Moral conviction in political text

14/08/2023

Word count 9995

Abstract

Research on the effects of morality in political rhetoric and as an attitudinal dimension has accelerated in the last decade, but our ability to measure morality in political texts remains methodologically limited and conceptually tied to a specific framework of morality. At the same time, the emergence of large language models has transformed natural language processing, as new models are significantly more capable and flexible than previous tools. This paper tests the ability of a large language model—GPT—to identify moral conviction in legislative speech and open-ended survey response across the US and UK. I perform a series of initial experiments demonstrating GPT can identify moral conviction similarly to a human. I then test its ability at scale, first in a novel survey experiment, and second in a replication of a previous study on the effects of moralising rhetoric. I find that GPT's understanding of moral conviction in real legislative speech aligns with that of humans, and that it is better at recovering the effects of moral rhetoric in observational contexts than alternative approaches. My findings have implications for the measurement of morality in political text, the use of LLMs in political science, and study of the effects of moralising political rhetoric.

30-page version

The original version of this paper has three studies. The first study validates the GPT measure against human labellers, the second against in a survey experiment, comparing it to the moral foundations dictionary, and the third study replicates Jung (2020). The full paper is 35 pages long.

To keep within the page limit, this 30-page version removes:

- 1.6.2 Qualitative error analysis: thematic analysis of GPT's classification errors

The full paper is available here:
https://github.com/joehigton/diss/blob/main/full_dissertation.pdf

Table of contents

Tables and figures

i Introduction and literature review

Morality is a central issue in politics. Politicians, depending on their party and audience, emphasise different moral values (Kraft & Klemmensen, 2023; Lipsitz, 2018), while different people often hold wildly different moral values (Haidt & Graham, 2007; Haidt & Joseph, 2004).

In the context of increasing polarisation, researchers have started to look beyond variation in moral values and consider which issues are matters of morality in the first place. This is termed 'moralisation' or 'moral conviction'.[1] While related to moral values, it is conceptually distinct: moral conviction concerns to what extent is an issue a matter of right and wrong, whereas moral values underpin specific ideas of right and wrong. For instance, abortion is almost universally regarded as a moral issue, but liberals and conservatives highlight different moral values to justify their positions. An issue like housing, on the other hand, is sometimes moralised, but sometimes not. A moralised view of housing might focus on fairness, while a non-moralised view might stress its economic impact. Contrast a Guardian editorial on UK housing: "house prices rising faster than incomes are a gross injustice,"[2] to a Telegraph article: "a failure to build enough housing has trapped more money in land rather than productive investments."[3] The former implies a moral stance, the latter a pragmatic one.

Moralisation in politics has a number of important consequences. Moralising rhetoric by politicians has been shown to impact outcomes from turnout (Jung, 2020) to perceived candidate sincerity (Clifford & Simas, 2022). On the citizen side, viewing an issue as a matter of right and wrong has been linked to polarisation and motivated reasoning (Ryan, 2017, 2019) and moral conviction has been shown to shape activism behaviour (Sabucedo et al., 2018) and other forms of participation (Morgan et al., 2010; Ryan, 2014; Skitka et al., 2015).

Measuring concepts like moral conviction is difficult: it is subjective, context-dependent and nuanced (Skitka et al., 2021). Studies typically use vignettes to measure the effects of moralising rhetoric (e.g. Simonsen & Bonikowski, 2022) and survey instruments to measure attitudinal moral conviction in individuals (e.g. Hornsey et al., 2003; Ryan, 2014; Skitka et al.,

[1] I use both terms in this paper.

[2] Guardian Editorial Board (2022), "The Guardian view on housing costs: a grave and growing injustice", *Guardian,* 31 July 2022 https://www.theguardian.com/commentisfree/2022/jul/31/the-guardian-view-on-housing-costs-a-grave-and-growing-injustice

[3] Lawford, Melissa (2023), "How Britain's broken housing market is crushing growth" *Telegraph,* 12 February 2023, https://www.telegraph.co.uk/business/2023/02/12/how-britains-broken-housing-market-crushing-growth/

2005). Though widespread in research on moral values, text-as-data methods are rarely applied to moralisation, even though they have been applied to similar rhetorical and attitudinal constructs (e.g. emotions, Seyeditabari et al., 2018). Previous research that does measure moralisation in political text uses rule-based methods designed for moral values (Jung, 2020). These are suboptimal for measuring moralisation.

At the same time, following the development of the Transformer architecture (Vaswani et al., 2017) complex large language models have been developed (e.g. OpenAI's GPT-3, or Google's BERT) that are able to engage in complex reasoning patterns and understand nuanced or subtextual implications of text. These models have been shown to perform very well across a diverse range of baseline natural language processing (NLP) tasks (Brown et al., 2020) and may present a useful tool for measuring moral conviction in text.

This paper tests the capability of GPT, a large language model, to measure moral conviction in political texts. I address the research question "Can GPT be used to identify moral conviction in political texts?" and the sub-question "is it useful to identify moral conviction in political texts?". I do this through three studies. The first tests GPT's ability to identify moral conviction in legislative speech data and open-ended survey response data. I find that GPT is as reliable as a human coder at these tasks. Next, I conduct a survey experiment to test GPT's labels against wider perceptions of moral conviction and respondents' emotional response to moralisation. I find that GPT's labels align well with respondents' views on what texts express moral conviction, and that these texts provoke a stronger emotional reaction than non-moral texts. Finally, I replicate Jung's (2020) study on the effect of moralisation on mobilisation with a GPT-based metric, finding that moral conviction suppresses turnout in low-education voters. Across the three studies I find that GPT demonstrates human-like behaviour when coding for a subjective concept like moral conviction, suggesting it can be used in a manner akin to a human labeller.

This paper offers a novel contribution in several areas. Methodologically, it adds to a growing literature on the uses of LLMs in political science research (e.g. Argyle et al., 2023; Chiu et al., 2022; Mellon et al., 2022; Ornstein et al., 2023). It also contributes to methodological literature on measuring morality in text (Araque et al., 2020; Garten et al., 2017; Hoover et al., 2019; Lin et al., 2017), as one of few papers that attempts to measure moral concepts in text without moral foundations theory. Finally, my survey findings, demonstrating the impact of moralisation in parliament, have implications for substantive

work on morality and rhetoric in politics (Clifford & Jerit, 2013; Clifford & Simas, 2022; Lipsitz, 2018).

i.1 Morality in rhetoric and attitudes

In politics, moral conviction has a discursive and psychological component. It is invoked in political rhetoric (a politician may or may not choose to frame housing as a matter of right and wrong) and is also an attitudinal dimension (a voter may or may not consider housing to be a matter of right and wrong). The utility of measuring moral convictions as an attitudinal dimension in individuals is well established. Individuals who hold strong moral convictions on an issue are less likely to be open to compromise (Ryan, 2017), less likely to listen to new information about an issue (Ryan, 2019) and more likely to be hostile to those who do not share their views (Garrett & Bankert, 2020). At the same time, moral conviction is predictive of activism behaviour (Sabucedo et al., 2018) and other forms of participation, a relationship mediated via emotional response (Skitka & Wisneski, 2011).

Generally, research focusing on the effects of moral rhetoric tends to focus on the differential effects of stressing various moral values, rather than the effects of moral conviction. For example, Clifford et al. (2015) show that when politicians express moral values in line with voters' it reinforces the link between voters' moral values and political attitudes, while Feinberg and Willer (Feinberg & Willer, 2015, 2019) demonstrate that politicians can effectively persuade voters from rival parties by emphasising the correct moral values. However, a small literature is emerging which explores the effects of moral conviction in political rhetoric. Simonsen and Bonikowski (2022) look at moralised messaging on immigration, finding that it contributes to affective polarisation, while Jung (2020) finds that moralised rhetoric mobilises educated co-partisan voters.

i.2 Measuring morality in text

i.2.1 Using moral foundations to measure conviction

Whether we are looking to measure moral conviction as an attitude dimension or a rhetorical tool, there are clear advantages to being able to extract this information from text rather than relying on surveys or vignettes.[4] Despite this, there is no established framework for measuring moral conviction in text, though tools exist to measure moral values in text. Researchers have adapted moral value measurement tools to try and capture moral conviction, but this has several weaknesses.

The Moral Foundations Dictionary (MFD, Graham & Haidt, 2012) is the most common way that researchers measure moral values in text. The MFD is based on moral foundations theory, a popular framework in social psychology that aims to map the contours of morality according to six 'foundations', each with a virtue and a vice, that capture the spectrum of morality across individuals, cultures and groups: care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, sanctity/degradation, and liberty/oppression. The MFD consists of lemmas corresponding to each vice/virtue of the six foundations and texts are usually scored in each foundation according to the number of matches with the dictionary. First developed in 2009 to analyse differences between liberal and conservative religious sermons (Graham et al., 2009), political scientists have used the MFD to identify moral foundations in political adverts (Lipsitz, 2018), media (Clifford et al., 2015; Clifford & Jerit, 2013) and survey responses (Kraft, 2018). In the methodological literature, there have been a slew of developments since the original dictionary that improve performance and better capture the foundations. For example, Garten et al. (2017) use pre-trained distributed representations over the dictionary, Lin et al., (2017) use entity-recognition tools to obtain background knowledge, and Araque et al., (2020) extend the MFD using WordNet synsets.

The MFD (and methods which build on it) is designed—like MFT generally—to characterise morality, not identify moral conviction, but it has been used for this purpose. Jung (2020) uses the MFD to construct a binary measure of moralisation, coding sentences in political manifestos as 'moral' if they contain any of the words from the MFD, and non-moral

[4] Measuring politicians' moralising rhetoric in text allows us to access its potential effects in an observational context. As an attitudinal dimension, social media text as well as open-ended survey responses have been shown to be a useful window into citizens attitudes (Amaya et al., 2020; Kayes et al., 2020; Kraft, 2018; Zollinger, 2022)

if they do not, to measure "whether and how much [parties] frame their positions as fundamental, moral beliefs about right and wrong" (p. 341).

Using the MFD to code moral conviction like this presents a few problems. The bag-of-words nature of dictionary methods cannot capture syntactic and semantic structures of text, which can be pivotal in expressing complex, nuanced concepts like morality, as Kraft and Klemmensen (2023) find. They show that semantic context is crucial in identifying the meaning politicians attach to moral terms, arguing that dictionary-based methods fail because this meaning is more important than the use of moral terms itself.

This shortcoming is a larger problem in contexts where moral sentiment is rare (such as parliamentary speech), as in such cases as in such cases the risk of misinterpreting or overlooking subtle expressions of morality is heightened. The standard criticisms of dictionary methods[5] also become a larger problem in these contexts, because even minor misclassifications will significantly impact the signal-to-noise ratio. Indeed, in their paper introducing the MFD, Haidt and Graham had initially wanted to analyse Democrat and Republican convention speeches but found that "those speeches were so full of policy proposals, and of moral appeals to the political center of the country, that extracting distinctive moral content was unfeasible." (Graham et al., 2009, p. 1038).

i.2.2 Measuring moral conviction with large language models

Large language models, often referred to as LLMs, are a recent development in natural language processing (NLP). In this paper I use models from OpenAI's Generative Pre-trained Transformer (GPT) suite, namely GPT-3 and GPT-3.5, which are among the most well-known and widely researched LLMs. GPT-3 has 175 billion parameters and was trained on more than 45 terabytes of text (Zhang & Li, 2021). It can understand complex semantic and syntactic meaning in text and compose writing that is difficult to distinguish from human output (Argyle et al., 2023). In a review of text-of-data methods in political science, Grimmer and Stewart (2013) write that the fundamental problem facing researchers is that we cannot read every text we wish to label, analyse or understand, but "the complexity of language implies that automated content analysis methods will never replace careful and close reading of texts." (2013, p. 268). With LLMs, this may no longer be the case.

[5] Dictionaries cannot account for negation or context, meaning they are prone to misclassification or over-generalisation (Grimmer & Stewart, 2013).

To understand why, it is worth briefly explaining how GPT and similar LLMs grasp and generate intricate linguistic structures and semantics. Unlike traditional text-as-data methods that use word frequencies, GPT employs 'word embeddings'—multi-dimensional vectors representing words based on co-occurrence with other words. While earlier word embedding models like word2vec produced fixed vectors, GPT's attention mechanism generates contextualised embeddings, adjusting word vectors based on surrounding context. This allows GPT to discern meanings, like distinguishing "prime" in "prime minister" differently from in "prime example," enabling GPT to interpret and generate text with depth and adaptability previously unseen in automated systems.

i.2.2.1 Applications of LLMs

LLMs have been shown to perform well on academic tests like the GRE or SAT (OpenAI, 2023) as well as baseline NLP tasks (Brown et al., 2020) and even tests designed to probe Theory of Mind capabilities (Kosinski, 2023; Shapira et al., 2023). When it comes to moral conviction, most relevant is their ability to identify complex psychological constructs and rhetorical devices in political texts. Ornstein et al. (2023) test GPT-3's ability on a wide variety of text-as-data political tasks, finding that GPT outperforms supervised learning approaches in all cases. Research has also found that LLM-based emotion detection in text outperforms other approaches (Acheampong et al., 2021; Del Arco et al., 2022), that GPT is able to detect hate speech with a high degree of accuracy (Chiu et al., 2022) and that it is accurate at classifying topics in open-ended survey responses (Mellon et al., 2022). As such, there are theoretical and practical reasons to expect GPT to be adept at identifying moralisation in political text.

There are some potential dangers of using LLMs to identify moralisation. Schramowski et al (2022) show that while GPT-based models can demonstrate biased or "degenerated" behaviour due to the unfiltered nature of training data. In a similar vein, research has demonstrated that GPT effectively has its own set of values and personality traits (Li et al., 2023; Miotto et al., 2022). These values and traits could be imposed on the identification process, creating an implicit bias.

Study 1: using GPT to measure moral conviction

To test GPT's ability to identify moral conviction in political texts I perform a series of classification and labelling experiments across four novel corpora from the UK and US containing either legislative speech or open-ended survey responses. By testing these two different domains, I assess the capabilities of GPT at identifying moral conviction in politicians' rhetoric (via legislative speech) as well as in an attitudinal context (via open-ended survey responses) in comparative context.

No existing text dataset contains labels for moral conviction, so I obtain human labels across the four corpora to evaluate GPT's performance.[6] I test four different prompts across two different GPT models. In addition to labels from human annotators, I also assess GPT's performance against two baseline MFD-based binary classification methods.

1.1 Labelling subjective concepts

Concepts like moral conviction are inherently subjective and contested (Skitka, 2010), meaning labelling text for moral conviction is not a classification task with an objective ground truth. Contrast it with a task like labelling x-rays for broken bones, where there is an objective truth (the bone is either broken in real life or not), and expert annotators (trained medical professionals) are better at accessing this truth. Moral conviction, however, is inherently subjective. When annotators disagree neither is necessarily *wrong,* instead, each annotator's perspective can be a valid interpretation.[7] In these kind of scenarios crowdsourcing labels is often preferred, because they provide a better approximation of the range of interpretations in the broader population (Aroyo et al., 2019; Reidsma & op den Akker, 2008), but I do not have the resources to crowdsource labels. Instead, in study 2 I test GPT's labels against crowdsourced evaluations.

Though it is subjective, I nevertheless need a minimal definition of moral conviction to guide labellers. I borrow one from the social psychology literature: "moral conviction refers to a strong and absolute belief that something is right or wrong, moral or immoral" (Skitka et al., 2005). From this, researchers usually measure the attitudinal side of moral conviction with

[6] The only existing labelled text dataset for morality is the Moral Foundations Twitter Corpus (Hoover et al., 2019) which contains tweets from six corpora labelled according to the moral foundations.

[7] See Potter and Levine-Donnerstein (1999) for a discussion of subjective concepts in content analysis.

a survey item along the lines of "How much are your feelings about x based on fundamental questions of right and wrong" (Skitka et al., 2021). To measure it in text, we can ask whether a text expresses 'a belief that something is right or wrong' in this way.

I use a binary classification schema (moral vs non-moral) rather than a scale or rating. I experimented with asking GPT to provide a confidence score for its classification, but this was unfruitful, as it provided only 0.8 or 0.9 every time. A rating scale might be preferred, but the binary schema provides a suitable initial test.

1.2 Data and pre-processing

For the experiments I build four corpora. Two use legislative speech data (from the UK House of Commons and United States Congress), and two use open-ended survey responses (from the British Election Study and American National Election Studies). I select texts according to specific criteria for each corpus, explained below, before randomly sampling 250 texts to undergo human and GPT annotation. Details on the nature of these 1000 texts can be found in A.1.

1.2.1 Legislative speech data

Within the legislative data I select speeches from high-profile debates and label them at the sentence level, as sentences typically cover only one topic, so are more amenable to a binary classification than a whole speech.[8] I use speeches from high-profile debates because I expect them to have a higher prevalence of moralisation[9] making it easier to evaluate GPT's performance.

1.2.1.1 UK House of Commons

To test GPT on speech data from the House of Commons, I build a dataset of speeches from the period 2001-2019 with metadata from Osnabrugge et al. (2021) and text data collated from the Hansard record by Odell (2021). I limit my analysis to speeches from Prime Ministers' Questions (PMQs), a weekly high-profile debate. In total, there are 121,788 sentences from 34,030 PMQs speeches. I sample 250 sentences at random for study 1.

[8] See A.2 for details on how I split texts into sentences and further detail on pre-processing for all corpora.

[9] Osnabrugge et al. (2021) finds that politicians use more emotive rhetoric in high-profile debates, so it seems likely they will use more moral rhetoric in them too.

1.2.1.2 US Congress

Speech data for the US Congress comes from the daily edition of the United States Congressional Record (Gentzkow et al., 2018). I use data from the 107th to 114th Congress (2001-2017), covering speeches from House of Representatives and the Senate. There is no equivalent to PMQs in the US, so to capture high-profile debates I use speeches from the State of the Union address or the day of the State of the Union address, giving 57,645 sentences from 6,130 speeches. I sample 250 sentences at random for study 1.

1.2.2 Open-ended survey data

In addition to legislative speech data, I test GPT on open-ended survey data. Specifically, I look at GPT's ability to classify responses to 'most important issue' (MII) questions. MII questions ask respondents along the lines of "What is the most important issue facing the country right now?". This question is well-suited for measuring attitudinal moral conviction.

One issue with open-ended MII question is the variable length of responses. Many responses are single-word answers. I restrict my analysis to answers that are over 30 characters long to give room for respondents to justify their view.[10] Because MII answers are typically short and focussed on one issue, I do not split them into sentences, instead labelling the entire response.

1.2.2.1 British Election Study

The British Election Study (BES) is a panel study with 23 waves from 2014-2022. There are a total of 727,614 observations from 107,796 individuals, with 43,864 observations from 23,395 individuals giving MII responses over 30 characters long. I sample 250 responses at random for study 1.

1.2.2.2 American National Election Study

The American National Election Study is a time series survey that has run since 1948. Open-ended MII responses are available in data from 2008, 2012, 2016 and 2020. In each of these, respondents were asked for their first, second and third most important issue, giving a total of 46,284 observations from 16,441 individuals. I sample 250 responses at random for study 1.

[10] There is likely self-selection into providing a long survey response, so it would be unwise to generalise trends from the 30+ character sample to the entire population, but that should not be a major problem in a labelling task.

1.3 Human labelling procedure

Each text was labelled by me and two other annotators. In total there were six annotators. Before undertaking the labelling task, annotators were shown a series of discriminatory examples and a coding guide for moral conviction that I built (see appendix A.3). I include more details about annotators including recruitment and profiles in appendix A.4.

1.3.1 Human labelling results

Annotators labelled the 250 texts in each test corpus, aside from the ANES corpus where 7 MII responses were excluded because they were in Spanish. Table 1 shows the results of the annotation procedure. The legislative corpora show a much lower prevalence of moral texts than the MII corpora. This is consistent with expectations—most of the content of parliamentary speeches is not moral grandstanding but policy details and procedural matters, whereas moral conviction in an attitudinal context is more common.

Table 1: annotation results

	HoC			USC			BES			ANES		
Annotator	1	2	3	1	5	6	1	3	4	1	2	5
Moral	22	54	20	25	18	11	40	54	51	34	90	43
Non-moral	228	196	230	225	232	239	210	196	199	209	153	200
% moral	9%	22%	8%	10%	7%	4%	16%	22%	20%	14%	37%	18%

Table 2: IAA statistics

	PABAK	Cohen's Kappa
HoC	0.72	0.37
USC	0.83	0.40
BES	0.56	0.30
ANES	0.58	0.44

Note: mean score across all annotators. For full matrices of agreement see A.5

To measure inter-annotator agreement (IAA) I report the prevalence and bias-adjusted Kappa statistic (PABAK). The Kappa statistic measures the agreement between two annotators who each classify items into mutually exclusive categories. PABAK adjusts the Kappa statistic to account for imbalances in the prevalence of categories and the bias of annotators, making it a useful metric for imbalanced and subjective data like this (Sim & Wright, 2005).

In Table 2 I include the mean score across all three annotators for each corpus; for full matrices of agreement see A.5. Overall, using the accepted standards for Kappa from Landis and Koch (1977), IAA scores between individual annotators range from 0.44 (*moderate*) to 0.91 (*almost perfect*). These are comparable to the level of IAA found in annotated datasets for moral foundations (Hoover et al., 2019).

1.4 Prompt design, few-shot learning and GPT models

To test the performance of GPT, I pass each text to the GPT API as part of a prompt. GPT only sees one text at a time and does not remember previous texts. I test four different prompts on each corpus, each with two different models, giving a total of 32 different configurations. I test two different GPT models: GPT-3.5-turbo and GPT-3. In all cases I provide GPT with two examples in the prompt, one for each class. This is called 'few-shot learning' (as opposed to 'zero-shot learning' when no examples are provided) and typically yields better results (Brown et al., 2020).

1.4.1 GPT prompt design

Users interface with LLMs through prompts: passages of text that the model 'reads' and responds to. Good prompt design is key for getting good outputs from an LLM: there is a growing literature on prompt design (e.g. Xu et al., 2023; Zamfirescu-Pereira et al., 2023) and it can radically effect the quality of responses. I test two models – GPT-3.5-turbo and GPT-3 – because each use different prompt formats. GPT-3.5-turbo uses the 'chat' framework. In this framework, the user provides GPT with a 'system' prompt to define its role, before passing examples of user prompts and system responses. GPT-3 uses the 'completions' framework, in which the user provides one prompt and the system responds. Figures 1 and 2 demonstrate these structures. I change the examples for each corpus to examples that correspond more closely to the text content for that corpus. All prompts are in appendix A.6.

Figure 1: sample GPT-3.5 chat prompt (researcher prompt, ANES)

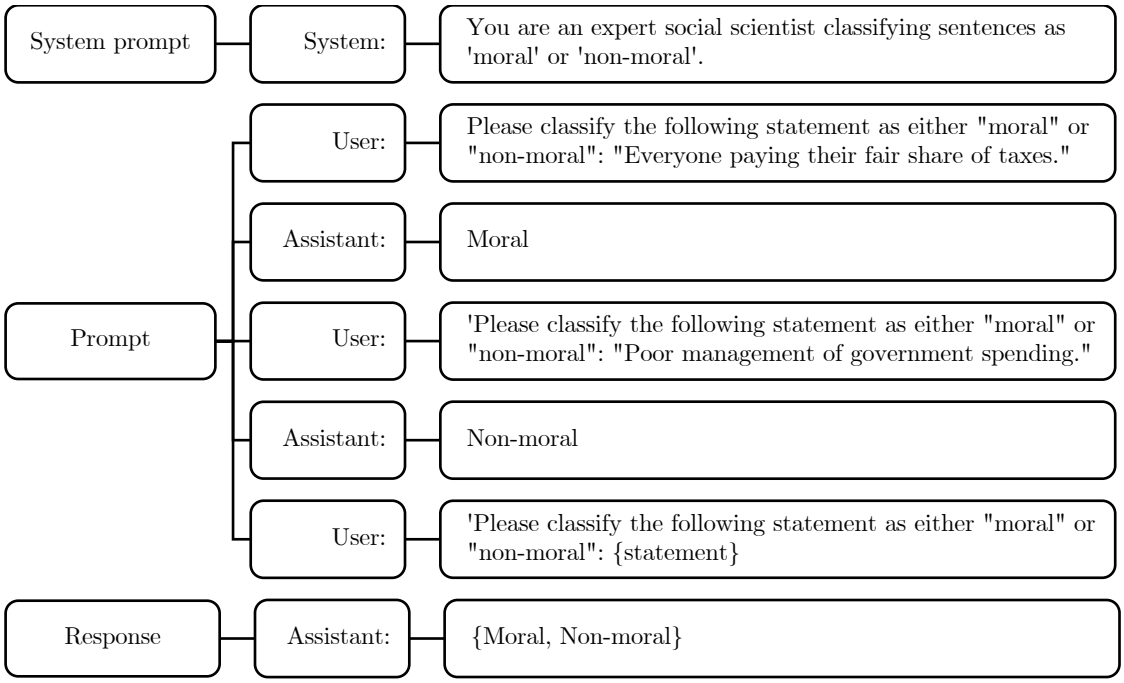
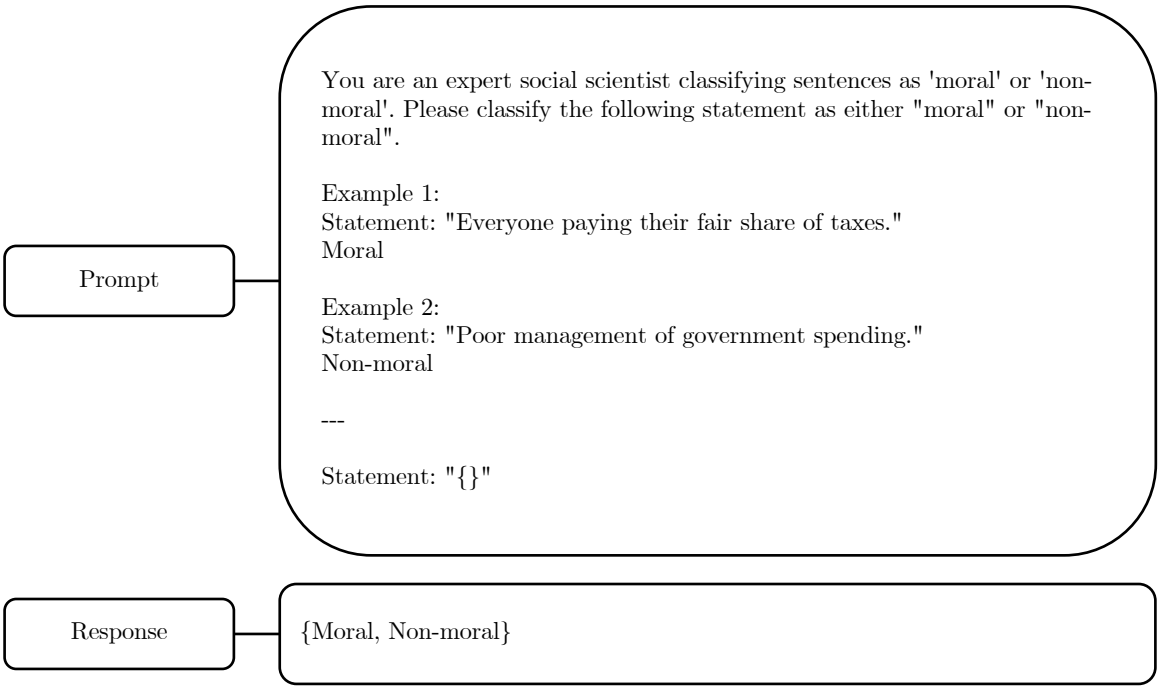


Figure 2: GPT-3 completion prompt (researcher prompt, ANES)



I test four different prompts to test whether GPT's performance can be enhanced by combining domain expertise with clear task-specific directives (Table 3). In researcher prompts, I tell GPT that it is "an expert social scientist". While labelling for moral conviction is not a task that benefits from expertise (see 1.1 for a discussion of this), the goal is to determine if framing the model as an expert influences its performance, following research showing that prompting GPT to 'act as an expert' has been shown to improve performance in various tasks (Niszczota & Abbas, 2023; Xu et al., 2023). The guidance prompts provide explicit instructions about what constitutes a moral statement to assess if making the definition of moral conviction explicit affects GPT's ability to label sentences. Combining both the researcher and guidance elements in the R+G prompt aims to maximise the potential benefits of both role assignment and explicit instruction. I vary the system prompt in the chat framework, and the first sentence in the completion prompts.

Table 3: GPT prompt variations

	System prompt
Minimal (M)	You are an assistant that classifies statements as 'moral' or 'non-moral.
Researcher (R)	You are an expert social scientist classifying sentences as 'moral' or 'non-moral'.
Guidance (G)	You are an assistant that classifies statements as 'moral' or 'non-moral'. A sentence should be classified as moral if it expresses a belief that something is right or wrong.
Researcher and guidance (R+G)	You are an expert social scientist classifying sentences as 'moral' or 'non-moral'. A sentence should be classified as moral if it expresses a belief that something is right or wrong.

1.5 Baseline metrics

To assess the performance of GPT-generated labels I treat it first as a classification task, reporting F1 scores, precision and recall, using human labels as the 'true' value for each text.[11] I use the mean scores across GPT and each human annotator rather than construct a single ground truth standard for each text because of the subjective nature of moral conviction (see 1.2).

I also report PABAK scores between GPT and human annotators, treating the task as a labelling task. This is how GPT is often used in practice—to label data in place of human annotators—because of its low cost and ability to simulate human-like behaviour (see i.2.2). These GPT-generated labels can then be used to train supervised classification models (Wang et al. 2021). PABAK scores, usually used to measure IAA, are a useful way to evaluate GPT's agreement with human labellers in this way.

To compare GPT against a baseline, I report the same scores for a simple MFD classifier, as in Jung (2020), and a similar classifier using MoralStrength[12] (Araque et al., 2020). The baseline MFD measurement codes a text as 'moral' if any of the terms from Jung's (2020) modified moral foundation dictionary[13] are found in the text, and 'non-moral' if not. For the MoralStrength classifier, I rated each text using the 'unigram+moral stat' model as described in Araque et al. (2020). The model scores each text on each of the six moral foundations and also on 'non-moral'. I code texts as 'moral' if the highest rating is any of the moral foundations, and 'non-moral' if it is 'non-moral'. I rank all the models according to the Friedman statistical test rank (Demšar, 2006), and test if results are statistically significant over the baseline MFD measure using McNemar's test (Pembury Smith & Ruxton, 2020).

[11] Precision is the fraction of predicted positives that are true positives (i.e., labelled as moral by the human labeller). Recall is the fraction of true positives correctly identified as positive. The F1 score is the harmonic mean of precision and recall, balancing their trade-off, and is valid for uneven class distributions (Lin et al., 2017).

[12] MoralStrength is Python package that contains pre-trained models to identify moral foundations in text based on an augmented version of the MFD. The authors show it outperforms the MFD at identifying moral foundations, so it is a useful benchmark here.

[13] Jung (2020) modifies the MFD to make it more suited for political texts. Her alterations are minor, so I use her version over the original MFD.

1.6 Results and discussion

Table 4 shows F1 scores for each GPT model/prompt, as well as baseline scores for the MFD and MoralStrength classification approaches. The best performing model for each corpus is in bold. For the ANES, House of Commons and US Congress corpora GPT outperforms the MFD and MoralStrength approaches. For the BES, the MFD and MS both outperform all GPT methods. Overall, the highest ranked model/prompt combination is the researcher+guidance prompt using the GPT-3 completion endpoint. The highest average F1 score is the researcher+guidance prompt using the GPT-3.5-turbo chat endpoint. This difference is statistically significant at the 0.05 level in all cases.

These F1 scores are in the same range as best-practice machine learning methods for classifying text according to the six moral foundations (e.g. Garten et al., 2017). This may initially seem disappointing, as we would expect GPT to be able to improve over these approaches. However, comparing these F1 scores to the F1 scores between each human labeller (Table 5) shows that in every case other than the BES corpus, the best performing GPT model is close to the level of performance consistency between labellers. This suggests that as a classifier, GPT is as effective as a minimally trained human. Likewise, the PABAK scores between GPT and human labellers in Table 6 are similar to the PABAK scores between human labellers from Table 2. This shows that GPT agrees with human labellers as much as human labellers agree with each other.

Table 4: average F1 scores between GPT configurations and human labels

	Chat model (GPT-3.5)				Completion model (GPT-3)				Non-GPT	
	M	G	R	R+G	M	G	R	R+G	MFD	MS
BES	0.17	0.29	0.25	0.25	0.27	0.28	0.24	0.28	**0.31**	0.31
ANES	0.41	0.49	0.47	**0.52***	0.47	0.48	0.44	0.50	0.45	0.42
HoC	0.35	0.35	0.32	0.30	0.38	0.40	0.35	**0.41***	0.23	0.22
USC	0.51	0.43	0.51	**0.55***	0.31	0.33	0.28	0.29	0.16	0.10
Average	0.36	0.39	0.39	**0.41**	0.36	0.37	0.33	0.37	0.29	0.26
Rank	6.9	3.8	5.6	4.4	5.1	3.9	7.5	**3.6**	6.6	7.6

Note: M: minimal prompt; G: guidance prompt; R: researcher prompt; R+G researcher and guidance prompt; MFD: Moral Foundations Dictionary following procedure in Jung (2020); MS: MoralStrength using unigram + moral stats (Araque et al. 2020). Results are the average F1 scores across GPT and each annotator. 'Moral' is the positive class. * indicates statistically significant performance over MFD baseline according to McNemar's test ($p<0.05$).

Table 5: average F1 scores between human labellers

	F1 score
BES	0.41
ANES	0.56
HoC	0.44
USC	0.45

Note: these F1 scores do not involve GPT—they are an average of all the possible combinations of annotator—the average F1 score between annotator 1 and annotator 2, annotator 2 and annotator 3 and so on. This is given by the formula $\frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} F1(a_i,a_j)}{\binom{n}{2}}$ where $F1(a_i, a_j)$ is the F1 score between annotator a_i and a_j.

Table 6: PABAK score for different GPT prompts

	Chat model (GPT-3.5)				Completion model (GPT-3)				Non-GPT	
	M	G	R	R+G	M	G	R	R+G	MFD	MS
BES	**0.52**	0.39	0.48	0.48	0.14	0.00	-0.14	-0.12	0.43	0.06
ANES	0.47	0.35	**0.56**	0.49	0.41	0.29	0.32	0.29	0.30	-0.18
HoC	0.76	0.72	**0.76**	0.74	0.71	0.71	0.59	0.66	0.19	-0.7
USC	0.85	0.77	**0.86**	0.86	0.68	0.67	0.51	0.61	0.14	-0.75

Note: M: minimal prompt; G: guidance prompt; R: researcher prompt; R+G researcher and guidance prompt; MFD: Moral Foundations Dictionary following procedure in Jung (2020); MS: MoralStrength using unigram + moral stats (Araque et al. 2020). Results are the average PABAK scores across GPT and each annotator. A score of 1 would indicate perfect agreement, -1 perfect disagreement, and 0 agreement no better than chance.

1.6.1 GPT labelling dynamics

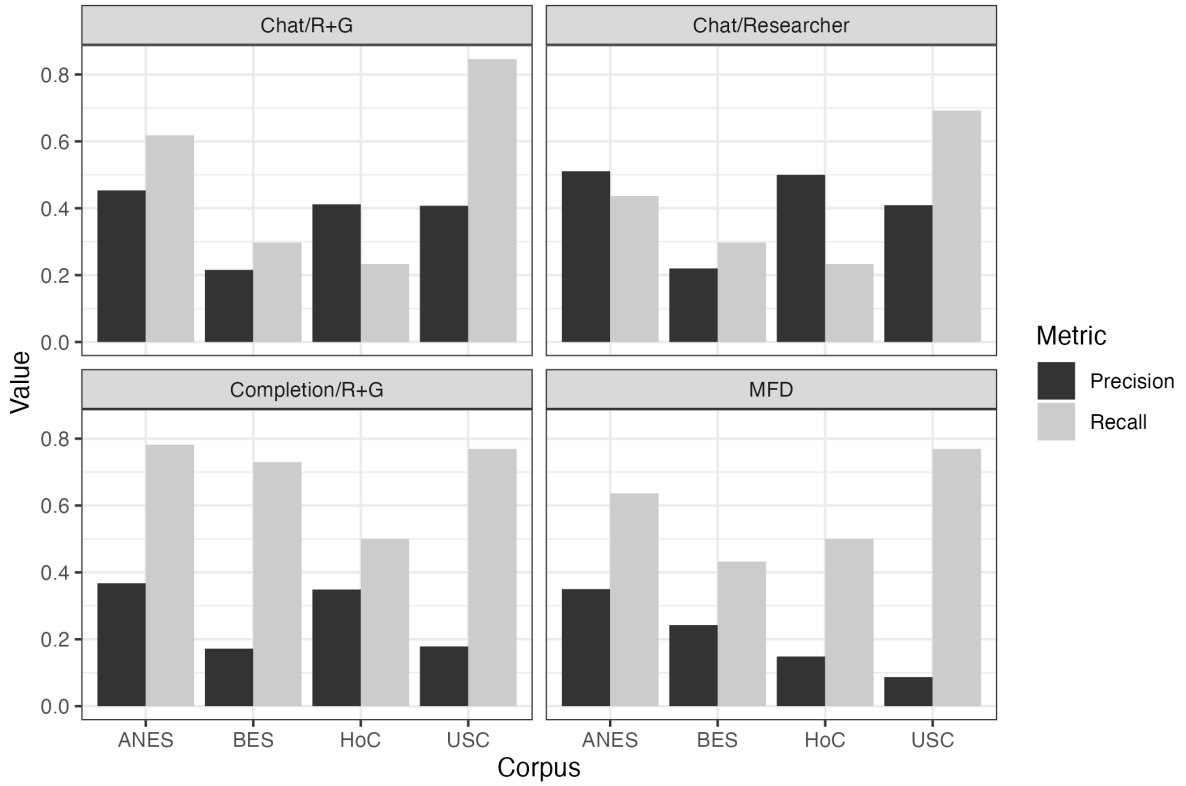
The similar F1 scores between different GPT models conceal different approaches to labelling. Handling low prevalence is a particular advantage of the chat-based models (Table 7). Chat models label statements as 'moral' at a very similar rate to human labellers, whereas the completion models do so at a much higher rate. This is reflected in the recall and precision of the models (Figure 3), with the chat models showing improved precision over the completion models but worse recall. Baseline approaches are less discriminatory overall in their labelling, with the MFD approach showing consistently low precision throughout.

Table 7: positive class proportion by labelling method

	Human annotators	Chat models	Completion models	MFD	MS
BES/ANES	18.7%	24.5%	48.3%	33.8%	66.3%
HoC/USC	8.6%	9.1%	18.6%	43.2%	94.8%

Note: positive class is 'moral'.

Figure 3: precision and recall for select models



Note: average precision and recall between human labellers and GPT.

To assess when and why GPT disagrees with human labellers, I looked at the relationship between mislabelled texts and covariates, as well as a simple MFD score that counts how many MFD terms are in each text. For the sake of simplicity, I define as mislabelled texts here as when GPT disagrees with the majority vote across the three annotators.

I ran a logistic regression between a binary indicator of labelling error and each covariate. Table 8 shows which covariates have a statistically significant effect on the likelihood of GPT mislabelling a text. Across the ANES, HoC and USC corpora a higher count of words from the MFD are positively associated with GPT mislabelling the text, with each extra word from the MFD in a text increasing the probability of an error by values ranging from 2.5% (for the USC corpus) to 5.5% (for the HoC corpus). In each of these three corpora, MFD count is more strongly associated with false positives than false negatives.[14] This could indicate an over-sensitivity or bias in GPT's classification mechanism when encountering dense moral terminology.[15] None of the other variables are statistically significant, aside from the BES wave variable, where responses from more recent waves are less likely to be misclassified by GPT.

Table 8: statistical significance and effect direction of covariates on likelihood of GPT labelling mistake

	BES	ANES	HoC	USC
Party	-	-	-	?
Date	✓(-)	-	-	-
MFD count	-	✓(+)	✓(+)	✓(+)
Gender	-	-	-	-

Note: ticks mean that the covariate has a statistically significant ($p<0.05$) impact on the likelihood of GPT making a labelling error. The sign indicates the effect direction. No data available for USC party. A.7 shows differential effects on false positives and false negatives.

Across the different corpora GPT models were consistently better on US-based than UK-based data, with performance on the BES corpus notably poor. The majority of GPT's training data comes from US-based sources (OpenAI, 2023), and previous research has shown GPT to

[14] See A.7 for results of a multinomial logistic model demonstrating this.

[15] This could be caused by the presence of MFD-related literature in GPT's training data, meaning GPT might have learned to associate MFD terms with morality directly. This is an issue in LLMs called 'data contamination', and often happens when tests used to evaluate LLMs are included in their training data (Magar & Schwartz, 2022).

have culturally specific conceptions of morality and other socio-psychological constructs (Schramowski et al., 2022). All human annotators were all British. This could explain this variation, as human annotators may have labelled texts as moral based on concepts which read as moral issues only to people from the UK. Indeed, inspecting the most common unigrams and bigrams in misclassifications of UK data (Table 9) for false negatives "health_service" is in the top five for parliamentary data and "EU" and "Brexit" are in the top five for BES data. These terms may have a moral valence to British people that is not obvious to GPT.

Table 9: five most common uni/bigrams in GPT misclassifications

	False positive	False negative
BES	People, government, wealth, vulnerable, poor	Government, getting, immigration, eu, brexit
ANES	Lack, greed, healthcare, people, mental_health	People, healthcare, jobs, need, enough
HoC	Prime_minister, people, hon, numbers, work	People, minister, need, health_service, sure
USC	Willing, war, today, thousands, people	Can, americans, one, pay, legislation

Note: a misclassification here is based on human annotators' majority vote.

1.6.2 Qualitative error analysis

In addition to quantitative analysis, I conducted a qualitative analysis of mislabelled texts from the ANES and HoC corpora (as the highest performers from each country). To identify commonalities in the mislabelled data, I followed the thematic analysis procedure as described in Braun and Clarke (2006), closely reading and re-reading each mislabelled statement to identify commonalities between them, paying attention to semantic qualities like sentence structure or keyword presence and latent themes underpinning the meaning of the text. I include in A.8 the full results of the coding process.

I found four main themes in mislabelled texts across the two corpora. In false negatives, the most common theme was 'implicit morality' — texts which did not use overt moral language but referenced abstract concepts that imply moral conviction, for example, lack of transparency or loss of personal freedoms. False negatives also occurred in texts that referenced specific events or subjects without context ('lack of context'). For example, in "we are not free and i have no faith in our voting system. it was fixed" (Table 10) 'it' most likely refers to the 2020

election but could also refer to the voting system. Lack of context is an issue inherent to the survey corpora but could be mitigated in the legislative data by showing GPT entire speeches and asking for labels at the sentence level.

A lot of mislabelled statements (especially in the ANES corpus) used unusual sentence structure or word choice ('complexity and ambiguity'). In false positives, GPT often seemed to be misled by morally charged language ('overemphasis on keywords', e.g., 'pain' and 'hurt' in the HoC false positive in table 10). This is in keeping with the results from Table 8 that a higher count of words from the MFD is positively associated with classification errors. Importantly, all of these errors are of the kind a human labeller would make (Sandri et al., 2023), in keeping with GPT's quantitatively similar performance to human annotators.

Throughout the thematic analysis, I found that the distinction between moral conviction and its absence is much blurrier in legislative speech compared to survey responses. For example, in my view all the HoC examples in Table 10 could be viewed as expressing a moral conviction or not. In the case of the survey data, however, misclassifications are much more flagrant. The ANES texts in Table 10 illustrate this. The false positives do not express or even hint at a moral conviction, while the false negatives explicitly state moral problems.[16] This shows the increased difficulty GPT had labelling the survey data. GPT appears to be better suited for the vague distinction between moral and non-moral in legislative speech.

[16] Of course, we might say that at a second order, invoking climate change is *always* moral because climate change is inherently a matter of right and wrong. I take the view of Skitka (2010) that no issue is inherently moral, so prefer to rely on invocations of abstract concepts linked to morality (e.g. freedom, injustice, guilt—all present in the false negative column).

Table 10: select false positive and negatives from the HoC and ANES corpora

	False positive	False negative
HoC	"It is important that risk be seen to lie with the banks and the lenders and not be underwritten by the taxpayer." (lack of context)	"The Prime Minister has also said that too many of the guilty are going free." (implicit morality)
	"The pain that has been caused to the Windrush generation needs to be resolved very rapidly, with full compensation paid as quickly as it can possibly be done and an understanding of the hurt that they feel." (Overemphasis on keywords)	"Increasing numbers of people in the west country feel that the freedoms for which they fought have been eroded by the European Union and that their homes are no longer their castles." (lack of context)
ANES	"stupidity, americans don't use brains" (lack of context)	"lack of transparency enabling corruption-lack of ethics-morals" (complexity and ambiguity)
	"climate change-- it needs to be talked about as much as possible to set an example for other countries and create new industries" (Overemphasis on keywords)	"we are not free and i have no faith in our voting system. it was fixed" (lack of context)

Note: theme from thematic analysis in brackets. See A.8 for more details.

1.7 Summary

This study shows that GPT can classify political texts for moral conviction as well as minimally trained human labellers. Alternative approaches cannot. The advantage of GPT over baseline methods is especially marked in the case of legislative speech data, where the low prevalence of the 'moral' class and the blurrier distinction between moral conviction and its absence make it particularly challenging for traditional approaches. These findings are in line with previous research on GPT's classification abilities (Ornstein et al., 2023; Wang et al., 2021).

Moral conviction is not an objective classification task with a strong ground truth, but as a labeller, GPT agrees with human labellers as much as human labellers agree with each other, supporting the use of GPT to label for moral conviction. In addition, my thematic analysis shows GPT is fallible in a human way. Its errors are similar to errors a human labeller might make, rather than the mechanistic errors classification models exhibit. On this view, we might conceive GPT as a cheap, fast, and reliable labeller (as in Wang et al., 2021). GPT could possibly then be used to crowdsource labels for moral conviction, in keeping with previous research that has found conditioning GPT on demographic backstories makes it able to mimic human survey samples accurately (Argyle et al., 2023). In appendix A.9 I provide supporting evidence for this by conditioning GPT on demographic backstories and finding that GPT's labelling behaviour varies when prompted to act as strong liberal vs a strong conservative. Future research should explore this potential further.

Study 2: testing potential misclassifications

To test whether GPT's labels align with human understandings of moral conviction in political text I design a survey experiment that shows respondents labelled parliamentary speeches from the House of Commons at random. I test respondent perceptions of moral conviction explicitly, with a direct survey question, and implicitly, by testing emotional reactions to the texts. I find that GPT's classifications reliably align with respondents' perceptions of moral conviction and provoke an emotional response in line with the emotional response recovered by vignettes in previous research (Jung, 2020; Lipsitz, 2018; Simonsen & Bonikowski, 2022). This supports the use of GPT to identify moral conviction in text.

2.1 Method

I show respondents real parliamentary speeches from the House of Commons corpus as described in Study 1. This allows me to test respondents' reactions to texts classified by GPT and directly demonstrate the validity of GPT's labels for observational research.

Instead of simply showing respondents random labelled texts, I use texts where GPT disagrees with a baseline classification method (i.e., GPT thinks the text shows moral conviction but a baseline method does not, or vice-versa). This offers a stronger test of GPT's labels because texts where the two approaches disagree are more likely to be misclassifications. As such, if these texts can recover an effect, it is strong evidence in favour of GPT's ability to identify moral conviction.[17]

2.1.1 Selecting parliamentary speeches

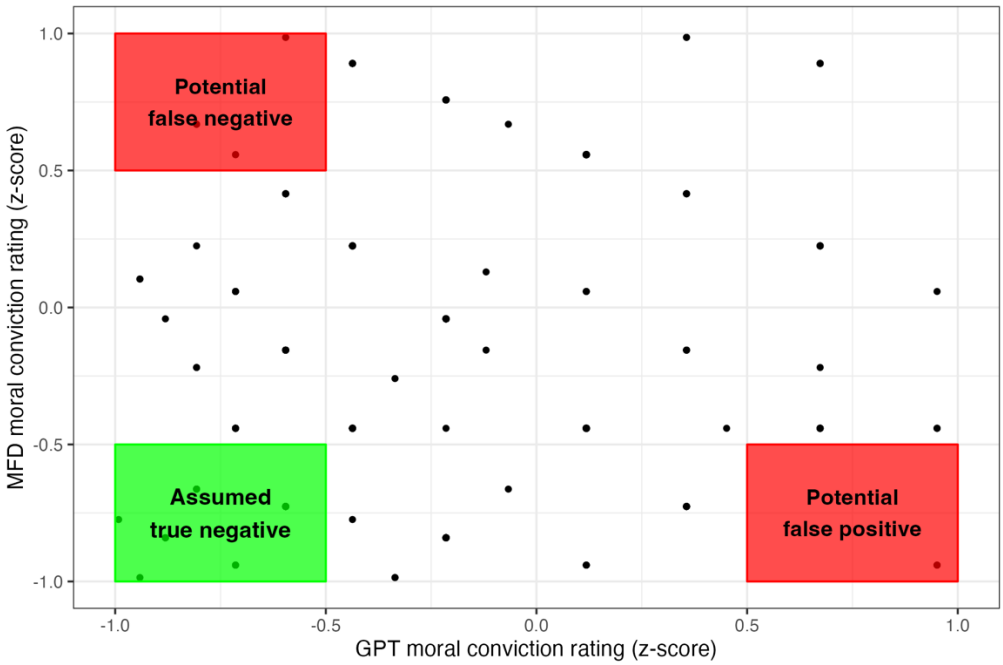
To select texts to show to respondents I label the complete House of Commons corpus from 2001-2019. This consists of 34,030 speeches. I use the HoC corpus because the data is cleaner than the US Congress data, making it more suitable for a survey experiment. I labelled every sentence in the speeches with two methods: GPT-3.5-turbo with the 'full guidance' prompt and the MFD approach described in section 1.5. I then calculated two 'moral conviction ratings' for each speech by calculating the proportion of sentences in that speech that were classified as moral by either GPT or the MFD. Scores range from 0 (no sentences classified as moral) to 1 (all sentences classified as moral). After calculating these ratings I restricted the

[17] If GPT provided confidence scores for its labels I would use texts with lower confidence scores, but as previously discussed this is not possible.

sample to speeches with four or more sentences (n = 6279) to remove procedural speeches. I then calculated z-scores for the GPT rating and MFD rating and calculated the difference between each speech's z-score. This identifies speeches where the MFD and GPT ratings disagree most strongly.

From these difference scores, I established three categories of texts: 'potential false positives' (PFP), 'potential false negatives' (PFN) and 'assumed true negatives' (ATNs), based on the level of agreement between the two methods. PFPs have a high GPT score and low MFD score. The logic here is that while GPT thinks the text shows moral conviction, the MFD method does not, suggesting GPT may be mistaken. If GPT is mistaken, then the text is a false positive. The inverse applies to PFNs. In the case of ATNs the two methods agree on a non-moral classification. I say 'potential' false positives and negatives because there are no human labels for these texts, so we do not know their true class. I used the 200 speeches with the biggest difference between GPT and MFD z-scores in each direction for the false negative and false positive categories. For the ATNs more than 200 speeches received 0 scores from both methods so I randomly sampled 200 of these speeches. Figure 4 illustrates these categories graphically. I do not include would-be 'assumed true positives' in the survey sample, only comparing potential false negatives/false positives from GPT and the MFD to assumed true negatives. This is due to limits I faced on funding given targets for sample size. I include information about speeches in each category in A.10.

Figure 4: categorisation of parliamentary speeches for survey experiment



Note: this figure is illustrative, speeches were categorised as explained above, not using the cutoffs denoted in this figure.

2.1.2 Survey design

Participants were recruited via survey platform Prolific, a crowdsourcing platform designed for academic research. I restricted my sample to participants from the UK. In total 219 participants completed the survey.

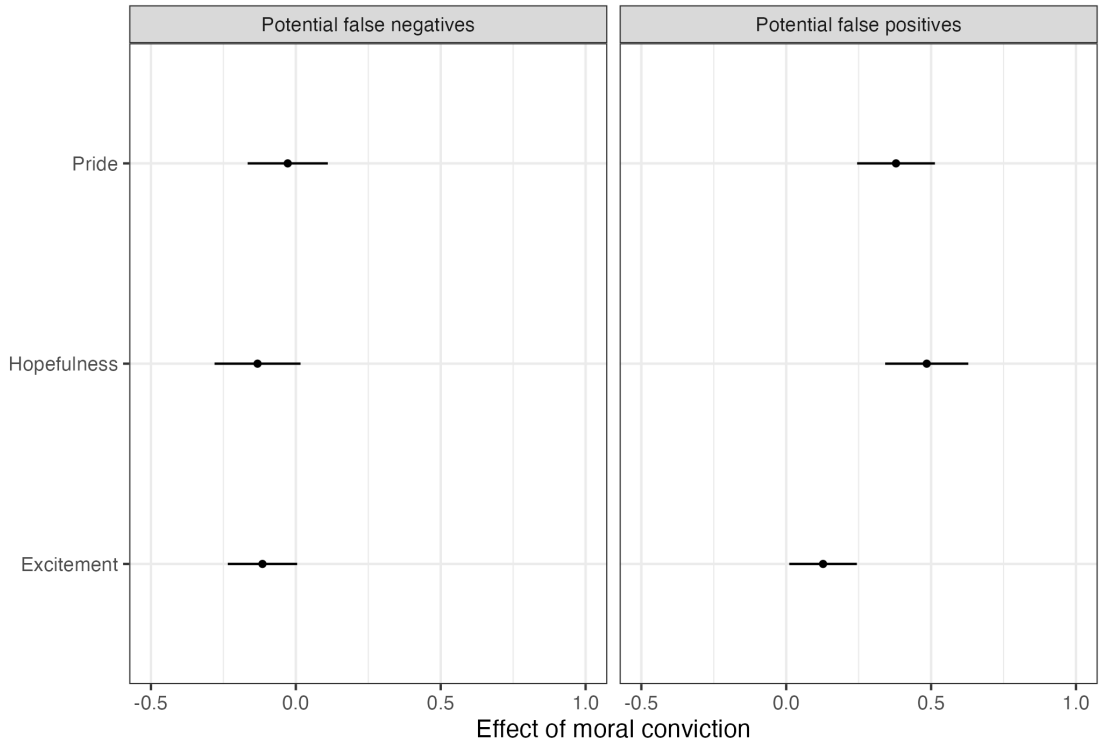
In the survey, participants were shown seven texts at random. Initially they were shown a random text from one of the three groups (PFP, PFN, ATN), before being shown six more random texts, two from each group. The group order was random.

After each text, respondents were asked two questions: first, "Consider the main issues being discussed in the text. To what extent do you think they are a matter of right and wrong, versus a matter of personal opinion?". This question is designed to measure the extent to which participants' perceptions of moral conviction agree with GPT. I juxtapose 'right and wrong' with 'personal opinion' in keeping with findings in social psychology that moral convictions contain a strong universalising component (Skitka 2021) and survey research showing respondents find it easier to answer judgement questions when given an alternative (Krosnick, 1999). Respondents rated the text on a scale of 1 'Entirely personal opinion' to 7 'Entirely right and wrong'. If GPT can identify moral conviction in text, we would expect higher scores for texts with a high GPT score (i.e., PFPs).

Second, respondents were asked "To what extent does the text make you feel proud / hopeful / excited / angry / disgusted?" to which they could respond "not at all," "slightly," "moderately," "much," and "very much." This design follows the format used in research on emotions (Valentino et al., 2011) and that used in Jung (2020). Previous research finds that moral vignettes provoke a stronger positive emotional response than non-moral vignettes (Jung, 2020; Lipsitz, 2018). If GPT can identify moral conviction in text, we would expect higher scores for texts with a high GPT score (PFPs).

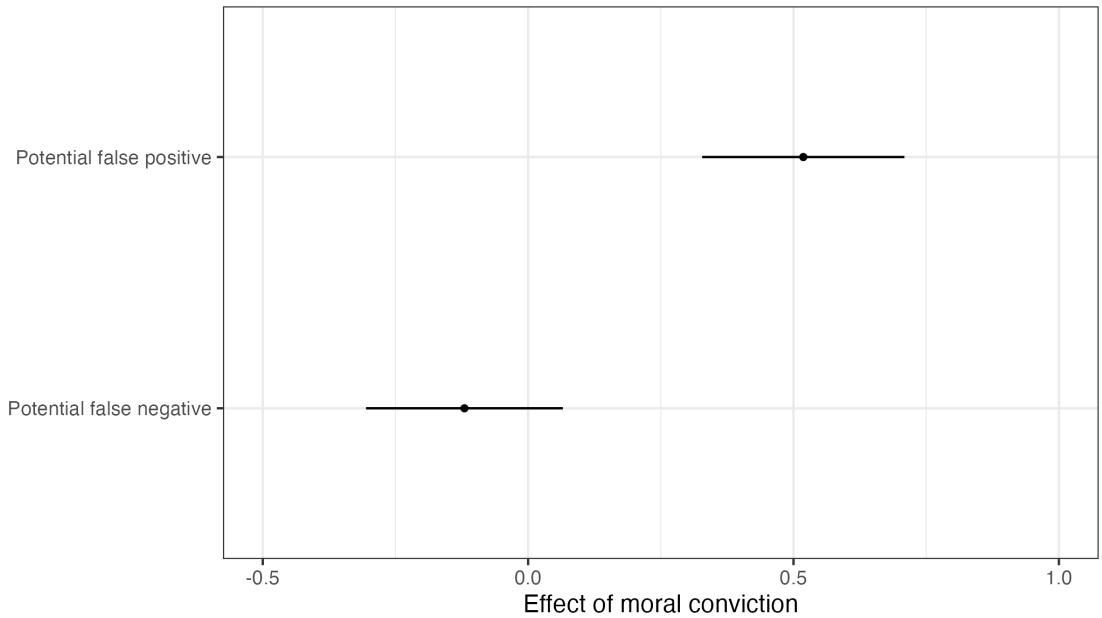
2.2 Results and discussion

Figure 5: the effect of moral conviction on positive emotions in parliamentary texts



Note: each point is the estimated difference in means between texts in the label category (potential false positives or potential false negatives) and assumed true negatives. Horizontal lines are 95% confidence intervals. Higher scores mean a stronger emotional reaction. The scale was from 1 to 5.

Figure 6: the effect of moral conviction in parliamentary texts on perceptions of whether a text is about right and wrong



Note: each point is the estimated difference in means between texts in the label category on the y-axis (potential false positives or potential false negatives) and assumed true negatives. Horizontal lines are 95% confidence intervals. Higher scores indicate a text is viewed as concerning right and wrong. The scale was from 1 to 7.

Figure 5 shows the effect of parliamentary speeches labelled as 'moral' on pride, hopefulness, and excitement with 95% confidence intervals. The results are the difference in means compared to true negatives. Potential false positives generate a positive emotional reaction—the expected response of a speech containing moral conviction (Jung, 2020; Lipsitz, 2018). This difference is statistically significant for all three positive emotions.[18] Potential false negatives do not provoke a different reaction than true negatives. Recall that potential false positives have a high GPT score and a low MFD score, with potential false negatives having the inverse. This suggests that GPT is identifying moral conviction in these texts.

PFP speeches are also more likely to be rated as concerning matters of right and wrong, whereas PFNs show no difference from true negatives (Figure 6). This demonstrates alignment in GPT with respondents' understanding of moral conviction. It also helps guard against the possibility that GPT is identifying something other than moral conviction in the texts which is generating the emotional response in the other questions.

One issue with using parliamentary speeches rather than vignettes is potential confounding. For example, one party's speeches may provoke a stronger emotional response, and the prevalence of speeches from this party may systematically vary across classifications. To test this I include in A.10 results of ordinal regression measuring the effect of PFPs on emotional response with controls for respondent left-right orientation, respondent party ID, respondent age, the party of the speaker giving the speech in parliament, and speech year. All results are robust to these tests.[19]

2.2.1 Overview

These results suggest that GPT is consistently capturing both moral conviction and its absence even when the MFD disagrees with its ratings. Compared to texts that the GPT and the MFD both classify as non-moral, texts rated as moral by GPT are consistently viewed as concerning right and wrong and provoke an emotional reaction consistent with moralisation in text. The inverse also holds: texts rated as non-moral by GPT are not considered a matter of right and wrong, and provoke an emotional reaction consistent with a lack of moralisation in text.

[18] Moral conviction has no effect on these negative emotions as expected (Jung, 2020)

[19] These tests do not guard against all possible confounding, as it remains possible that other unobserved speaker, speech or respondent-level characteristics are driving results. The other survey question helps protects against this.

These results support my earlier claim that GPT captures a version of moral conviction that qualitatively aligns with what people understand as issues of right and wrong. In addition, the effect magnitude for the emotional response questions is similar to that in Jung (2020). This does not speak to the actual accuracy of GPT as a classifier per se, but rather provides supporting evidence that it is measuring what we want to measure when we think about moral conviction, and that it is better at doing this than the MFD.

As well as supporting the use of GPT to measure moral conviction in legislative text, my results have substantive implications. Jung theorises that only co-partisan moralising rhetoric will provoke a positive emotional response, but in the appendix (A.10) I show the effect holds across party lines, which calls into question her causal mechanism for how moral conviction increases turnout. I explore this further in the next study. On a wider level, the emotional response prompted by moral conviction shows that it is a useful concept to measure, because emotional responses to politics has been linked to a host of important outcomes from polarisation (Osborne & Sibley, 2022) to participation (Valentino et al., 2011). While it goes beyond the scope of this paper, I include in A.11 information from 34,030 labelled HoC speeches showing moral conviction has been increasing in the House of Commons from 2001-2019, suggesting this is an area ripe for further research.

Study 3: moral conviction and mobilisation

This study tests GPT-generated labels in an observational setting. I do this by replicating the main study from The Mobilizing Effect of Parties' Moral Rhetoric (Jung 2020). As one of few papers that measures the effects of moral conviction using textual data, Jung's paper is a good test case.

Jung finds that moral conviction in political manifestos (taken as a proxy for moral conviction generally) has no effect on turnout on average but mobilises educated co-partisans, who Jung suggests are more exposed to rhetoric espousing moral conviction. Her theory is that moralisation "makes salient in the minds of voters that politics is a matter of moral right and wrong" (p. 5) which activates the emotions of voters, motivating participation. To operationalise moral conviction Jung uses a modified version of the Moral Foundation Dictionary, calculating for each manifesto the proportion of sentences containing one or more matching terms.

I discuss issues with the MFD used this way in section i.2.1. In study 2 I show that GPT's understanding of moral conviction aligns with survey respondents more closely than the MFD method, and many texts that the MFD labels as moral provoke no emotional reaction, while many texts that the MFD labels as non-moral do provoke a reaction. Jung's causal mechanism for increased mobilisation relies on the emotional response moral conviction provokes, so I expect replicating her findings with the GPT-based measure to produce stronger results. At the same time, it may complicate her findings: Jung's causal mechanism relies on the positive effects of specifically co-partisan moral rhetoric, whereas I find in study 2 that these effects hold across party lines (see A.10).

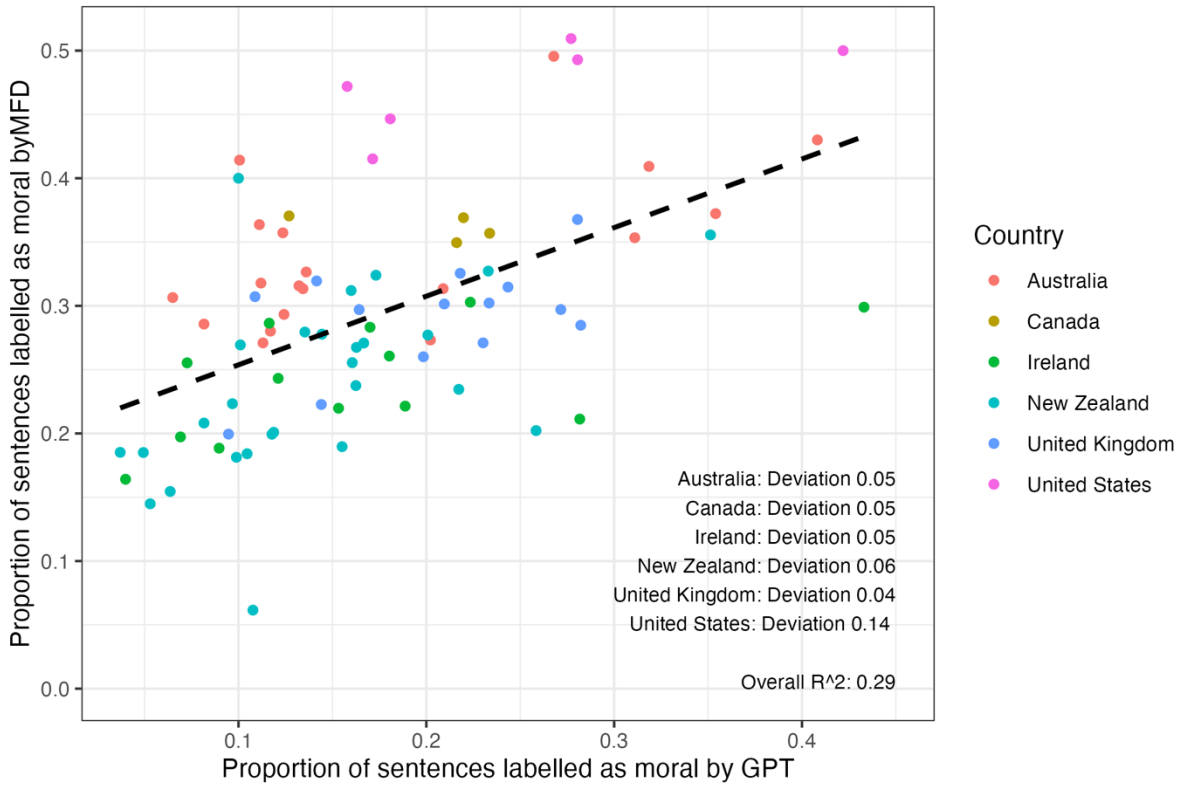
Below I briefly lay out information on the data sources from Jung, compare my labelling results to those obtained by her method, and then replicate the findings from her first study using GPT labels. I find that using GPT labels increases the magnitude of the main effect she reports but shifts its incidence across the interaction with education, resulting in significantly different implications.

3.1 Research design

3.1.1 (Re)measuring moral conviction

In her original study Jung measures moral conviction in party manifestos with the MFD as explained in section 1.5. She aggregates the measure to the manifesto level, calculating the proportion of sentences in each manifesto that contains one or more terms from her modified MFD. There are a total of 62,544 sentences in the corpus taken from party manifestos in Australia, Canada, Ireland, New Zealand, the UK and the US from 1993 to 2015. To obtain GPT labels, I pass sentences individually to the 'researcher+guidance' prompt and chat model (see section 1.4.1). I used the chat model because it is significantly cheaper than the completion model, and 'researcher+guidance' is the best-performing prompt overall. Figure 7 shows the differences in the proportion of sentences coded as 'moral' by GPT and Jung's MFD method in each manifesto. The correlation is visible but modest. The deviation from the trendline is strongest in the US, in keeping with results from study 1 where the differences between the MFD and GPT are stronger on US data than UK data.

Figure 7: Moral conviction score by manifesto (GPT and MFD)



Note: deviation measures the average size of residuals from the trendline (dashed black line).

3.1.2 Mobilisation and covariates

I use co-partisan turnout in election surveys to measure mobilisation, using data collected by Jung from the Comparative Study of Electoral Systems and country-specific data sources. Data on education comes from these same sources and is normalised to between 0 and 1. In total, 64 parties across 18 elections enter the analysis.[20] Following Jung I control for age, sex and income at the individual level, niche party at the party level and effective number of parties at the previous election.

3.2 Results

Table 11 shows the results of the replication in the 'GPT' columns, along with the original results from Jung 2020 in the MFD columns. Model 1 is a logistic regression with standard errors clustered at the party-election level and includes country-level fixed effects. Clustering at this level handles correlation in the error term between voters exposed to the same campaign. Model 2 is a logistic multilevel model with intercepts varying at the country, election and party level, helping to account for issues with small sample sizes at the country-election-party level by partially pooling group averages (Gelman & Hill, 2006). These are identical model specifications to in Jung (2020).

In both GPT models the main effect (the moral conviction × education interaction) is larger than in the original paper. This suggests that the GPT measure is picking up textual features that relate more closely to this interaction. The GPT measure introduces several new relationships into the data that are not present in Jung's original paper. Firstly, while the original paper finds no effect for moral conviction not interacted with education, I find that moral conviction suppresses co-partisan turnout overall (given by the negative and statistically significant coefficients for 'moral conviction'). This is demonstrated in Figure 8, which uses model 2 to show the marginal effect of moral conviction on turnout over different education levels. Using Jung's measure, the effect of moral conviction is positive and statistically significant only for those with an education level over 0.95 (15% of respondents). At other levels of education, there is no effect. Using GPT, the effect never becomes positive and statistically significant, but is negative and statistically significant below an education level of 0.32, effecting 20% of respondents.

[20] Australia (2004-2013); Canada (2015); Ireland (2007, 2011); New Zealand (1996-2011); U.K. (1997); U.S. (1992-2012).

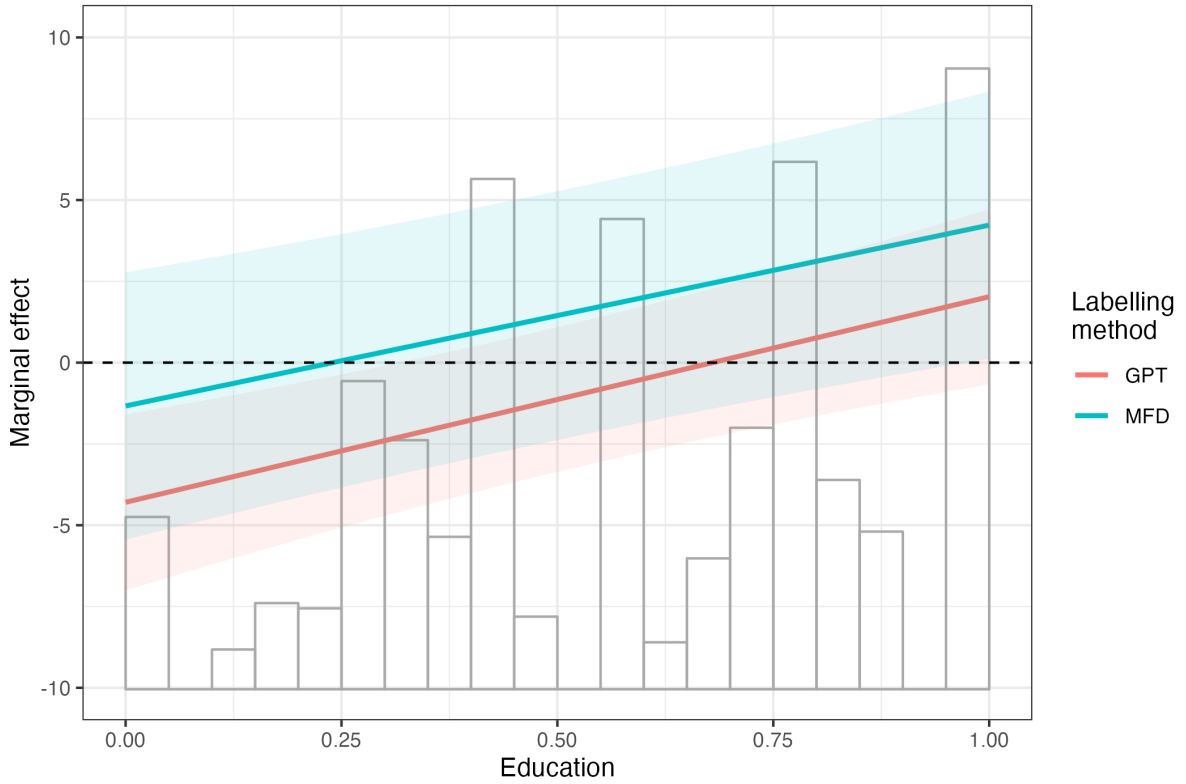
Table 11: effect of moral conviction on turnout across two labelling approaches

	MFD (Jung 2020)		GPT	
	Model 1	Model 2	Model 1	Model 2
Moral conviction	2.95	-1.19	-2.05*	-4.30*
	(3.10)	(2.09)	(0.80)	(1.38)
Moral conviction × education	6.00*	5.46*	7.70*	6.32*
	(2.18)	(1.51)	(2.14)	(1.56)
Education	-0.79	-0.40	-0.03	0.32
	(0.72)	(0.59)	(0.39)	(0.38)
Age (decades)	0.25*	0.26*	0.26*	0.27*
	(0.04)	(0.02)	(0.04)	(0.02)
Male	-0.12	-0.11	-0.11	-0.11
	(0.09)	(0.08)	(0.09)	(0.08)
Income	1.11*	0.99*	1.11*	0.98*
	(0.18)	(0.14)	(0.18)	(0.14)
Niche party	-0.42		-0.31	
	(0.22)		(0.20)	
ENEP (t-1)	-0.00		0.15	
	(0.26)		(0.22)	
Country fixed effects	Yes		Yes	
(Intercept)	1.54	1.14	2.15*	1.50*
	(1.20)	(0.84)	(0.93)	(0.50)
N	14835	14835	14835	14835
Likelihood ratio	1278.34		1268.62	
Log Likelihood		-2269.88		-2267.47

Note: the dependent variable is turnout, where 1 = voted, 0 = did not vote. Model 1 is a logistic regression with standard errors clustered at the party-election level; model 2 is a logistic multilevel model with varying intercepts (Jung 2020). * $p<0.05$.

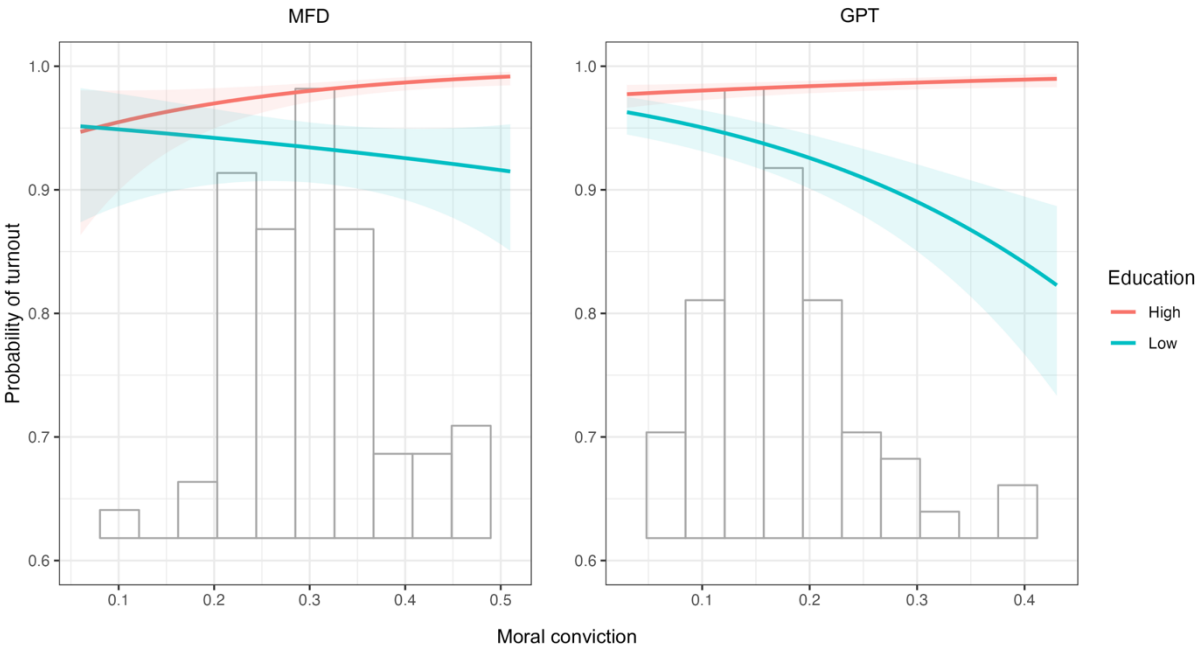
This difference is illustrated in Figure 9, which shows predicted probabilities of turnout across different levels of moral conviction for a voter with education levels 1 or 0 (the maximum and minimum) with other variables held constant at their mean. Using the MFD measure, for a high education voter the likelihood of turnout increases, while for a low education voter it decreases, but the change is not statistically significant. Using the GPT labels, there is no statistically significant increase for a high education voter, whereas the probability of turnout for a low education voter collapses from 0.96 to 0.82. This change is statistically significant and a much larger fall. This suggests that co-partisan turnout among low-education voters is suppressed by moralising rhetoric.

Figure 8: marginal effect of moral conviction on turnout by labelling method



Note: the solid lines denote the marginal effect of moral conviction (according to GPT or the MFD approach) on turnout along levels of education. Shaded areas are 95% confidence intervals. The histogram is the distribution of education levels (Jung 2020).

Figure 9: substantive effect of moral conviction on turnout by labelling method



Note: the solid lines denote the predicted probabilities of turnout for a voter with maximum and minimum education (1 and 0 across the range of moral conviction values. The shaded areas are 95% confidence intervals. Results are based on model 2. The histograms are the distribution of moral conviction according to each labelling approach.

3.3 Discussion

My findings show that using GPT to identify moral conviction (rather than the MFD method) increases the strength of the moral conviction × education interaction. It also shifts the effect such that moral conviction is associated with a reduced likelihood of turnout among low-education voters. Without further testing, it is difficult to know whether GPT this is because GPT's labels contain less noise than the MFD approach, or if GPT is capturing noise that is correlated with a confounder. The shift in the interaction could also suggest that GPT is picking up on specific moral tones that resonate differently with different education groups, capturing nuances in moralising rhetoric that are particularly off-putting or demobilising for less educated voters.

These findings align with my survey results from study 2. There I find that moral conviction causes a positive emotional response across party lines (see A.10). This is a problem for Jung's theory, because her link between education and moral conviction is based on a co-partisan effect. For Jung, education mediates the effect of moral conviction because educated voters are more likely to be exposed to in-party messaging and more likely to have the cognitive tools to understand this messaging (Jung 2020). Study 2 shows that the positive effect of moral conviction is not a co-partisan effect, and this study backs up those results by finding an effect on mobilisation for low-education voters, who – on Jung's account – ought to be minimally exposed to political rhetoric at all. The consistent results across my two studies suggests that GPT is consistent in its identification of moral conviction.

Moral conviction reducing turnout among low-education voters (who are typically lower income) is consistent with theories that contrast values-based competition with interests-based competition (e.g. Tavits & Letki, 2014), where lower-income voters are less receptive to values-based messaging. It also aligns with certain journalistic narratives in the UK that suggest the moralising rhetoric of the Labour party under Jeremy Corbyn alienated working-class voters.[21] This relationship would benefit from further study looking into the differential effects of moral conviction across different parties.

These results demonstrate that GPT can be used to recover moral conviction in political messaging at scale, and that its labels produce results consistent with expectations.

[21] E.g. Gilbert (2020): "Corbyn was intensely moral, but never a working class hero."

Conclusion

In practical terms, my findings suggest that using GPT to identify moral conviction in political texts may be the best option when moral conviction is rare or heavily context dependent such as legislative speech or political manifestos. In these cases, GPT outperforms a dictionary-based approach, and the cost of labelling a large corpus is relatively cheap — it cost me £13 to label 62,544 sentences from the CMP. This also has advantages over a traditional approach of training a model on a large number of human labels, which are more expensive and difficult to generalise to different corpora (Wang et al., 2021). In the case of open-ended survey data, the advantages of using GPT over a simple MFD-based approach are less clear. This advances our understanding of the specific use-cases of LLMs in political science across different corpora and tasks, building on more general earlier work (Ornstein et al., 2023).

My findings also contribute to research on measuring morality in text. Not only do I demonstrate that LLMs can be used for this task, I also expand the measurement of morality beyond moral foundations and into a new domain of text, parliamentary speech, as previous research has overwhelmingly focused on measuring moral foundations expressed in tweets (Araque et al., 2020; Garten et al., 2017; Lin et al., 2017). This shows the promise of using LLMs for new data sources, especially those where we would expect traditional methods to struggle. I also demonstrate the utility of explicitly measuring moral conviction in political text, showing that moral conviction has important consequences in terms of emotional responses to politics and turnout. Future research should make use of the measurement strategy developed in this paper to investigate further impacts of moral conviction in political discourse. My survey findings suggest that parliamentary speech may impact voters emotionally, a contribution to literature on rhetoric in politics that has previously found such effects from adverts and manifestos (Fernandez-Vazquez, 2014; Lipsitz, 2018).

More generally, this paper shows the potential for LLMs like GPT to simulate human labellers when labelling for a subjective concept without a genuine ground truth. GPT is able to closely mimic a human labeller, which aligns with previous research (Argyle et al., 2023; Wang et al., 2021) that uses LLMs as a replacement for human samples. At the same time, given the inbuilt biases and black-box nature of GPT, we should be careful about adopting it at scale before we can answer questions about how it fits into existing social science frameworks: as Spirling (2023) writes, we might want to view variation in labels generated by GPT "as random sampling error with the LLM as a 'population' [or] more like Monte Carlo

error in estimation." In reality neither is exactly true, as GPT is neither an estimator nor a population. This points to further theoretical work that needs to be done on the ontology of LLMs before they are adapted at scale in the social sciences.

Bibliography

Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: A review of BERT-based approaches. *Artificial Intelligence Review*, *54*(8), 5789–5829. https://doi.org/10.1007/s10462-021-09958-2

Amaya, A., Bach, R., Kreuter, F., & Keusch, F. (2020). Measuring the Strength of Attitudes in Social Media Data. In *Big Data Meets Survey Science* (pp. 163–192). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118976357.ch5

Araque, O., Gatti, L., & Kalimeri, K. (2020). MoralStrength: Exploiting a Moral Lexicon and Embedding Similarity for Moral Foundations Prediction. *Knowledge-Based Systems*, *191*, 105184. https://doi.org/10.1016/j.knosys.2019.105184

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, *31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. http://arxiv.org/abs/2005.14165

Chiu, K.-L., Collins, A., & Alexander, R. (2022). *Detecting Hate Speech with GPT-3* (arXiv:2103.12407). arXiv. http://arxiv.org/abs/2103.12407

Clifford, S., & Jerit, J. (2013). How words do the work of politics: Moral foundations theory and the debate over stem cell research. *The Journal of Politics*, *75*(3), 659–671. https://doi.org/10.1017/S0022381613000492

Clifford, S., Jerit, J., Rainey, C., & Motyl, M. (2015). Moral Concerns and Policy Attitudes: Investigating the Influence of Elite Rhetoric. *Political Communication*, *32*(2), 229–248. https://doi.org/10.1080/10584609.2014.944320

Clifford, S., & Simas, E. N. (2022). Moral Rhetoric, Extreme Positions, and Perceptions of Candidate Sincerity. *Political Behavior*. https://doi.org/10.1007/s11109-022-09835-w

Del Arco, F. M., Collado-Montañez, J., Ureña, L. A., & Martín-Valdivia, M.-T. (2022). Empathy and Distress Prediction using Transformer Multi-output Regression and Emotion Analysis with an Ensemble of Supervised and Zero-Shot Learning Models. *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, 239–244. https://doi.org/10.18653/v1/2022.wassa-1.23

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

Editorial board. (2022, July 31). The Guardian view on housing costs: A grave and growing injustice. *The Guardian*. https://www.theguardian.com/commentisfree/2022/jul/31/the-guardian-view-on-housing-costs-a-grave-and-growing-injustice

Feinberg, M., & Willer, R. (2015). From Gulf to Bridge: When Do Moral Arguments Facilitate Political Influence? *Personality and Social Psychology Bulletin*, *41*(12), 1665–1681. https://doi.org/10.1177/0146167215607842

Feinberg, M., & Willer, R. (2019). Moral reframing: A technique for effective and persuasive communication across political divides. *Social and Personality Psychology Compass*, *13*(12), e12501. https://doi.org/10.1111/spc3.12501

Fernandez-Vazquez, P. (2014). *And Yet It Moves*: The Effect of Election Platforms on Party Policy Images. *Comparative Political Studies*, *47*(14), 1919–1944. https://doi.org/10.1177/0010414013516067

Garrett, K. N., & Bankert, A. (2020). The Moral Roots of Partisan Division: How Moral Conviction Heightens Affective Polarization. *British Journal of Political Science*, *50*(2), 621–640. https://doi.org/10.1017/S000712341700059X

Garten, J., Boghrati, R., Hoover, J., Johnson, K., & Dehghani, M. (2017). Morality between the lines: Detecting moral sentiment in text. *Proceedings of IJCAI 2016 Workshop on Computational Modeling of Attitudes*.

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Gentzkow, M., Shapiro, J. M., & Taddy, M. (2018). *Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts* [dataset]. https://data.stanford.edu/congress_text.

Gilbert, J. (2020, January 14). *Corbyn was intensely moral, but never a working class hero.* OpenDemocracy. https://www.opendemocracy.net/en/opendemocracyuk/corbyn-was-intensely-moral-never-working-class-hero/

Graham, J., & Haidt, J. (2012). *Moral Foundations Dictionary* [dataset]. http://moralfoundations.orgReturn

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), 1029–1046. https://doi.org/10.1037/a0015141

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Haidt, J., & Graham, J. (2007). When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize. *Social Justice Research*, *20*(1), 98–116. https://doi.org/10.1007/s11211-007-0034-z

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, *133*(4), 55–66. https://doi.org/10.1162/0011526042365555

Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A. M., Lin, Y., Kennedy, B., Atari, M., Kamel, Z., Mendlen, M., Moreno, G., Park, C., Chang, T. E., Chin, J., Leong, C., Leung, J. Y., Mirinjian, A., & Dehghani, M. (2019). *Moral Foundations Twitter Corpus: A collection of 35k tweets annotated for moral sentiment.* PsyArXiv. https://doi.org/10.31234/osf.io/w4f72

Hornsey, M. J., Majkut, L., Terry, D. J., & McKimmie, B. M. (2003). On being loud and proud: Non-conformity and counter-conformity to group norms. *The British Journal of Social Psychology*, *42*(Pt 3), 319–335. https://doi.org/10.1348/014466603322438189

Inglehart, R. (1971). The Silent Revolution in Europe: Intergenerational Change in Post-Industrial Societies. *American Political Science Review*, *65*(4), 991–1017. https://doi.org/10.2307/1953494

Inglehart, R. (1997). *Modernization and postmodernization: Cultural, economic, and political change in 43 societies*. Princeton University Press.

Jung, J. (2020). The Mobilizing Effect of Parties' Moral Rhetoric. *American Journal of Political Science*, *64*(2), 341–355. https://doi.org/10.1111/ajps.12476

Kayes, A. S. M., Islam, M. S., Watters, P. A., Ng, A., & Kayesh, H. (2020). *Automated Measurement of Attitudes Towards Social Distancing Using Social Media: A COVID-19 Case Study* (2020040057). Preprints. https://doi.org/10.20944/preprints202004.0057.v1

Kosinski, M. (2023). *Theory of Mind May Have Spontaneously Emerged in Large Language Models* (arXiv:2302.02083). arXiv. https://doi.org/10.48550/arXiv.2302.02083

Kraft, P. W. (2018). Measuring Morality in Political Attitude Expression. *The Journal of Politics*, *80*(3), 1028–1033. https://doi.org/10.1086/696862

Kraft, P. W., & Klemmensen, R. (2023). Lexical Ambiguity in Political Rhetoric: Why Morality Doesn't Fit in a Bag of Words. *British Journal of Political Science*, 1–19. https://doi.org/10.1017/S000712342300008X

Krosnick, J. A. (1999). Survey Research. *Annual Review of Psychology*, *50*(1), 537–567. https://doi.org/10.1146/annurev.psych.50.1.537

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lawford, M. (2023, February 12). How Britain's broken housing market is crushing growth. *The Telegraph*. https://www.telegraph.co.uk/business/2023/02/12/how-britains-broken-housing-market-crushing-growth/

Li, X., Li, Y., Joty, S., Liu, L., Huang, F., Qiu, L., & Bing, L. (2023). *Does GPT-3 Demonstrate Psychopathy? Evaluating Large Language Models from a Psychological Perspective* (arXiv:2212.10529). arXiv. https://doi.org/10.48550/arXiv.2212.10529

Lin, Y., Hoover, J., Dehghani, M., Mooijman, M., & Ji, H. (2017). *Acquiring Background Knowledge to Improve Moral Value Prediction* (arXiv:1709.05467). arXiv. http://arxiv.org/abs/1709.05467

Lipsitz, K. (2018). Playing with Emotions: The Effect of Moral Appeals in Elite Rhetoric. *Political Behavior*, *40*(1), 57–78. https://doi.org/10.1007/s11109-017-9394-8

Magar, I., & Schwartz, R. (2022). *Data Contamination: From Memorization to Exploitation* (arXiv:2203.08242). arXiv. https://doi.org/10.48550/arXiv.2203.08242

Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., & Miori, M. (2022). *Does GPT-3 know what the Most Important Issue is? Using Large Language Models to Code Open-Text Social Survey Responses At Scale* (SSRN Scholarly Paper 4310154). https://doi.org/10.2139/ssrn.4310154

Miotto, M., Rossberg, N., & Kleinberg, B. (2022). *Who is GPT-3? An Exploration of Personality, Values and Demographics* (arXiv:2209.14338). arXiv. https://doi.org/10.48550/arXiv.2209.14338

Morgan, G. S., Skitka, L. J., & Wisneski, D. C. (2010). Moral and religious convictions and intentions to vote in the 2008 presidential election. *Analyses of Social Issues and Public Policy (ASAP)*, *10*(1), 307–320. https://doi.org/10.1111/j.1530-2415.2010.01204.x

Niszczota, P., & Abbas, S. (2023). *GPT as a Financial Advisor* (SSRN Scholarly Paper 4384861). https://doi.org/10.2139/ssrn.4384861

Odell, E. (2021). *Hansard Speeches 1979-2021* (3.1.0) [dataset]. https://doi.org/10.5281/zenodo.4843485

OpenAI. (2023). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

Ornstein, J. T., Blasingame, E. N., & Truscott, J. S. (2023). *How to Train Your Stochastic Parrot: Large Language Models for Political Texts.* https://joeornstein.github.io/publications/ornstein-blasingame-truscott.pdf

Osborne, D., & Sibley, C. G. (2022). *The Cambridge Handbook of Political Psychology.* Cambridge University Press.

Osnabrügge, M., Hobolt, S. B., & Rodon, T. (2021). Playing to the Gallery: Emotive Rhetoric in Parliaments. *American Political Science Review*, *115*(3), 885–899. https://doi.org/10.1017/S0003055421000356

Pembury Smith, M. Q. R., & Ruxton, G. D. (2020). Effective use of the McNemar test. *Behavioral Ecology and Sociobiology*, *74*(11), 133. https://doi.org/10.1007/s00265-020-02916-y

Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, *27*(3), 258–284. https://doi.org/10.1080/00909889909365539

Ryan, T. J. (2014). Reconsidering Moral Issues in Politics. *The Journal of Politics*, *76*(2), 380–397. https://doi.org/10.1017/S0022381613001357

Ryan, T. J. (2017). No Compromise: Political Consequences of Moralized Attitudes. *American Journal of Political Science*, *61*(2), 409–423. https://doi.org/10.1111/ajps.12248

Ryan, T. J. (2019). Actions versus Consequences in Political Arguments: Insights from Moral Psychology. *The Journal of Politics*, *81*(2), 426–440. https://doi.org/10.1086/701494

Sabucedo, J.-M., Dono, M., Alzate, M., & Seoane, G. (2018). The Importance of Protesters' Morals: Moral Obligation as a Key Variable to Understand Collective Action. *Frontiers in Psychology*, *9*. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00418

Sandri, M., Leonardelli, E., Tonelli, S., & Jezek, E. (2023). Why Don't You Do It Right? Analysing Annotators' Disagreement in Subjective Tasks. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2428–2441. https://aclanthology.org/2023.eacl-main.178

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), Article 3. https://doi.org/10.1038/s42256-022-00458-8

Seyeditabari, A., Levens, S., Maestas, C. D., Shaikh, S., Walsh, J. I., Zadrozny, W., Danis, C., & Thompson, O. P. (2018). *Cross Corpus Emotion Classification Using Survey Data* (SSRN Scholarly Paper 3108133). https://doi.org/10.2139/ssrn.3108133

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., & Shwartz, V. (2023). *Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models* (arXiv:2305.14763). arXiv. https://doi.org/10.48550/arXiv.2305.14763

Sim, J., & Wright, C. C. (2005). The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy*, *85*(3), 257–268. https://doi.org/10.1093/ptj/85.3.257

Simonsen, K. B., & Bonikowski, B. (2022). Moralizing Immigration: Political Framing, Moral Conviction, and Polarization in the United States and Denmark. *Comparative Political Studies*, *55*(8), 1403–1436. https://doi.org/10.1177/00104140211060284

Skitka, L. J. (2010). The Psychology of Moral Conviction. *Social and Personality Psychology Compass*, *4*(4), 267–281. https://doi.org/10.1111/j.1751-9004.2010.00254.x

Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral Conviction: Another Contributor to Attitude Strength or Something More? *Journal of Personality and Social Psychology*, *88*(6), 895–917. https://doi.org/10.1037/0022-3514.88.6.895

Skitka, L. J., Hanson, B. E., Morgan, G. S., & Wisneski, D. C. (2021). The Psychology of Moral Conviction. *Annual Review of Psychology*, *72*(1), 347–366. https://doi.org/10.1146/annurev-psych-063020-030612

Skitka, L. J., Morgan, G. S., & Wisneski, D. C. (2015). Political orientation and moral conviction: A conservative advantage or an equal opportunity motivator of political engagement? In *Social psychology and politics* (pp. 57–74). Psychology Press.

Skitka, L. J., & Wisneski, D. C. (2011). Moral conviction and emotion. *Emotion Review*, *3*(3), 328–330. https://doi.org/10.1177/1754073911402374

Spirling, A. (2023, July 27). Interesting to read different takes on what the variance in LLM responses. *Twitter*. https://twitter.com/arthur_spirling/status/1684591605196652544

Tavits, M., & Letki, N. (2014). From Values to Interests? The Evolution of Party Competition in New Democracies. *The Journal of Politics*, *76*(1), 246–258. https://doi.org/10.1017/S002238161300131X

Valentino, N. A., Brader, T., Groenendyk, E. W., Gregorowicz, K., & Hutchings, V. L. (2011). Election Night's Alright for Fighting: The Role of Emotions in Political Participation. *The Journal of Politics*, *73*(1), 156–170. https://doi.org/10.1017/S0022381610000939

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762

Wang, S., Liu, Y., Xu, Y., Zhu, C., & Zeng, M. (2021). *Want To Reduce Labeling Cost? GPT-3 Can Help* (arXiv:2108.13487). arXiv. http://arxiv.org/abs/2108.13487

Xu, B., Yang, A., Lin, J., Wang, Q., Zhou, C., Zhang, Y., & Mao, Z. (2023). *ExpertPrompting: Instructing Large Language Models to be Distinguished Experts* (arXiv:2305.14688). arXiv. https://doi.org/10.48550/arXiv.2305.14688

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–21. https://doi.org/10.1145/3544548.3581388

Zhang, M., & Li, J. (2021). A commentary of GPT-3 in MIT Technology Review 2021. *Fundamental Research*, *1*(6), 831–833. https://doi.org/10.1016/j.fmre.2021.11.011

Zollinger, D. (2022). Cleavage Identities in Voters' Own Words: Harnessing Open-Ended Survey Responses. *American Journal of Political Science*. https://doi.org/10.1111/ajps.12743

Appendix

A.1 Corpora details

This section contains information on the 250 texts randomly selected from each corpus to be labelled by humans and GPT compared to the entire corpus.

A.1.1 ANES

All data is all MII responses with > 30 characters

	All data (n=42397)	Sample (n=250)
% male	46.5%	47.2%
% party id= Dem	47.9%	44.5%
Year = 2008	10.4%	9.3%
Year = 2012	38.1%	39.5%
Year = 2016	10.1%	12.2%
Year = 2020	41.4%	39.1%
Mean number of characters	45.67	44.26

A.1.2 BES

All data is all MII responses with > 30 characters

	All data (n=43864)	Sample (n=250)
Median wave	10 (5, 17)	9 (4.25, 17)
% party id= Lab	26.4%	29.2%
% male	51.1%	50.9%
% voted Remain	48.5%	46.4%
Mean number of characters	53.13	51.01

A.1.3 House of Commons

All data is all sentences from PMQs

	All data (n=121788)	Sample (n=250)
Period 2001-2010	43.9%	45.2%
]% party = Lab	45.3%	44.0%
% female	19.1%	18.2%
Mean number of characters	112.95	111.28

A.1.4 US Congress

All data is all sentences from SOU addresses and day of SOU addresses

	All data (n=57645)	Sample (n=250)
Period 2001-2010	40.4%	39.5%
% female	17.0%	19.2%
Mean number of characters	111.84	114.50

A.2 Text preprocessing

I split texts into sentences using the corpus_reshape function from the quanteda package in R. I couldn't find details on exactly how corpus_reshape works, but it can handle abbreviations like Dr. or Hon. to split speeches into sentences correctly.

The US Congress corpus required further pre-processing because in the downloaded data from Gentzkow et al. (2018) commas have been mis-parsed as full stops, meaning there are many more full stops than there should be. To fix this, I replaced all full stops that are followed by a lowercase letter with commas, and then used the corpus_reshape function.

A.3 Human coding guide

Human annotators received the following guidance. Examples were changed depending on the corpus. This is an example for the ANES corpus.

***

- Your task is to label texts for 'moral conviction'. This is when a text expresses a belief that something is right or wrong.
- Put either 'moral' or 'non-moral' in the label column.
- Moral sentences typically contain value judgment or speaks to issues of right and wrong, fairness, justice, or ethics.
- Non-moral sentences are any that don't. They might be procedural, or making a judgement not based on any moral claims.
- Sometimes sentences might imply a moral claim, but if they aren't actually making one, they should be labelled as non-moral.
- There will be some that you are unsure about. In this case, just go with whichever you think is most likely. Don't worry too much about it.
- A lot of the sentences have typos and formatting errors. Ignore these.
- You don't need to justify the labels, just put them down.
- Ignore any texts not in English.
- Don't spend too long considering your answer.

Example 1:
Statement: "we need more jobs. not too many people are working."
Classification: Non-moral. *While it might hint at intrinsic value of work and work being good, there is no explicit or really even implicit moral claim.*

Example 2:
Statement: "climate change - us needs to make progress."
Classification: Non-moral. *No moral reasoning present.*

Example 3:
Statement: "treatment of immigrants, climate change, racial injustice"
Classification: Moral. *Injustice mentioned which is explicitly moral, and treatment of immigrants is gesturing at fairness and decency which are moral concepts.*

Example 4:
Statement: "capitalism and the enslavement of the poor in other countries that make our junk."
Classification: Moral. *The issue with capitalism being presented here is a moral one — that it enslaves people.*

Example 5:
Statement: "In the African American community in particular, 17 percent are unemployed."
Classification: Non-moral. *While unemployment is negative, there is no moral claim made about the unemployment. It is purely factual. If it said '17 percent are unemployed in a grave injustice' that would be moral.*

Example 6:
Statement: "illegal immigrants using benefits.abortion as a form of eugenics; minorities are not informed of this.."
Classification: Moral. *The issue with illegal immigrants using benefits is clearly one of desert/justice so it is moral.*

Example 7:
Statement: "transition of power in jeopardy, election integrity being questioned by a fascist president"
Classification: Non-moral. *While this is charged language, it is not making a moral claim.*

A.4 Human annotator information

Annotator	Age	Highest education level	Political orientation (Left-right)
1	25	Post-graduate	1
2	26	Post-graduate	1
3	25	Post-graduate	2
4	26	Post-graduate	3
5	26	Post-graduate	1
6	26	Undergraduate	2

Notes: Left-right scale is from 1 (left) to 7 (right)

A.5 Full IAA agreement and other metrics

Corpus	Annotator 1	Annotator 2	Accuracy	Precision	Recall	F1	PABAK	kappa
BES	One	Three	0.82	0.44	0.58	0.50	0.65	0.40
BES	One	Four	0.78	0.33	0.45	0.38	0.56	0.25
BES	Three	Four	0.74	0.35	0.36	0.36	0.48	0.19
HoC	One	Two	0.85	0.35	0.86	0.50	0.70	0.43
HoC	One	Three	0.89	0.35	0.32	0.33	0.78	0.27
HoC	Two	Three	0.85	0.90	0.33	0.49	0.70	0.42
USC	One	Five	0.88	0.39	0.28	0.33	0.77	0.26
USC	One	Six	0.91	0.64	0.28	0.39	0.82	0.35
USC	Five	Six	0.96	0.82	0.50	0.62	0.91	0.60
ANES	One	Two	0.77	0.40	1.00	0.57	0.54	0.45
ANES	One	Five	0.84	0.49	0.59	0.54	0.69	0.44
ANES	Two	Five	0.76	0.89	0.43	0.58	0.52	0.44

A.6 Full GPT prompts

A.6.1 Chat prompts (16)

A.6.1.1 US Congress

```
usc_prompt_chat_full = [
        {"role": "system", "content": "You are an expert social scientist classifying sentences as 'moral' or 'non-moral'. A sentence should be classified as moral if it expresses a belief that something is right or wrong."},
        {"role": "user", "content": 'Please classify the following statement as either "moral" or "non-moral": "It is wrong that hardworking men and women can not afford to put food on the table or heat their homes."'},
        {"role": "assistant", "content": "Moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral" or "non-moral": "It has been argued by some on the other side that this would undermine the ability to reach a comprehensive settlement on entitlement reform."'},
        {"role": "assistant", "content": "Non-moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral" or "non-moral". Do not respond with any classification other than "moral" or "non-moral": "{}"'},
    ]

usc_prompt_chat_res = [
        {"role": "system", "content": "You are an expert social science researcher classifying statements as 'moral' or ‘non-moral’."},
        {"role": "user", "content": 'Please classify the following statement as either "moral" or "non-moral": "It is wrong that hardworking men and women can not afford to put food on the table or heat their homes."'},
        {"role": "assistant", "content": "Moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral" or "non-moral": "It has been argued by some on the other side that this would undermine the ability to reach a comprehensive settlement on entitlement reform."'},
        {"role": "assistant", "content": "Non-moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral" or "non-moral". Do not respond with any classification other than "moral" or "non-moral": "{}"'},
    ]

usc_prompt_chat_guide = [
        {"role": "system", "content": "You are an assistant that classifies statements as ‘moral' or ‘non-moral’. A sentence should be classified as moral if it expresses a belief that something is right or wrong."},
        {"role": "user", "content": 'Please classify the following statement as either "moral" or "non-moral": "It is wrong that hardworking men and women can not afford to put food on the table or heat their homes."'},
        {"role": "assistant", "content": "Moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral" or "non-moral": "It has been argued by some on the other side that this would undermine the ability to reach a comprehensive settlement on entitlement reform."'},
```

```
        {"role": "assistant", "content": "Non-moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral". Do not respond with any classification other than "moral" or "non-moral":
"{}"'},
    ]

usc_prompt_chat_min = [
        {"role": "system", "content": "You are an assistant that classifies statements as
‘moral' or ‘non-moral."},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral": "It is wrong that hardworking men and women can not afford to put food on the
table or heat their homes."'},
        {"role": "assistant", "content": "Moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral": "It has been argued by some on the other side that this would undermine the
ability to reach a comprehensive settlement on entitlement reform."'},
        {"role": "assistant", "content": "Non-moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral". Do not respond with any classification other than "moral" or "non-moral":
"{}"'},
    ]
```

A.6.1.2 House of Commons

```
ukp_prompt_chat_full = [
        {"role": "system", "content": "You are an expert social scientist classifying sentences
as 'moral' or 'non-moral'. A sentence should be classified as moral if it expresses a belief
that something is right or wrong."},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral": "In a week when the amount of abuse of the elderly is being noted, I think it
right to say that no citizen should be engaged in anything that puts the dignity and security
of elderly people in our country at risk."'},
        {"role": "assistant", "content": "Moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral": "We want to see life sciences and these areas succeed in Britain, and Porton
Down has an important role to play."'},
        {"role": "assistant", "content": "Non-moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral". Do not respond with any classification other than "moral" or "non-moral":
"{}"'},
    ]

ukp_prompt_chat_res = [
        {"role": "system", "content": "You are an expert social science researcher classifying
statements as 'moral' or ‘non-moral’."},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral": "In a week when the amount of abuse of the elderly is being noted, I think it
right to say that no citizen should be engaged in anything that puts the dignity and security
of elderly people in our country at risk."'},
        {"role": "assistant", "content": "Moral"},
```

```
        {"role": "user", "content": 'Please classify the following statement as either "moral" or "non-moral": "We want to see life sciences and these areas succeed in Britain, and Porton Down has an important role to play."'},
        {"role": "assistant", "content": "Non-moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral" or "non-moral". Do not respond with any classification other than "moral" or "non-moral": "{}"'},
    ]

ukp_prompt_chat_guide = [
        {"role": "system", "content": "You are an assistant that classifies statements as ‘moral' or ‘non-moral’. A sentence should be classified as moral if it expresses a belief that something is right or wrong."},
        {"role": "user", "content": 'Please classify the following statement as either "moral" or "non-moral": "In a week when the amount of abuse of the elderly is being noted, I think it right to say that no citizen should be engaged in anything that puts the dignity and security of elderly people in our country at risk."'},
        {"role": "assistant", "content": "Moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral" or "non-moral": "We want to see life sciences and these areas succeed in Britain, and Porton Down has an important role to play."'},
        {"role": "assistant", "content": "Non-moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral" or "non-moral". Do not respond with any classification other than "moral" or "non-moral": "{}"'},
    ]

ukp_prompt_chat_min = [
        {"role": "system", "content": "You are an assistant that classifies statements as ‘moral' or ‘non-moral."},
        {"role": "user", "content": 'Please classify the following statement as either "moral" or "non-moral": "In a week when the amount of abuse of the elderly is being noted, I think it right to say that no citizen should be engaged in anything that puts the dignity and security of elderly people in our country at risk."'},
        {"role": "assistant", "content": "Moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral" or "non-moral": "We want to see life sciences and these areas succeed in Britain, and Porton Down has an important role to play."'},
        {"role": "assistant", "content": "Non-moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral" or "non-moral". Do not respond with any classification other than "moral" or "non-moral": "{}"'},
    ]
```

A.6.1.3 ANES

```
anes_prompt_chat_full = [
        {"role": "system", "content": "You are an expert social scientist classifying sentences as 'moral' or 'non-moral'. A sentence should be classified as moral if it expresses a belief that something is right or wrong."},
```

```
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral": "Everyone paying their fair share of taxes."'},
        {"role": "assistant", "content": "Moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral": "Poor management of government spending."'},
        {"role": "assistant", "content": "Non-moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral". Do not respond with any classification other than "moral" or "non-moral":
"{}"'},
    ]

anes_prompt_chat_res = [
        {"role": "system", "content": "You are an expert social science researcher classifying
statements as 'moral' or ‘non-moral’."},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral": "Everyone paying their fair share of taxes."'},
        {"role": "assistant", "content": "Moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral": "Poor management of government spending."'},
        {"role": "assistant", "content": "Non-moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral". Do not respond with any classification other than "moral" or "non-moral":
"{}"'},
    ]

anes_prompt_chat_guide = [
        {"role": "system", "content": "You are an assistant that classifies statements as
‘moral' or ‘non-moral’. A sentence should be classified as moral if it expresses a belief that
something is right or wrong."},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral": "Everyone paying their fair share of taxes."'},
        {"role": "assistant", "content": "Moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral": "Poor management of government spending."'},
        {"role": "assistant", "content": "Non-moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral". Do not respond with any classification other than "moral" or "non-moral":
"{}"'},
    ]

anes_prompt_chat_min = [
        {"role": "system", "content": "You are an assistant that classifies statements as
‘moral' or ‘non-moral."},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral": "Everyone paying their fair share of taxes."'},
        {"role": "assistant", "content": "Moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral": "Poor management of government spending."'},
        {"role": "assistant", "content": "Non-moral"},
```

```
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral". Do not respond with any classification other than "moral" or "non-moral":
"{}"'},
    ]
```

A.6.1.4 BES

```
bes_prompt_chat_full = [
        {"role": "system", "content": "You are an expert social scientist classifying sentences
as 'moral' or 'non-moral'. A sentence should be classified as moral if it expresses a belief
that something is right or wrong."},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral": "Unfairness of housing market and corporate greed."'},
        {"role": "assistant", "content": "Moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral": "Establishing stability post-referendum."'},
        {"role": "assistant", "content": "Non-moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral". Do not respond with any classification other than "moral" or "non-moral":
"{}"'},
    ]

bes_prompt_chat_res = [
        {"role": "system", "content": "You are an expert social science researcher classifying
statements as 'moral' or ‘non-moral’."},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral": "Unfairness of housing market and corporate greed."'},
        {"role": "assistant", "content": "Moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral": "Establishing stability post-referendum."'},
        {"role": "assistant", "content": "Non-moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral". Do not respond with any classification other than "moral" or "non-moral":
"{}"'},
    ]

bes_prompt_chat_guide = [
        {"role": "system", "content": "You are an assistant that classifies statements as
‘moral' or ‘non-moral’. A sentence should be classified as moral if it expresses a belief that
something is right or wrong."},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral": "Unfairness of housing market and corporate greed."'},
        {"role": "assistant", "content": "Moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral": "Establishing stability post-referendum."'},
        {"role": "assistant", "content": "Non-moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral". Do not respond with any classification other than "moral" or "non-moral":
"{}"'},
    ]
```

```
bes_prompt_chat_min = [
        {"role": "system", "content": "You are an assistant that classifies statements as
'moral' or 'non-moral."},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral": "Unfairness of housing market and corporate greed."'},
        {"role": "assistant", "content": "Moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral": "Establishing stability post-referendum."'},
        {"role": "assistant", "content": "Non-moral"},
        {"role": "user", "content": 'Please classify the following statement as either "moral"
or "non-moral". Do not respond with any classification other than "moral" or "non-moral":
"{}"'},
    ]
```

A.6.2 Completion prompts (16)

A.6.2.1 US Congress

```
usc_prompt_comp_full = """
You are an expert social scientist classifying sentences as 'moral' or 'non-moral'. A sentence
should be classified as moral if it expresses a belief that something is right or wrong. Please
classify the following statement as either "moral" or "non-moral".

Example 1:
Statement: "It is wrong that hardworking men and women can not afford to put food on the table
or heat their homes."
Moral

Example 2:
Statement: "It has been argued by some on the other side that this would undermine the ability
to reach a comprehensive settlement on entitlement reform."
Non-moral
---

Statement: "{}"
"""

usc_prompt_comp_res = """
You are an expert social scientist classifying sentences as 'moral' or 'non-moral'. Please
classify the following statement as either "moral" or "non-moral".

Example 1:
Statement: "It is wrong that hardworking men and women can not afford to put food on the table
or heat their homes."
Moral

Example 2:
Statement: "It has been argued by some on the other side that this would undermine the ability
to reach a comprehensive settlement on entitlement reform."
Non-moral
---
```

```
Statement: "{}"
"""

usc_prompt_comp_guide = """
Please classify the following statement as either "moral" or "non-moral". A sentence should be classified as moral if it expresses a belief that something is right or wrong.

Example 1:
Statement: "It is wrong that hardworking men and women can not afford to put food on the table or heat their homes."
Moral

Example 2:
Statement: "It has been argued by some on the other side that this would undermine the ability to reach a comprehensive settlement on entitlement reform."
Non-moral
---

Statement: "{}"
"""


usc_prompt_comp_min = """
Please classify the following statement as either "moral" or "non-moral".

Example 1:
Statement: "It is wrong that hardworking men and women can not afford to put food on the table or heat their homes."
Moral

Example 2:
Statement: "It has been argued by some on the other side that this would undermine the ability to reach a comprehensive settlement on entitlement reform."
Non-moral
---

Statement: "{}"
"""
```

A.6.2.2 House of Commons

```
ukp_prompt_comp_full = """
You are an expert social scientist classifying sentences as 'moral' or 'non-moral'. A sentence should be classified as moral if it expresses a belief that something is right or wrong. Please classify the following statement as either "moral" or "non-moral".

Example 1:
Statement: "In a week when the amount of abuse of the elderly is being noted, I think it right to say that no citizen should be engaged in anything that puts the dignity and security of elderly people in our country at risk."
```

```
Moral

Example 2:
Statement: "We want to see life sciences and these areas succeed in Britain, and Porton Down has an important role to play."
Non-moral
---

Statement: "{}"
"""

ukp_prompt_comp_res = """
You are an expert social scientist classifying sentences as 'moral' or 'non-moral'. Please classify the following statement as either "moral" or "non-moral".

Example 1:
Statement: "In a week when the amount of abuse of the elderly is being noted, I think it right to say that no citizen should be engaged in anything that puts the dignity and security of elderly people in our country at risk."
Moral

Example 2:
Statement: "We want to see life sciences and these areas succeed in Britain, and Porton Down has an important role to play."
Non-moral
---

Statement: "{}"
"""

ukp_prompt_comp_guide = """
Please classify the following statement as either "moral" or "non-moral". A sentence should be classified as moral if it expresses a belief that something is right or wrong.

Example 1:
Statement: "In a week when the amount of abuse of the elderly is being noted, I think it right to say that no citizen should be engaged in anything that puts the dignity and security of elderly people in our country at risk."
Moral

Example 2:
Statement: "We want to see life sciences and these areas succeed in Britain, and Porton Down has an important role to play."
Non-moral
---

Statement: "{}"
"""
```

```
ukp_prompt_comp_min = """
Please classify the following statement as either "moral" or "non-moral".

Example 1:
Statement: "In a week when the amount of abuse of the elderly is being noted, I think it right to say that no citizen should be engaged in anything that puts the dignity and security of elderly people in our country at risk."
Moral

Example 2:
Statement: "We want to see life sciences and these areas succeed in Britain, and Porton Down has an important role to play."
Non-moral
---

Statement: "{}"
"""
```

A.6.2.3 ANES

```
anes_prompt_comp_full = """
You are an expert social scientist classifying sentences as 'moral' or 'non-moral'. A sentence should be classified as moral if it expresses a belief that something is right or wrong. Please classify the following statement as either "moral" or "non-moral".

Example 1:
Statement: "Everyone paying their fair share of taxes."
Moral

Example 2:
Statement: "Poor management of government spending."
Non-moral
---

Statement: "{}"
"""

anes_prompt_comp_res = """
You are an expert social scientist classifying sentences as 'moral' or 'non-moral'. Please classify the following statement as either "moral" or "non-moral".

Example 1:
Statement: "Everyone paying their fair share of taxes."
Moral

Example 2:
Statement: "Poor management of government spending."
Non-moral
---

Statement: "{}"
```

```
"""

anes_prompt_comp_guide = """
Please classify the following statement as either "moral" or "non-moral". A sentence should be
classified as moral if it expresses a belief that something is right or wrong.

Example 1:
Statement: "Everyone paying their fair share of taxes."
Moral

Example 2:
Statement: "Poor management of government spending."
Non-moral
---

Statement: "{}"
"""



anes_prompt_comp_min = """
Please classify the following statement as either "moral" or "non-moral".

Example 1:
Statement: "Everyone paying their fair share of taxes."
Moral

Example 2:
Statement: "Poor management of government spending."
Non-moral
---

Statement: "{}"
"""
```

A.6.2.4 BES

```
bes_prompt_comp_full = """
You are an expert social scientist classifying sentences as 'moral' or 'non-moral'. A sentence
should be classified as moral if it expresses a belief that something is right or wrong. Please
classify the following statement as either "moral" or "non-moral".

Example 1:
Statement: "Unfairness of housing market and corporate greed."
Moral

Example 2:
Statement: "Establishing stability post-referendum."
Non-moral
---

Statement: "{}"
```

```
"""

bes_prompt_comp_res = """
You are an expert social scientist classifying sentences as 'moral' or 'non-moral'. Please
classify the following statement as either "moral" or "non-moral".

Example 1:
Statement: "Unfairness of housing market and corporate greed."
Moral

Example 2:
Statement: "Establishing stability post-referendum."
Non-moral
---

Statement: "{}"
"""

bes_prompt_comp_guide = """
Please classify the following statement as either "moral" or "non-moral". A sentence should be
classified as moral if it expresses a belief that something is right or wrong.

Example 1:
Statement: "Unfairness of housing market and corporate greed."
Moral

Example 2:
Statement: "Establishing stability post-referendum."
Non-moral
---

Statement: "{}"
"""


bes_prompt_comp_min = """
Please classify the following statement as either "moral" or "non-moral".

Example 1:
Statement: "Unfairness of housing market and corporate greed."
Moral

Example 2:
Statement: "Establishing stability post-referendum."
Non-moral
---

Statement: "{}"
"""
```

A.7 Error analysis regression models

These are log-odds results from multinomial logit models to compare the effects of the MFD count variable on false positives and negatives.

Corpus	Effect of MFD count on likelihood of false positive	Effect of MFD count on likelihood of false negative
ANES	0.55 (0.172)	-0.22 (0.295)
USC	1.00 (0.468)	0.36 (0.214)
HoC	0.61 (0.352)	0.43 (0.196)

A.8 Thematic analysis

Corpus	Theme	N (% of errors)	Example
ANES	Implicit morality	18 (33%)	"social injustice, illegal immigrants, drugs, healthcare"
ANES	Lack of context	29 (53%)	"we are not free and i have no faith in our voting system. it was fixed"
ANES	Complexity and ambiguity	12 (22%)	"jobs there's, going to be more people with out jobs doctors and medical staff charge too much for medical care , congress should have a limit on there salary and also they shouldn't be able to raise there salary it should be someone other than them that does."
ANES	Overemphasis on keywords	10 (18%)	"climate change-- it needs to be talked about as much as possible to set an example for other countries and create new industries"
HoC	Implicit morality	15 (50%)	"The Prime Minister has also said that too many of the guilty are going free."
HoC	Lack of context	13 (43%)	"It is important that risk be seen to lie with the banks and the lenders and not be underwritten by the taxpayer."
HoC	Complexity and ambiguity	4 (13%)	"Is it getting stuck in with Pfizer and AstraZeneca, battling for the British interest, or is it standing back like him, doing absolutely nothing apart from playing politics?"
HoC	Overemphasis on keywords	5 (17%)	"The pain that has been caused to the Windrush generation needs to be resolved very rapidly, with full compensation paid as quickly as it can possibly be done and an understanding of the hurt that they feel."

A.9 GPT liberal and conservative prompts

The variation in performance across UK and US data, as well as variation in PABAK scores across different GPT prompts, suggests that we may be able to systematically alter how GPT identifies moral conviction. This would be consistent with previous literature (Argyle et al. 2023). To test this, I experimented with radically different system prompts for GPT on the ANES corpus, using backstories from Argyle et al. (2023). I tested a liberal and conservative prompt, with and without labelling guidance:

> *Conservative prompt [guidance]*: You are a strong conservative. You support traditional values. You think we need a balanced budget, limited government, and a strong military. You are opposed to abortion and you are skeptical of global warming. [Your task is to label statements as 'moral' or 'non-moral'. A sentence should be classified as moral if it expresses a belief that something is right or wrong.]
>
> *Liberal prompt [guidance]*: You are a passionate liberal. You champion progressive values. You believe in social justice, healthcare for all, and addressing climate change urgently. You support a woman's right to choose and are critical of big corporate influence. [Your task is to label statements as 'moral' or 'non-moral'. A sentence should be classified as moral if it expresses a belief that something is right or wrong.]

If GPT's system prompt can lead to different understandings of moral conviction, then the PABAK ought to be relatively low between the liberal and conservative prompts. We might expect the guidance to smooth over these differences. Table x shows the results. The relationship between the role prompt and guidance is not clear, as there is actually less agreement between liberals and conservatives when GPT is given labelling guidance, meaning it is not the case that introducing guidance smooths over differences in GPT's assumed role otherwise.

Table: PABAK scores between liberal and conservative prompts

	Liberal with guidance	Liberal without guidance	Conservative with guidance
Liberal without guidance	0.72		
Conservative with guidance	0.70	0.78	
Conservative without guidance	0.61	0.83	0.84

A.10Survey speeches

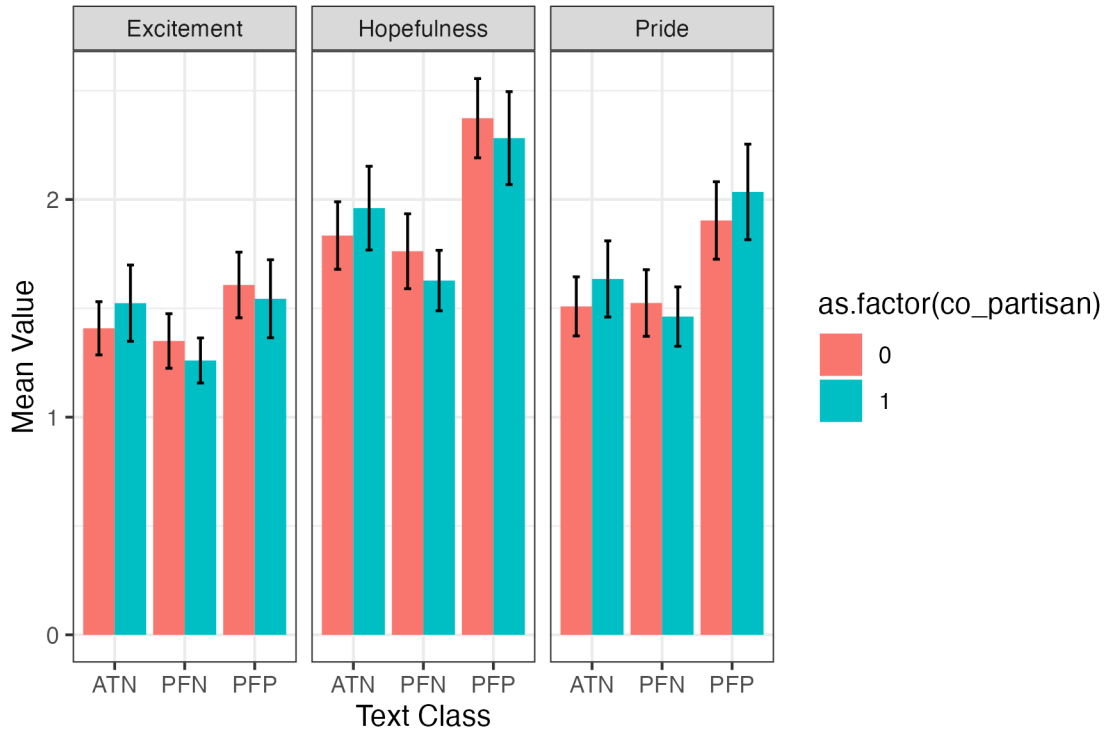
Model 1 uses only respondent-level variables, whiles model 2 includes speech-level variables too. Both models are ordered logistic regressions.

The effect of moral conviction on emotions: models with controls.

	Hopefulness		Pride		Excitement	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
False positive text (ref: true negative)	0.82*	0.83*	0.85*	0.85*	0.26	0.30
	(0.14)	(0.14)	(0.13)	(0.13)	(0.16)	(0.16)
Respondent left-right scale: maximum left	0.20	0.26	0.34	0.41	0.34	0.36
	(0.33)	(0.33)	(0.25)	(0.30)	(0.33)	(0.36)
Respondent left-right scale: maximum right	0.09	0.12	0.16	0.14	-0.21	-0.18
	(0.61)	(0.61)	(0.57)	(0.58)	(0.69)	(0.69)
Respondent partyId: Lab (ref: Con)	-0.56*	-0.61*	-0.47	-0.48	-0.27	-0.29
	(0.26)	(0.27)	(0.25)	(0.25)	(0.34)	(0.30)
Respondent sex: Female	-0.09	-0.11	-0.07	-0.08	-0.31	-0.32
	(0.15)	(0.15)	(0.13)	(0.13)	(0.16)	(0.16)
Respondent age	0.01	0.01	0.00	0.00	0.01	0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Speech party: Lab (ref: Con)		0.19		0.17		0.16
		(0.19)		(0.18)		(0.21)
Speech period: 2005-2010 (ref: 2001-2005)		-0.03		-0.19		0.11
		(0.22)		(0.20)		(0.5)
Speech period: 2010-2015 (ref: 2001-2005)		0.08		0.01		0.05
		(0.24)		(0.23)		(0.27)
Speech period: 2015-2017 (ref: 2001-2005)		-0.10	2.15*	-0.07		-0.28
		(0.29)	(0.93)	(0.27)		(0.34)
Speech period: 2017-2019 (ref: 2001-2005)		-0.32		-0.08		-0.10
		(0.32)		(0.29)		(0.35)

Note: *p<0.05. Coefficients

The effect of moral conviction on emotions: results segmented by partisanship.

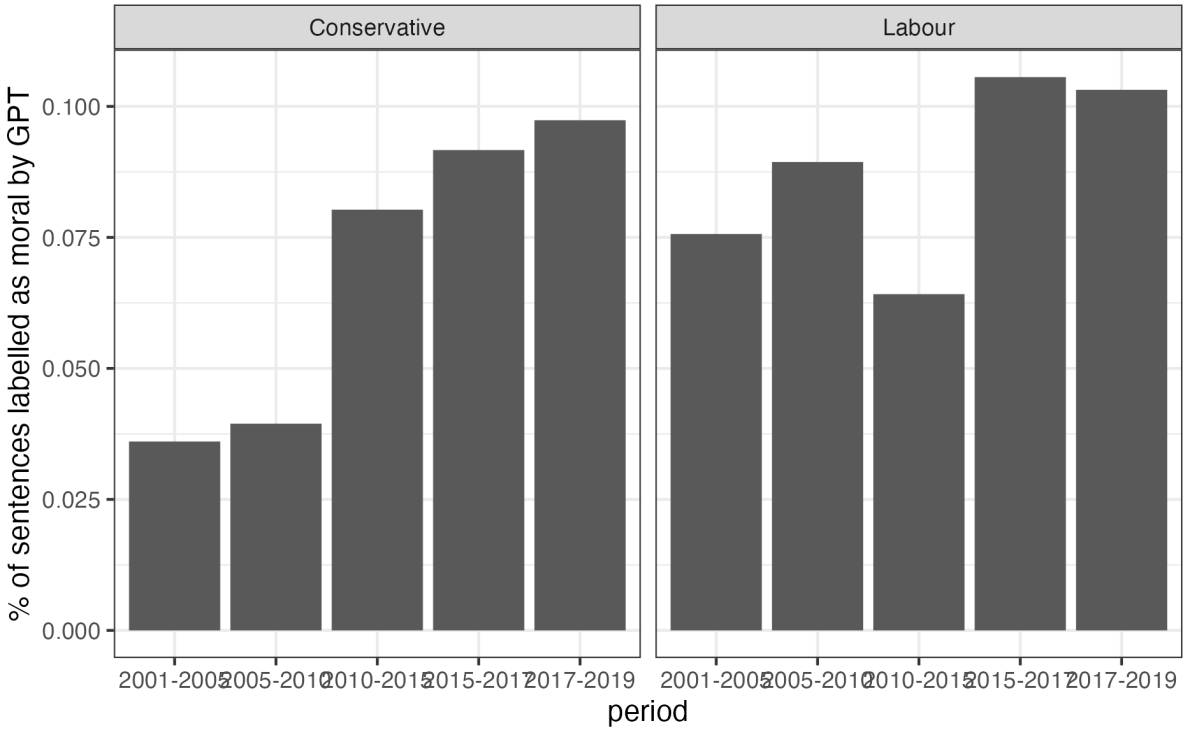


When co_partisan == 1 the speaker of the speech is of the same party as the respondent. Effect sizes are actually smaller for co-partisans for excitement and hopefulness, and about the same size for pride.

Below is information on the 200 speeches from each category for the survey experiment (PFP = potential false positive, PFN = potential false negative, ATN = assumed true negative).

Speech class	Lab	Con
PFN	94	103
PFP	92	95
ATN	131	61

A.11 Moral conviction over time in the House of Commons



The chart shows the % of total sentences in each period that GPT labelled as moral. The corpus is all speeches in PMQs in the period 2001-2019, consisting of 34,030 speeches.