes, T. *Leviathan, or the Matter, Forme, and Power of a Commonwealth, Ecclesiasticall and Civil* (Andrew Crooke, London, 1651).

[4] Dennett, D. C. *Darwin's Dangerous Idea: Evolution and the Meanings of Life* (Simon & Schuster, New York, 1995).

[5] Wisneski, D. C. & Skitka, L. J. Moralization through moral shock: Exploring emotional antecedents to moral conviction. *Personality and Social Psychology Bulletin* **43**, 139–150 (2017).

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

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

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

[9] Wang, Y. *et al.* Moralization of e-cigarette use and regulation: A mixed-method computational analysis of opinion polarization. *Health Communication* **0**, 1–11 (2022).

[10] Rozin, P. & Singh, L. The moralization of cigarette smoking in the United States. *Journal of Consumer Psychology* **8**, 321–337 (1999).

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

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

[13] Appiah, K. A. *The honor code: How moral revolutions happen* (WW Norton & Company, 2011).

[14] Wheeler, M. A., McGrath, M. J. & Haslam, N. Twentieth century morality: The rise and fall of moral concepts from 1900 to 2007. *PLoS one* **14**, e0212267 (2019).

[15] Sagi, E. & Dehghani, M. Measuring moral rhetoric in text. *Social Science Computer Review* **32**, 132–144 (2014).

[16] Garten, J. *et al.* Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis: Distributed dictionary representation. *Behavior Research Methods* **50**, 344–361 (2018).

[17] Lin, Y. *et al. Acquiring background knowledge to improve moral value prediction. 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 552–559 (IEEE, 2018).

[18] Awad, E. *et al.* The Moral Machine experiment. *Nature* **563**, 59–64 (2018).

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

[20] Graham, J. *et al.* in *Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism Advances in Experimental Social Psychology*, Vol. 47 55–130 (Elsevier, 2013).

[21] Xie, J. Y., Ferreira Pinto Junior, R., Hirst, G. & Xu, Y. *Text-based inference of moral sentiment change. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4654–4663 (Association for Computational Linguistics, Hong Kong, China, 2019).

[22] Haidt, J. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review* **108**, 814 (2001).

[23] Nelson, D. L., McEvoy, C. L. & Dennis, S. What is free association and what does it measure? *Memory & cognition* **28**, 887–899 (2000).

[24] De Deyne, S. & Storms, G. Word associations: Network and semantic properties. *Behavior Research Methods* **40**, 213–231 (2008).

[25] Ramezani, A. & Xu, Y. Moral association graph: A cognitive model for automated moral inference. *Topics in Cognitive Science* **17**, 120–138 (2025).

[26] De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M. & Storms, G. The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods* **51**, 987–1006 (2019).

[27] Wettler, M. & Rapp, R. *Computation of word associations based on co-occurrences of words in large corpora. Very Large Corpora: Academic and Industrial Perspectives* (1993).

[28] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Burstein, J., Doran, C. & Solorio, T. (eds) *BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019).

[29] Kipf, T. N. & Welling, M. *Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations* (2017).

[30] Davies, M. *The Corpus of Historical American English: 400 million words, 1810–2009* (2010).

[31] Sandhaus, E. The New York Times Annotated Corpus. *Linguistic Data Consortium, Philadelphia* **6**, e26752 (2008).

[32] Kraaijeveld, S. R. & Jamrozik, E. Moralization and mismoralization in public health. *Medicine, Health Care and Philosophy* **25**, 655–669 (2022).

[33] Cochran, J. K., Will, J. A. & Garner, J. *The moralization of illness: The role of moral values in the religious framing of the aids problem. Research in the Social Scientific Study of Religion, Volume 6*, 209–228 (Brill, 1999).

[34] Brooks, E. "Don't Be a Knucklehead": Moralizing Disability in New Jersey's Pandemic Response and Rhetoric. *Disability Studies Quarterly* **41** (2021).

[35] Haidt, J. & Graham, J. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research* **20**, 98–116 (2007).

[36] Amin, A. B. *et al.* Association of moral values with vaccine hesitancy. *Nature Human Behaviour* **1**, 873–880 (2017).

[37] Voelkel, J. G. & Feinberg, M. Morally reframed arguments can affect support for political candidates. *Social Psychological and Personality Science* **9**, 917–924 (2018).

[38] Kreitzer, R. J., Kane, K. A. & Mooney, C. Z. The evolution of morality policy debate: Moralization and demoralization **17(1)**, 3–24 (2019).

[39] Strimling, P., Vartanova, I., Jansson, F. & Eriksson, K. The connection between moral positions and moral arguments drives opinion change. *Nature Human Behaviour* **3**, 922–930 (2019).

[40] Everett, J. A. C. *et al.* Political differences in free will belief are associated with differences in moralization. *Journal of Personality and Social Psychology* **120**, 461 (2021).

[41] Miller, C. & Bakar, K. S. Conflict events worldwide since 1468bc: Introducing the historical conflict event dataset. *Journal of Conflict Resolution* **67**, 522–554 (2023).

[42] Battig, W. F. & Montague, W. E. Category norms of verbal items in 56 categories a replication and extension of the connecticut category norms. *Journal of Experimental Psychology* **80**, 1 (1969).

[43] Capitani, E., Laiacona, M., Mahon, B. & Caramazza, A. What are the facts of semantic category-specific deficits? A critical review of the clinical evidence. *Cognitive Neuropsychology* **20**, 213–261 (2003).

[44] Larochelle, S., Richard, S. & Soulieres, I. What some effects might not be: The time to verify membership in "well-defined" categories. *The Quarterly Journal of Experimental Psychology Section A* **53**, 929–961 (2000).

[45] McEvoy, C. L. & Nelson, D. L. Category name and instance norms for 106 categories of various sizes. *The American Journal of Psychology* **95**, 581–634 (1982).

[46] Rosch, E. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General* **104**, 192 (1975).

[47] Uyeda, K. M. & Mandler, G. Prototypicality norms for 28 semantic categories. *Behavior Research Methods & Instrumentation* **12**, 587–595 (1980).

[48] Van Overschelde, J. P., Rawson, K. A. & Dunlosky, J. Category norms: An updated and expanded version of the norms. *Journal of Memory and Language* **50**, 289–335 (2004).

[49] Banks, B. & Connell, L. Category production norms for 117 concrete and abstract categories. *Behavior Research Methods* **55**, 1292–1313 (2023).

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

[51] Boyd, R. & Richerson, P. J. Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **364**, 3281–3288 (2009).

[52] Haidt, J. The righteous mind: Why good people are divided by politics and religion. *New York Pantheon* (2012).

[53] Enke, B. Moral values and voting. *Journal of Political Economy* **128**, 3679–3729 (2020).

[54] Enke, B. Market exposure and human morality. *Nature Human Behaviour* **7**, 134–141 (2023).

[55] Enke, B. Morality and political economy from the vantage point of economics. *PNAS Nexus* **3**, pgae309 (2024).

[56] Stellar, J. E. & Willer, R. The corruption of value: Negative moral associations diminish the value of money. *Social Psychological and Personality Science* **5**, 60–66 (2014).

[57] Bennett, R. M., Anderson, J. & Blaney, R. J. Moral intensity and willingness to pay concerning farm animal welfare issues and the implications for agricultural policy. *Journal of Agricultural and Environmental Ethics* **15**, 187–202 (2002).

[58] Liebe, U., Preisendörfer, P. & Meyerhoff, J. To pay or not to pay: Competing theories to explain individuals' willingness to pay for public environmental goods. *Environment and Behavior* **43**, 106–130 (2011).

[59] Gentzkow, M., Shapiro, J. M. & Taddy, M. *Congressional record for the 43rd-114th congresses: Parsed speeches and phrase counts* (2018).

[60] Wang, S.-Y. N. & Inbar, Y. Moral-language use by us political elites. *Psychological Science* **32**, 14–26 (2021).

[61] Alvarez-Galvez, J., Cruz, F. L. & Troyano, J. A. Discovery and characterisation of socially polarised communities on social media. *Scientific Reports* **13**, 15439 (2023).

[62] Coleman, R. The effects of visuals on ethical reasoning: What's a photograph worth to journalists making moral decisions? *Journalism & Mass Communication Quarterly* **83**, 835–850 (2006).

[63] Zhu, W., Ramezani, A. & Xu, Y. *Visual moral inference and communication. Proceedings for the 47nd Annual Meeting of the Cognitive Science Society* (2025).

[64] Haidt, J. & Joseph, C. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus* **133**, 55–66 (2004).

[65] Ramezani, A., Stellar, J. E., Feinberg, M. & Xu, Y. Evolution of the Moral Lexicon. *Open Mind* **8**, 1153–1169 (2024).

[66] Strimling, P., Vartanova, I. & Eriksson, K. Predicting how US public opinion on moral issues will change from 2018 to 2020 and beyond. *Royal Society Open Science* **9**, 211068 (2022).

[67] Bird, S., Klein, E. & Loper, E. *Natural language processing with Python: analyzing text with the natural language toolkit* ("O'Reilly Media, Inc.", 2009).

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

[69] Hoover, J. *et al.* Moral Foundations Twitter Corpus: A collection of 35k Tweets Annotated for Moral Sentiment. *Social Psychological and Personality Science* **11**, 1057–1071 (2020).

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

[71] Warriner, A. B., Kuperman, V. & Brysbaert, M. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods* **45**, 1191–1207 (2013).

[72] Lehmann, T. C. & Zhukov, Y. M. Until the bitter end? The diffusion of surrender across battles. *International Organization* **73**, 133–169 (2019).

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

[74] Ramezani, A., Stellar, J. E., Feinberg, M. & Yang, X. Code for historical reconstruction of moralization (2025). URL https://doi.org/10.5281/zenodo.17692655.

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

Acknowledgments

We thank Graeme Hirst, Marco Baroni, Spike Lee, and Pontus Strimling for their feedback on the manuscript. We are grateful to Warren Zhu for developing the interactive online visualizer. YX is funded by a NSERC Discovery Grant RGPIN-2018-05872, and an Ontario Early Researcher Award #ER19-15-050.

Author contributions

AR and YX conceptualized the study. AR, YX, JS, and MF designed the study and interpreted the analysis. AR developed the model and analyzed the data. AR and YX wrote the manuscript. AR, YX, JS, and MF edited and approved the submitted manuscript. YX acquired funding.

Competing interests

The authors have no competing interests to declare.

Model	Theoretical grounding	Historical inference	Semantic enrichment	Domain flexibility
Lexical models [14, 15] (e.g., word co-occurrence statistics)	✗	✓	✗	✓
Word embedding models [16, 21]	✗	✓	✓	✓
Neural networks [17, 75]	✗	✗	✓	✗
Large language models [19]	✗	✗	✓	✗
Moral association graph model [25]	✓	✗	✓	✗
HistMoral framework	✓	✓	✓	✓

Table 1: Comparison of existing models and HistMoral framework for moral reconstruction. Representative studies using each methodology are cited in the first column.

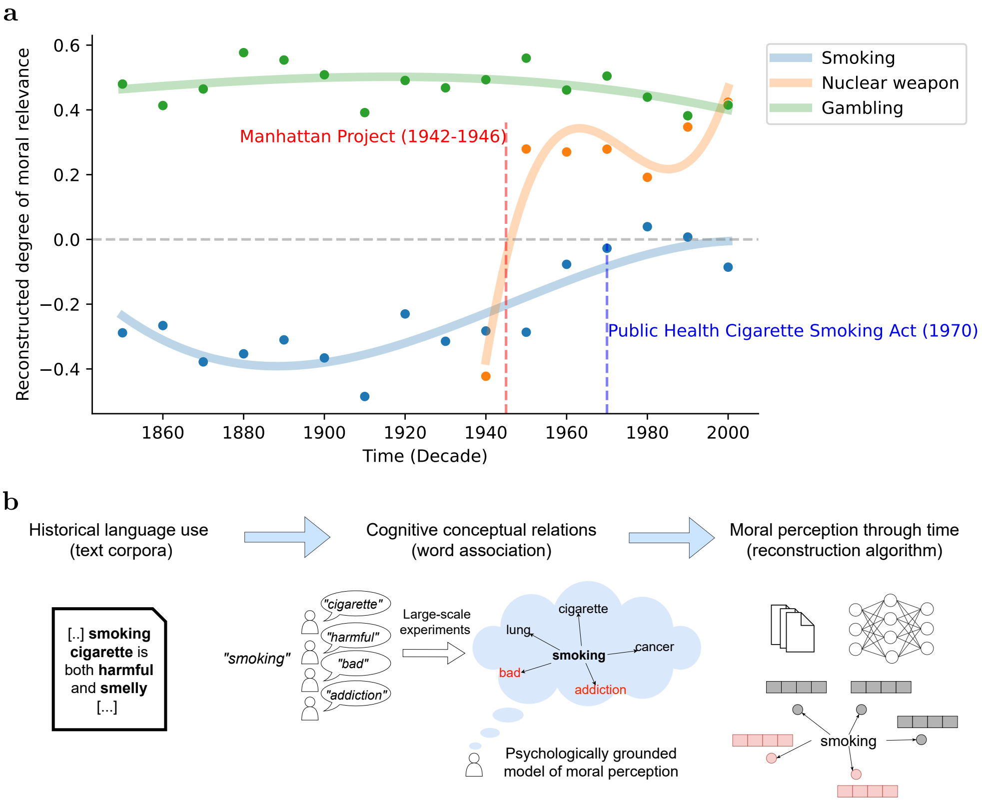

Fig. 1: Overview of HistMoral, a computational framework for historical moral reconstruction. a) Reconstructed moral time courses of the concepts *smoking*, *nuclear weapons*, and *gambling*. The time courses were reconstructed by estimating association networks across different historical periods using the *Corpus of Historical American English* (*COHA*). Vertical dashed lines indicate relevant historical events including the “Manhattan Project” and the “Public Health Cigarette Smoking Act” in the United States. The moral time courses are smoothed with a cubic spline for visualization. Values on the y-axis are normalized z-scores, where zero (horizontal dashed line) represents the average of estimated moral relevance for all concepts in the dataset. b) An illustration of how HistMoral reconstructs moral time courses of concepts in history. Our approach is grounded in conceptual relationships (operationalized through large-scale word association experiments), in which associations with moral terms (i.e., *bad*) reveal population-level moral perception. We use diachronic textual corpora to approximate the word association network in historical periods. In this approximated network, concepts are connected based on their degree of co-occurrences (edges) and represented using contextual semantic vectors. Our algorithm operates on these approximated networks to reconstruct moral perception through time.

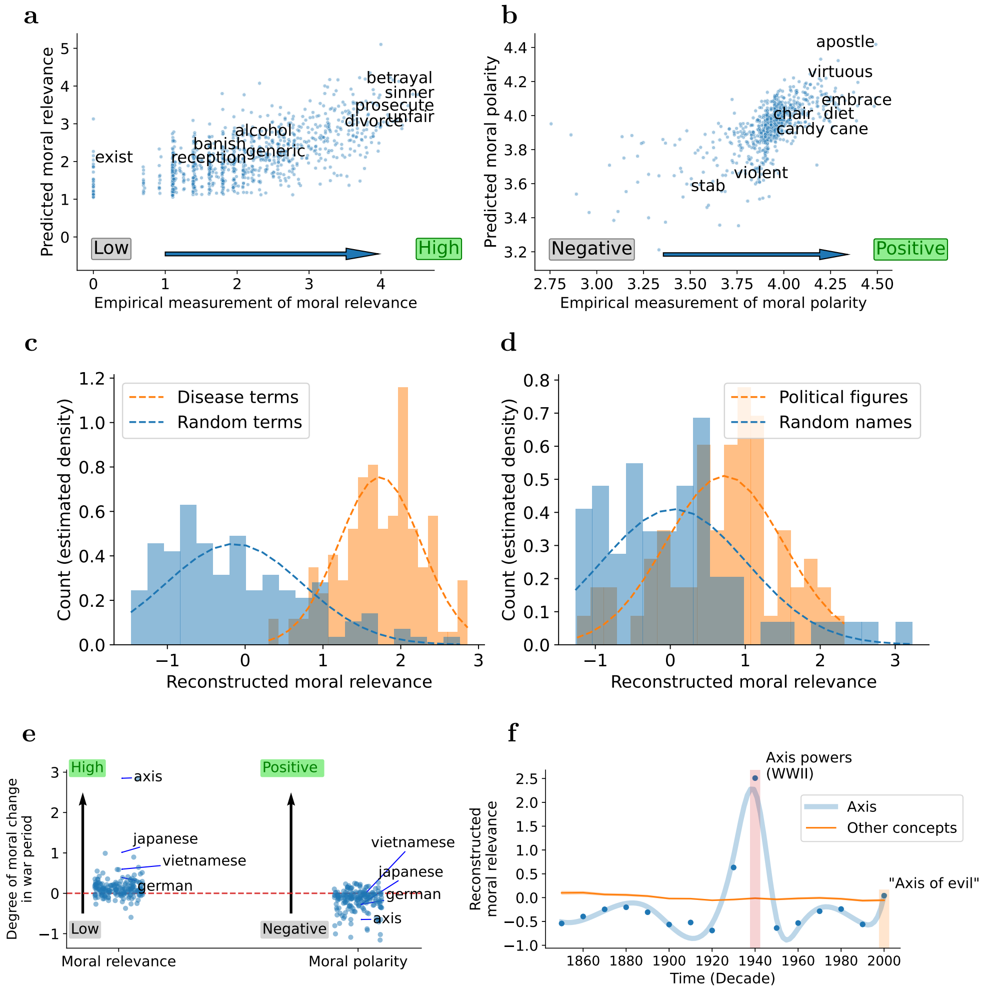

Fig. 2: Model reconstruction of empirical and historical trends of moralization. Validation of model-estimated scores against empirical human data based on word association in terms of a) moral relevance (sample size = 937), and b) moral polarity (sample size = 843). Each dot is a concept, and sample concepts are printed in black. c) Comparison of model-estimated historical moral relevance for disease-related terms (sample size = 135), and d) names of political figures (sample size = 65) against randomly selected control groups of equivalent sample size. The moral relevance scores are standardized by z-scores. e) Historical changes observed in the model-reconstructed moral relevance and polarity for the participants in international conflicts, where a change is quantified as the offset (in moral score) between war periods and peaceful periods. f) Reconstructed time course of moral relevance for the term *axis*, smoothed using a cubic spline. The baseline represents the average of estimated moral relevance for all concepts in COHA. The first vertical line marks the 1940s time point corresponding to World War II which took place between 1939 and 1945, and the second one corresponds to the first usage of the phrase “Axis of evil” by George W. Bush in 2002. All moral scores in subplots c, d, e, and f are shown after z-score standardization.

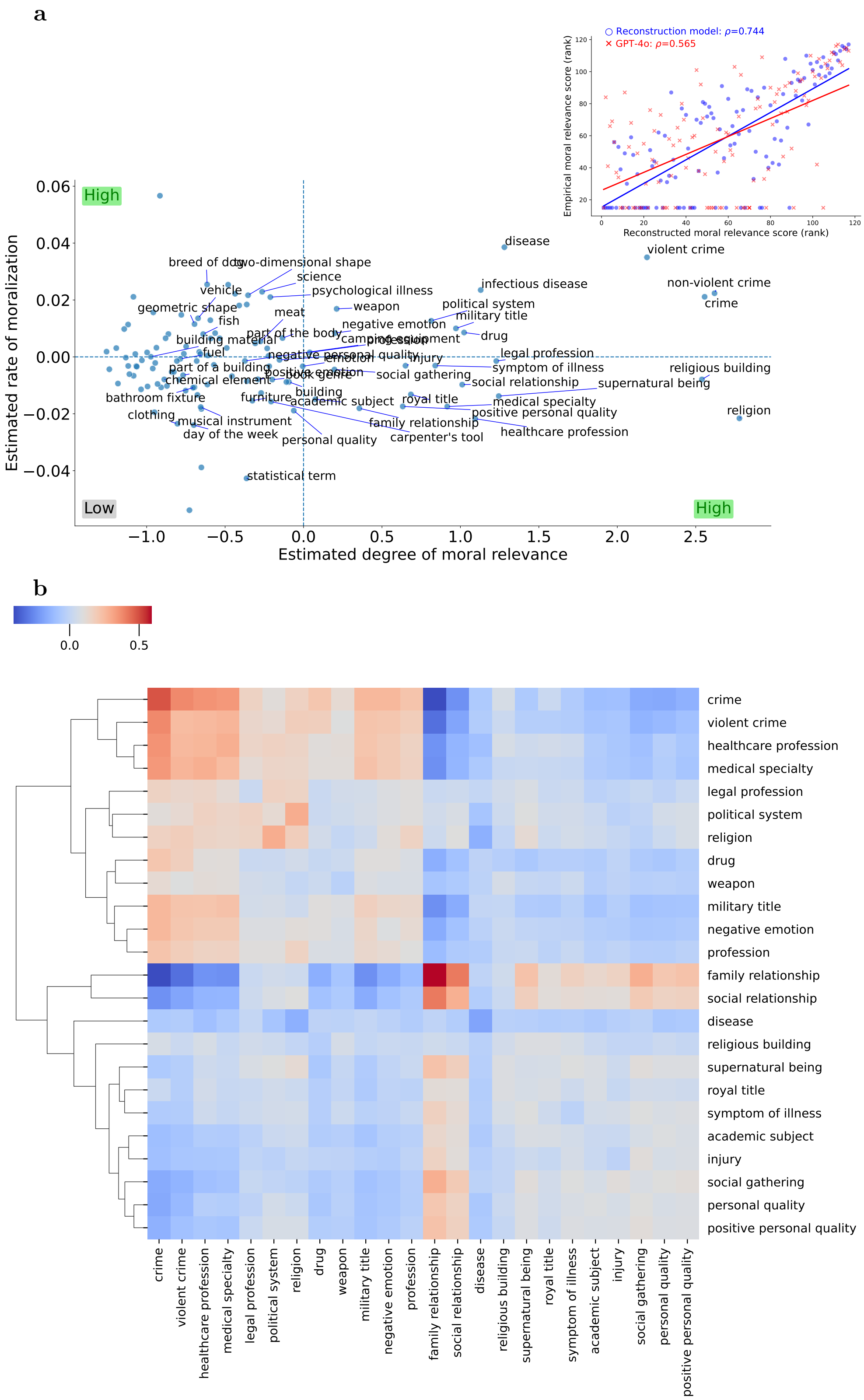

Fig. 3: Moralization within and across different conceptual categories. Comparison of model-estimated moral relevance and moralization rate for 117 conceptual categories. The inset scatter plot shows the Spearman correlation between empirical moral relevance scores for these 117 categories derived from word association data, and model-predicted moral relevance scores derived from our framework (reconstruction model) ($\rho = 0.744$, $P < 0.0001$, 95% CI $= [0.65, 0.82]$, $n = 117$), and by GPT-4o ($\rho = 0.565$, $P < 0.0001$, 95% CI $= [0.43, 0.68]$, $n = 117$). b) Clustered-heatmap of intra (diagonal entries) and inter (non-diagonal entries) pair-wise correlations in moral time courses for concepts drawn from different conceptual categories. Categories are selected such that their average moral relevance score across history is above zero.

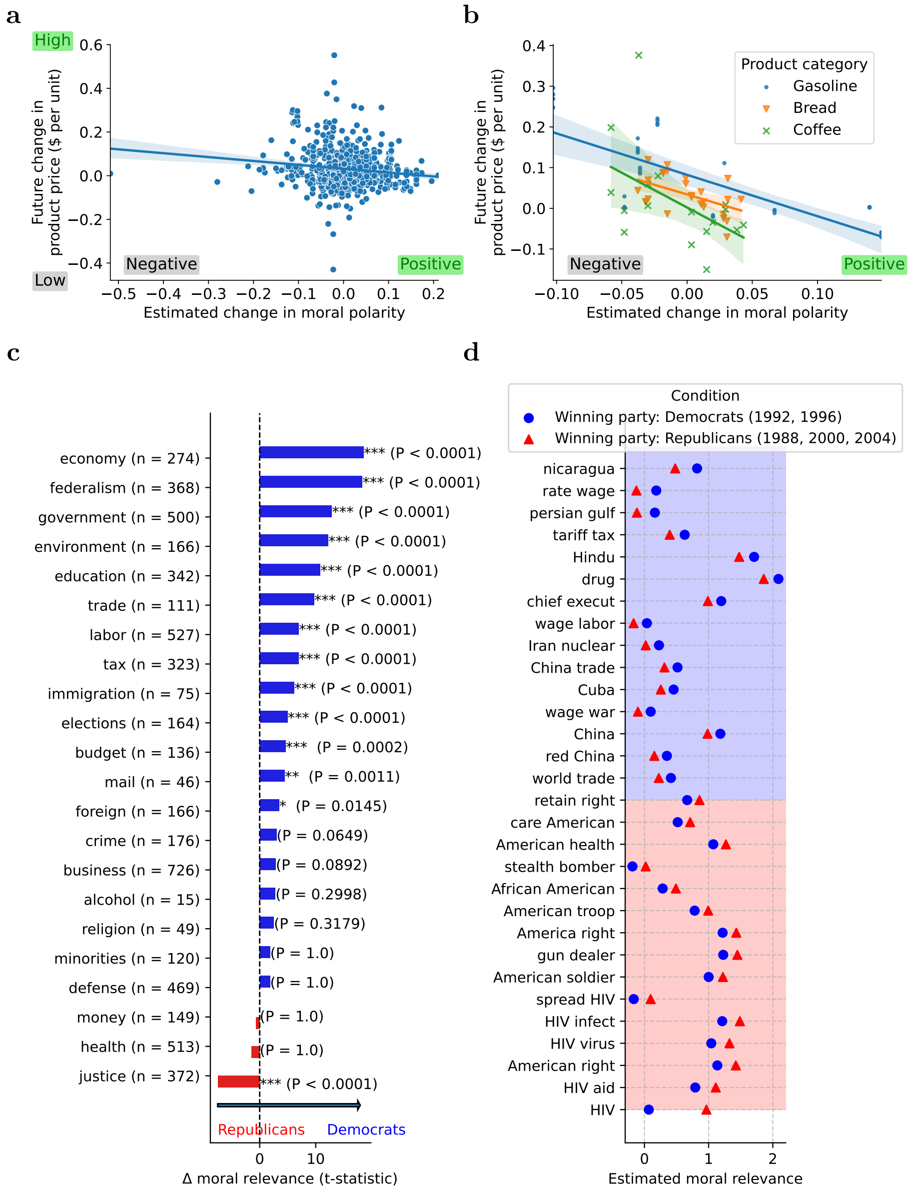

Fig. 4: Summary of results on moral change under economic-political shifts. a) A scatter plot showing the relationship between increases in annual retail price and changes in the perceived moral polarity toward products. Each dot represents the status of a retail product over two consecutive years. Shaded region represents the 95% confidence interval for the fitted regression line. b) Visualization of the model-estimated changes in moral polarity and the corresponding annual price increases for various products of gasoline, beef, and bread. Shaded regions represent the 95% confidence interval for the fitted regression lines. c) Degree of change in moral relevance for 22 topics discussed in U.S. Congressional speeches when different parties win the presidential election in the United States. Blue bars show topics that are more morally relevant when Democrats win the election, and red bars indicate topics that are more morally relevant when Republicans win. The asterisks indicate the Bonferroni-adjusted significance levels ("*," "**," "***" for $P < 0.05$, 0.01, 0.001 respectively) from multiple two-tailed t-tests. d) Phrases with the largest degree of moral relevance gain when Democrats win the presidential election (top blue plot), and when Republicans win (bottom red plot).