Data-Driven Understanding of Real-Life Moral Dilemmas via Topic Mapping and Moral Foundations

Tuan Dung Nguyen

A thesis submitted for the degree of
Master of Philosophy at
The Australian National University

Australian
National
University

October 2023

Except where otherwise indicated, this thesis is my own original work.

Tuan Dung Nguyen
October 30, 2023

To my family

Acknowledgments

Of all the incredible individuals who deserve the deepest appreciation, I would first like to thank Lexing Xie, my primary supervisor for her guidance, passion and remarkably creative approach to problems in computational social science. These virtues continue to inspire me to this day. I am equally grateful for the mentorship by Colin Klein, my co-supervisor from whom I have learned so much about moral psychology and philosophy.

I have had many an invaluable opportunity to work closely with other research students at the Australian National University. In particular, Alasdair Tran, Minjeong Shin, Nicholas Carroll, Georgiana Lyall and Ziyu Chen have been direct collaborators of mine on this "Moral Dilemmas" research project. Lydia Lucchesi, Alexander Soen, Siqi Wu, Jooyoung Lee and Alexander Matthews at the Computational Media Lab have been a relentless source of support—academically and socially—for me in the last two years.

Ever since its early days, my M.Phil. project has greatly benefited from the Humanising Machine Intelligence Grand Challenge, a collective group of researchers from many schools and colleges at ANU. Jenny Davis, Hanna Kurniawati, Seth Lazar, Claire Benn, Pamela Robinson and Nick Schuster, among many others, have given me useful feedback and suggestions on many of my manuscripts and presentations. Fellow research students, especially Rachel Aalders, Jake Stone and Xueyin Zha, have pointed me to important resources and facilitated many insightful discussions. A special thank you to Chelle Adamson and Yixuan Lei for their organizing service throughout the years I have been with HMI.

Many other people at ANU have been instrumental in my journey so far. I am grateful for the consultations by Alice Richardson during one of our human studies. My experiences in teaching, which early on became a passion of mine, were made available by Paul Scott, Lexing Xie, Dawei Chen, Stephen Gould and Yuan-Sen Ting. My conference travel was partially funded by the ANU Vice-Chancellor's HDR Travel Grant.

I have had the privilege of working at several academic institutions during my five memorable years in Australia. Charl Ras, Pauline Lin and Chris Ewin were great mentors and colleagues of mine at the University of Melbourne. Together with Nguyen Tran, Canh Dinh, Tung Anh Nguyen and Long Le at the University of Sydney, I have made—I hope—useful contributions to federated learning, distributed computing and optimization at large.

Finally, my family and friends deserve to know their indispensable role in this rewarding journey. I thank my mother, father and brother for their never-ending belief in my potential, even in my not-so-rare moments self-doubt. My friends, especially Claire Chau Do, Thao Phuong Dao, Sarah Zylstra, Sonia Tam, Kevin Linarto, Katy Yixue Li, Liam Saliba, Kevin Yuxuan Yang, Beau Annoptham and Lauren Song—thank you for always standing by my side and helping me make a better version of myself every day.

Abstract

The emergence of artificial intelligence systems capable of engaging in complex discourse with humans presents both an opportunity and a challenge for automated moral decision-making. While the relevant philosophical literature has largely focused on analyzing idealized moral dilemmas, we instead aim to investigate the patterns of human moral judgment and sentiment observable in large-scale online datasets, which can provide insight into real-world ethical issues. This thesis presents two in-depth studies of moral discussions on social media.

First, we explore the features of everyday ethical conflicts through an analysis of 100,000 discussion threads on Reddit. Using a combination of topic modeling, human validation and crowd-sourced labeling, we discover a set of 47 topics that sufficiently describe the content on this forum. Despite their complex nature, the moral stories in this dataset can be represented by very nominally neutral topics such as *money*, *work*, *appearance* and *communication*, suggesting a nuanced view of morality in daily life. Importantly, people tend to perceive each moral story with two topics—like *family* and *money*—giving rise to a rich thematic space of over 1,000 topic pairs throughout this discussion sphere. Downstream results suggest that topics and topic pairs can serve as an important covariate in examining how a moral story is framed and how its corresponding judgment is made.

Second, we analyze moral controversies on social media through the lens of moral foundations theory, a taxonomy of five fundamental moral intuitions. This theory is widely used in data-driven studies of online content, but existing methods used to detect moral foundations are surprisingly lacking in their consistency and cross-domain generalizability. In response, we fine-tune a language model to measure moral foundations in text based on datasets covering news media and online discussions. The resulting model, which we call Mformer, consistently outperforms current approaches across several in- and out-of-domain benchmarks, improving from the state-of-the-art by up to 17% in the AUC metric. Using Mformer to analyze Reddit and Twitter content, we discover that the relative importance of moral foundations can meaningfully describe people's stance on many social issues, and such variations are topic-dependent.

Altogether, these studies demonstrate the utility of a data-driven approach to practical ethics. In particular, modern text corpora capture an unprecedented record of human discourse on contemporary social and ethical issues. A topic- or moral foundation-driven approach to analyzing these data sources can prove useful by representing complex moral discussions in a low-dimensional and interpretable manner. This may enable researchers to meaningfully examine the nuanced but diverse patterns of human moral judgment and sentiment in everyday life. Methods and findings from this thesis can be used to inform the development of automated systems which engage in content filtering, dissemination and moderation, and, ultimately, moral conversations with humans.

Publications, Software, and Data

Publications

- **Tuan Dung Nguyen**, Georgiana Lyall, Alasdair Tran, Minjeong Shin, Nicholas George Carroll, Colin Klein, and Lexing Xie. "Mapping Topics in 100,000 Real-Life Moral Dilemmas." In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 16 (2022): 699–710. (Chapter 2)

  https://doi.org/10.1609/icwsm.v16i1.19327

- **Tuan Dung Nguyen**, Ziyu Chen, Nicholas George Carroll, Alasdair Tran, Colin Klein, and Lexing Xie. "Measuring Moral Dimensions in Social Media with Mformer." Forthcoming in *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 18 (2024). (Chapter 3)

  https://doi.org/10.48550/arXiv.2311.10219

Source code

- `moral_dilemma_topics`: an implementation of topic modeling methods to analyze moral discussions on Reddit. (Chapter 2)

  https://github.com/joshnguyen99/moral_dilemma_topics

- `moral_axes`: measuring moral foundations in text with Mformer and other existing approaches. (Chapter 3)

  https://github.com/joshnguyen99/moral_axes

Data

- All posts and comments on the `r/AmItheAsshole` subreddit until 2020/04. (Chapters 2 and 3)

  https://doi.org/10.5281/zenodo.6791835

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Background

Within the vast landscape of human life in which intelligent systems may be involved, the domain of morality—contentious questions and answers about what is right and wrong—presents a particularly interesting case. Consider the situation put forth by Philippa Foot (1972) and later famously dubbed the "trolley problem" by Judith Jarvis Thomson (1976). An unstoppable trolley is about to run over five people down its track if there is no intervention. On the side of the track, a bystander can pull a lever to divert the trolley to another track with only one person, thereby making the trolley run over that person instead. Should the bystander pull the lever and kill the one person or do nothing and let the five people die? This killing-versus-letting-die dilemma remains one of the most influential thought experiments in philosophical ethics and directly challenges the theories that guide our action—most notably whether the net outcome of saving four lives is justified (a consequentialist view) or whether such an action is inherently morally forbidden regardless of how good the outcome is (a deontic view).

Real life, however, is much more complicated. The sort of ethical conflicts we face every day often lack the clarity, simplicity and starkness of classic dilemmas like the trolley problem. It is difficult to imagine, on a practical basis, a life-and-death situation like pulling a lever to save five lives while letting one die; yet, scenarios like whether to exclude one's atheist friends from one's religious wedding arise much more commonly, and corresponding actions will arguably elicit the kinds of praise or blame we can expect in a philosophical dilemma. It is in these moral conflicts, however, that we realistically make decisions that affect our well-being and that of others. Moreover, it is in these situations that artificial intelligence (AI) can be meaningfully deployed to understand and assist human decision making, if that is indeed a goal of AI development.

Despite this, our current empirical understanding of moral life remains largely unexplored in several important ways. Classical dilemmas, which are aimed at teasing out the mechanisms in our ethical reasoning, are often too idealized and rare. The psychology literature has presented a multitude of studies on moral judgments, attitudes and intuitions; yet its data is often limited to laboratories or online surveys whose participating population is only up to the thousands. The development of AI systems—particularly "generative models" capable of making human-like responses like a normative judgment—often ignores the highly nuanced nature of moral conflicts experienced by laypeople daily. Overall, it is difficult to say exactly *where* moral dilemmas arise and *how* they are presented and deliberated among people

without a systematic account covering a sizable portion of the population at large.

This thesis examines the features of everyday moral dilemmas through a series of large-scale empirical studies of social media content. Online discussion forums offer a unique opportunity to study many kinds of social phenomena, of which moral debates are a growing source. They also come with a characteristic challenge: massive datasets collected from social media require scalable computational tools to extract useful patterns while accounting for a level of noise rarely seen in data collected in laboratory settings. In particular, we will present two main themes of investigation in the following chapters of this thesis. First, we conduct a content analysis of over 100,000 moral stories—the largest collection at the date of publication—and profile a number of salient moral topics frequently discussed by real internet users. Then, we adopt a popular framework for classifying moral intuitions to characterize the patterns of story framing and judgment, revealing novel insights about how people perceive and participate in real-life moral debates and how these judgments vary throughout their topical spheres.

1.1 Moral Dilemmas in Every Life

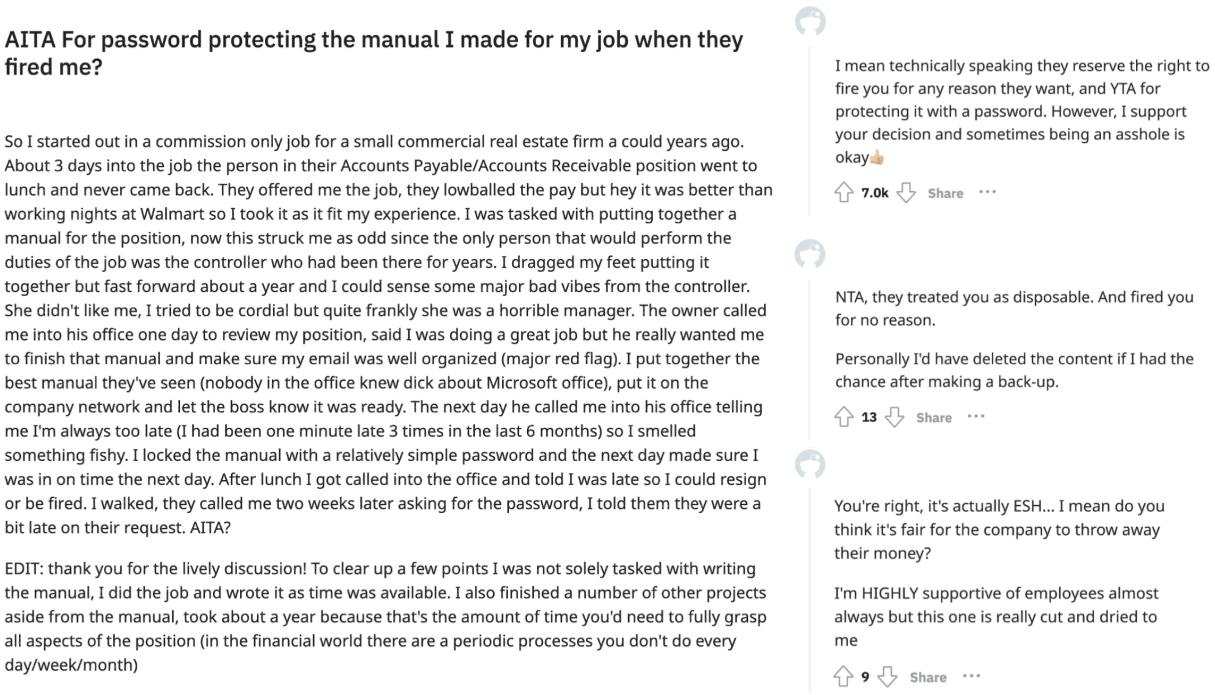
Many works in moral philosophy and moral psychology have extensively profiled a range of ethical issues and the theoretical foundations on which these conflicts may be resolved. Classic thought experiments that capture public imagination are pervasive: Should we sacrifice the life of one person to save the lives of five others? Which patient should be prioritized in getting a kidney transplant? By design, these dramatic and idealized moral dilemmas make the conflict between moral principles especially stark. Daily life, however, presents people with a wide variety of comparatively small-scale, low-stakes and messy moral conflicts. Few of us will allocate a kidney or sacrifice strangers, but nearly everyone will have to deal with uncomfortable in-laws at a wedding, or adjudicate bitter debates over the workplace fridge. Yet, even though these conflicts are equally important for understanding moral life, they remain under-studied because they lack the clarity of idealized dilemmas.

This thesis aims to fill this crucial gap in philosophical and empirical inquiries into moral dilemmas. Here, we will adopt the term "moral dilemma" in a broader and non-traditional sense, one which Driver (1992) calls a "morally charged situation":

> Morally charged situations resemble moral dilemmas, because whatever decision you make, you cannot back out of the situation into a morally free space. There is no neutral ground. In the ethicist's technical sense a moral dilemma is a situation in which one is required to do x, and required to do not-x. Thus, whatever one does is wrong. In a morally charged situation, however, one's options are between two alternatives, neither of which is required – one option carries approval, the other disapproval.

Morally charged situations, in which one action will elicit praise and the other will elicit blame, arguably well resemble such day-to-day conflicts. They have attracted less attention but are equally important for understanding moral life.

The subject of study in this thesis is moral content on social media. Platforms such as Reddit and Twitter—in addition to the news media—record an unprecedented amount of on-



AITA For password protecting the manual I made for my job when they fired me?

So I started out in a commission only job for a small commercial real estate firm a could years ago. About 3 days into the job the person in their Accounts Payable/Accounts Receivable position went to lunch and never came back. They offered me the job, they lowballed the pay but hey it was better than working nights at Walmart so I took it as it fit my experience. I was tasked with putting together a manual for the position, now this struck me as odd since the only person that would perform the duties of the job was the controller who had been there for years. I dragged my feet putting it together but fast forward about a year and I could sense some major bad vibes from the controller. She didn't like me, I tried to be cordial but quite frankly she was a horrible manager. The owner called me into his office one day to review my position, said I was doing a great job but he really wanted me to finish that manual and make sure my email was well organized (major red flag). I put together the best manual they've seen (nobody in the office knew dick about Microsoft office), put it on the company network and let the boss know it was ready. The next day he called me into his office telling me I'm always too late (I had been one minute late 3 times in the last 6 months) so I smelled something fishy. I locked the manual with a relatively simple password and the next day made sure I was in on time the next day. After lunch I got called into the office and told I was late so I could resign or be fired. I walked, they called me two weeks later asking for the password, I told them they were a bit late on their request. AITA?

EDIT: thank you for the lively discussion! To clear up a few points I was not solely tasked with writing the manual, I did the job and wrote it as time was available. I also finished a number of other projects aside from the manual, took about a year because that's the amount of time you'd need to fully grasp all aspects of the position (in the financial world there are a periodic processes you don't do every day/week/month)

I mean technically speaking they reserve the right to fire you for any reason they want, and YTA for protecting it with a password. However, I support your decision and sometimes being an asshole is okay👍

7.0k Share ···

NTA, they treated you as disposable. And fired you for no reason.

Personally I'd have deleted the content if I had the chance after making a back-up.

13 Share ···

You're right, it's actually ESH... I mean do you think it's fair for the company to throw away their money?

I'm HIGHLY supportive of employees almost always but this one is really cut and dried to me

9 Share ···

Figure 1.1: An example discussion thread on `r/AmItheAsshole`, a Reddit forum ("subreddit") where users post about their interpersonal moral conflicts and ask the community to judge their actions. This subreddit is the main subject of study in Chapter 2. Left: the original post with its title in bold, followed by body text. Right: three different judgments expressed in the comments below the post. The acronyms used in these judgments are `YTA` ("You're the AH"), `NTA` ("Not the AH") and `ESH` ("Everyone Sucks Here"). Below each comment is its score, equal to the number of upvotes minus downvotes. Usernames and profile pictures have been hidden. The entire post and all of its comments can be found at https://www.reddit.com/r/AmItheAsshole/comments/doknyz/aita_for_password_protecting_the_manual_i_made/.

line human interaction including, specifically, debates on moral issues. These discussions are much more diverse in topic than popular moral dilemmas like the trolley problem, encompassing interpersonal conflicts, political debates and social controversies, among others. They are also much more sophisticated in nature, often involving many parties and complex details elaborated by internet users. Figure 1.1 on page 3, which shows an example discussion thread on the Reddit forum `r/AmItheAsshole` (to be studied extensively in Chapter 2), illustrates this observation. Moreover, unlike classic moral dilemmas, real-life stories reflect the sort of conflicts people actually experience, which, in turn, will ultimately be the subject of moral evaluation by AI systems.

1.2 Empirical Studies of Moral Dilemmas and Judgments

Going beyond the subject of theorization, moral dilemmas have been a topic of empirical investigation in several disciplines such as social psychology and cognitive science. One of the ultimate contributions of such works is to machine ethics, an area concerned with designing automated agents—driverless cars and consulting chatbots, to name a few—that engage in morally charged situations and that may be tasked with making moral decisions. Large-scale studies of moral decision-making have the potential to offer a global perspective on moral preferences, enable the examination of cross-cultural variations in moral decision-making, and help identify common expectations about how to guide machine behavior.

A particularly influential line of work is the Moral Machine experiment (Awad et al., 2018), an investigation of moral preferences in trolley problem-like dilemmas across more than two million human participants. The study profiled global and cultural preferences in this moral dilemma over many dimensions such as age, gender, social status, and number of agents involved, which have important implications for the design and deployment of autonomous vehicles.

Another classic example of moral dilemma that has been studied empirically is kidney exchange. Freedman et al. (2018) considered the problem of estimating the "weights"—the importance of each factor, such as age, in each patient—in a kidney exchange market. These weights were then used for a market-clearing algorithm, one that aims to maximize the proportion of matched donor-patient pairs in a simulated population. This study presents a useful proof of concept for moral dilemmas of the sort, and provides a framework for eliciting moral judgments from human subjects to guide AI systems.

Building and probing systems that can engage with humans in conversations involving moral dilemmas is a growing area of research. These systems—often backed by advanced models in natural language processing—are not yet versed to "reason" about ethical norms, but they reflect a great deal of their training data, which records a massive number of human moral judgments. For example, Delphi (Jiang et al., 2022) is a research prototype that takes in a one-line natural language snippet and gives a moral judgment from a wide range of possibilities (e.g., expected, understandable, wrong, bad, rude, disgusting). Built as a model for commonsense moral reasoning (Choi, 2022), Delphi demonstrates good performance in assigning a moral judgment to many novel scenarios, although it has at times been shown to generate unfiltered and inconsistent responses, a common problem for many neural models.

To train and benchmark AI agents capable of engaging in morally charged situations, researchers have relied on datasets that capture moral stories and judgments across several domains. For example, the Social-Chem-101 dataset (Forbes et al., 2020) and Moral Integrity Corpus (Ziems et al., 2022) contain many textual examples of "rules-of-thumb," which are short statements reflecting a social norm such as "It is rude be selfish." Emelin et al. (2021) proposed Moral Stories, a dataset containing actions and consequences relating to specific norms and intentions to assess language models' ability to conform to predefined normative constraints. The Social Bias Inference Corpus (Sap et al., 2020) contains a large number of posts from social media annotated with implied bias relating to, for example, racism. Lourie et al. (2021) composed Scruples, a dataset of ethical anecdotes and their corresponding judgments posted by Reddit users. Finally, Hendrycks et al. (2021) proposed the ETHICS dataset for benchmarking large language models on a range of concepts such as commonsense morality and justice.

The majority of prior work that aims to evaluate moral judgments generated by AI has focused exclusively on language models—machine learning models typically trained on massive datasets to perform a range of tasks such as text generation and classification. For example, Zhou et al. (2021) and Haworth et al. (2021) built classifiers to assign a moral judgment to a textual input describing a moral conflict as posted on Reddit. In particular, Zhou et al. (2021) found that certain patterns in language usage, such as utilizing the first-person passive voice to describe the role of a victim, can reliably predict whether a post will receive a negative judgment. The Delphi prototype (Jiang et al., 2022), described above, is built based on a language model fine-tuned using several datasets including Scruples to generate moral judgments for short-form action descriptions. Large language models, such as GPT-3 (Brown et al., 2020), have been evaluated against a suite of benchmarks for ethical decision-making (Hendrycks et al., 2021; Krügel et al., 2023). Using the ETHICS dataset, for example, Hendrycks et al. (2021) found impressive capabilities of GPT-3 and other language models in a range of commonsense morality tasks, but noted their incomplete ability to estimate basic moral judgments.

This thesis, also presenting investigations of everyday moral conflicts, aims to fill two crucial gaps in this empirical literature. First, although data on moral dilemmas and judgments, especially generated on social media, has been growing, there exists no systematic account of where these stories emerge and how they are discussed and deliberated by online users. We provide a data-driven approach combining computational modeling and a two-stage human validation process to analyze this data, revealing the most common topics of discussion and how the patterns of framing and judgment vary within these topics. This methodology, to our knowledge, is novel and leads to high-quality, interpretable results for a range of downstream analyses.

Second, the empirical literature on moral dilemmas often relies on theoretical developments for its measurements; however, existing theory-driven computational tools to analyze this data often are brittle, lack human consensus and do not generalize well to novel data. Adopting a common framework called moral foundations theory—which represents the space of moral concerns in terms of five seemingly universal moral intuitions: *authority*, *care*, *fairness*, *loyalty* and *sanctity*—we extensively review and critique approaches that aim to detect foundations in text data. In response to their limitations, we propose a new machine learning-based model called Mformer that achieves state-of-the-art performance across a range of benchmarks

and does not suffer from the biases known in previous methods. We finally demonstrate the utility of this model in some case studies, some of which suggest that results reached from previous work may be subject to revision due to its use of unreliable tools.

1.3 Thesis Outline

It is based on large-scale datasets capturing moral debates by humans that we aim to examine the characteristics of everyday moral life. Specifically, we provide two in-depth studies of social media content through the lens of topic modeling and moral foundations, revealing non-trivial patterns of morality pertaining to story framing, judgments and moral stance across a wide variety of socially relevant topics. The rest of the thesis is organized as follows.

In Chapter 2, we address the question of what discussion themes are prevalent in everyday moral dilemmas and how moral valence varies across these topics. After curating a dataset of over 100,000 threads from the `r/AmItheAsshole` subreddit—the largest collection at the time of writing—we perform a rigorous analysis including topic modeling, human validation and crowd-sourced labeling to map this large domain into 47 interpretable topics. We find that the moral stories discussed in this forum present a nuanced view of morality by the topics from which they arise. In particular, most stories can be represented by very nominally neutral topics such as *money*, *work*, *appearance* and *communication*. Moreover, people tend to perceive each story with two topics—like *family* and *money*—giving rise to a rich thematic space of over 1,000 topic pairs throughout this discussion sphere. Using this result to further analyze this forum, we discover interesting patterns of moral framing and judgment, such as the observation that the rate at which an author is judged to be in the wrong is dependent on the topics that the story is about. We conclude this chapter by positing that topics and topic pairs can serve as an important covariate in examining how a moral dilemma is framed and how its corresponding judgment is made.

In Chapter 3, we examine moral foundations theory (Haidt and Joseph, 2004; Haidt, 2013), a popular framework widely adopted in computational social science, in uncovering important aspects of morality in daily life. We start by highlighting the limitations of existing computational tools used to detect moral foundations in text: Particularly, they are surprisingly lacking in their consistency and cross-domain generalizability. Advances in natural language processing, most notably in the development of powerful language models suitable for text classification, are a promising approach in this direction (Trager et al., 2022; Liscio et al., 2022; Guo et al., 2023). We present Mformer, a language model fine-tuned to measure moral foundations based on datasets covering news media and short- and long-form online discussions. Through an extensive suite of benchmarks containing both in- and out-of-domain data, we show that Mformer consistently improves from current approaches, including the state-of-the-art, by up to 17% in the AUC metric. Next, using Mformer to analyze Reddit and Twitter content on moral dilemmas and controversies, we find that moral foundations can meaningfully describe people's stance on many social issues, and such variations are topic-dependent. We conclude by reiterating the efficacy of this theoretical framework in studying a wide range of social phenomena, while emphasizing that the tools used to detect moral foundations must be chosen carefully to ensure the validity of downstream measurements

and conclusions. Finally, we release Mformer publicly as a better alternative for researchers who aim to incorporate moral foundations in their studies.

Finally, in Chapter 4, we summarize our work and discuss a range of future directions.

1.4 Key Contributions and Impact

A common theme throughout this thesis is a data-driven approach to understanding human moral life. As a result, our contributions reflect a meaningful interplay between theoretical foundations—mostly developed in traditional disciplines such as philosophy, sociology and psychology—and advances in computer science.

First, we introduce datasets and computational tools for studying everyday moral dilemmas:

- We curate a dataset of over 100,000 discussion threads containing over 8 million comments from the `r/AmItheAsshole` subreddit. This is among the largest collections of everyday moral dilemmas to date and consists of high-quality moral stories and judgments.
- In exploring a large domain of moral discussions, we present a novel topic discovery method with multiple stages of validation which results a set of meaningful, interpretable and well-verified topics. The resulting topic model is available online.
- In measuring moral foundations in text, we highlight the conceptual and practical limitations of existing tools, which can seriously affect downstream measurements and conclusions. We introduce and publicly release Mformer, an alternative language model fine-tuned on diverse datasets to recognize moral foundations that is demonstrated to consistently outperform current approaches on a range of benchmarks.

Second, we provide novel empirical insights into everyday moral conflicts through the lens of topic modeling and moral foundations such as:

- The sorts of moral dilemmas described and debated online may arise from very nominally neutral topics like *communication* and *family*, presenting a nuanced view of morality by laypeople.
- Most moral stories involve two nominal topics, and the thematic unit of topic pair can serve as an important factor in examining several aspects of moral judgment, such as the rate at which a moral story is judged negatively.
- The relative importance of moral foundations can explain the variations in people's moral stance on many social issues, such as the legalization of abortion, as such variations are topic-dependent.

Moral dilemmas play a crucial role in philosophical theorizing. Part of the motivation of this thesis is an empirical interest in the traditional moral-conventional distinction in ethical conflicts (Foot, 1972; Southwood, 2011), which may well be challenged by the data we have on everyday moral dilemmas. We present an examination of the morally charged situations

that laypeople experience in their daily lives, which differ from traditional, classic thought experiments and laboratory vignettes in several major ways but are equally important in understanding human morality. The observations on topic pairs in many moral discussions can help map out the space of morality in a more nuanced yet tractable manner, facilitating the examination and comparison of moral framing and judgment across many relevant thematic spheres. The findings using moral foundations illuminate the usefulness of current theoretical developments in characterizing moral controversies and people's moral stances on many contemporary social issues. It is our hope that this thesis provides valuable grounds on which researchers can build to further explore the intersection of practical and theoretical ethics, especially relating to the development of human-centered AI systems.

Several following lines of research have taken shape based on the contributions of this thesis. The topic-mapping approach to characterizing moral stories, for example, has been adopted by Xi and Singh (2023b) in their study of participants' self-reported gender on AITA. After identifying and labeling a number of "meaning clusters"—such as *judgment of appearance* and *law and order*—found to correlate with moral judgment, the authors found consistent associations of gender with these meaning clusters and such observations vary with the discussed topic. For instance, among posts belonging to the topic *safety*, female authors receive significantly more comments classified under *judgment of appearance*, whereas the same effect is observed for male authors under *law and order*.

Other following works have taken the *final judgment* of a post—see Figure 1.1 above for an example—as a supervised signal to study the effect of several linguistic and social features on moral sentiment. Among these features are topics: Xi and Singh (2023a) used the same computational approach as ours, albeit with a less extensive human validation procedure, to identify topics on `r/AmItheAsshole`. A statistical analysis found that some of these topics are highly correlated with the valence of a post, i.e., whether the author is judged to be in the wrong. For instance, while the presence of most topics is found to increase the odds of a negative judgment, some topics such as *school*, *holiday* and *gifts* have the opposite effect.

In another line of work, Giorgi et al. (2023) studied the differences in the role of an `r/AmItheAsshole` author—either as the *narrator* or a *character* in their story. After profiling a range of features related to demographics, sentiment and emotion, among others, the authors found that there exist posts whose topical content is highly associated with some moral judgment. An analysis of unigrams, for example, revealed that when a post's author acts as the narrator, they are more likely to be judged negatively if their story is about relationships with women. Although it was not confirmed by Giorgi et al., this effect may be more significant if the narrator is constrained to be male and older, suggesting a possibility of sexist undertones in their stories.

In Chapter 4, we further discuss the directions worthy of exploration in future work.

Chapter 2

Mapping Topics in 100,000 Real-Life Moral Dilemmas

2.1 Introduction

Should we sacrifice the life of one person to save the lives of five others? Which patient should be prioritized in getting a kidney transplant? The idealized moral dilemmas that capture public imagination are clear and dramatic. This is by design. Thought experiments like the trolley problem (Thomson, 1976) make the conflict between moral principles especially stark. Daily life also presents people with a wide variety of comparatively small-scale, low-stakes, messy moral dilemmas. These remain under-studied because they lack the clarity of idealized dilemmas, yet they are arguably the sort of dilemmas that preoccupy most people most of the time.

Philosophers define a moral dilemma as a situation in which an agent has a moral duty to perform two actions but can only perform one of them (Sinnott-Armstrong, 1988). Here we will use the term in a broader and non-traditional sense, encompassing *inter alia* what (Driver, 1992) calls a "morally charged situation." This is a situation in which an agent is faced with a non-obvious choice between performing one of two actions, neither of which is morally required, but where one will elicit praise while the other will elicit blame. Such situations have attracted less attention, but are equally important for understanding moral life.

In this work, we investigate *moral dilemmas that arise in daily life*. A broad study of such dilemmas will help to fill a crucial gap in philosophical and empirical inquiries into moral dilemmas. Few of us will allocate a kidney or sacrifice strangers; nearly everyone will have to deal with uncomfortable in-laws at a wedding, or adjudicate bitter debates over the workplace fridge. Many moral conflicts arise in pedestrian contexts from familiar concerns. A better understanding of everyday moral dilemmas will provide a novel foundation for testing philosophical and social scientific theories about the nature and taxonomy of our moral judgments such as the moral foundations theory (Graham et al., 2011), morality as cooperation (Curry, 2016), or forms of moral particularism (Dancy, 1983; Kagan, 1988). Our analysis also shows that many moral dilemmas result from the interaction of what are traditionally considered conventional norms, suggesting that the moral/conventional distinction is less stark than some have supposed. Finally, a better understanding of everyday moral dilemmas could help shape the design of next-generation AI systems that are capable of fluid interaction in

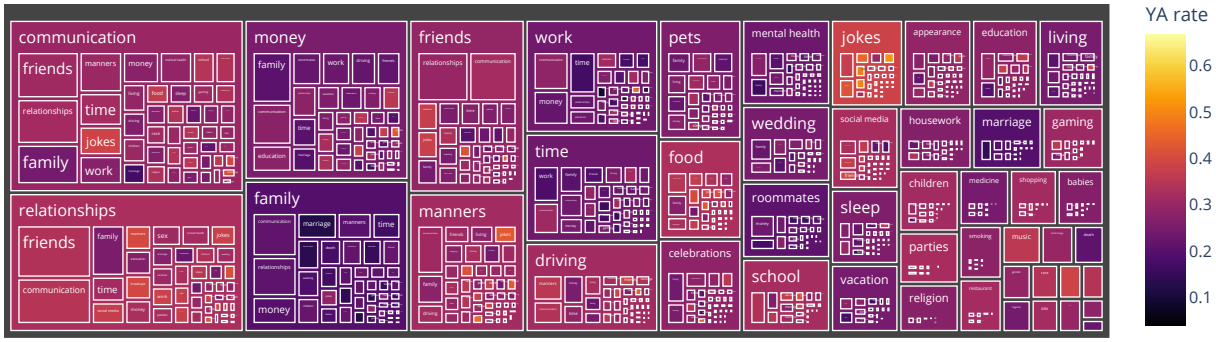


Figure 2.1: Treemap showing the top-1 (outer rectangles) and top-2 (inner rectangles) topics discovered by latent Dirichlet allocation (LDA) on the AITA dataset. The size of a block corresponds to the number of posts in a topic or a topic pair. *Communication* and *relationships* are the two most prevalent topics. A lighter color corresponds to a higher proportion of `YA` judgments (*you are the asshole* or *everyone sucks here*) within a topic/topic pair. For example, posts about *family* receives mostly a *not the asshole* judgment.

complex human environments.

To this end, we turn to Reddit, a social network on which members can rate and discuss content submitted by other members. Reddit consists of user-created communities called *subreddits*, each of which focuses on a single topic of discussion. The `r/AmItheAsshole` (AITA) subreddit allows members to describe a non-violent moral conflict that they have recently experienced, and ask the community to decide if they were in the right. AITA is, as put by its community, *"a catharsis for the frustrated moral philosopher in all of us"*.[1] It is a popular subreddit: at the time of writing, it has over 3 million members[2] and regularly ranks in the top 10 for volume of comments per day.[3] This makes it an excellent source for real-life moral dilemmas and the discussion that surrounds them.

We extract more than 100,000 real-life moral dilemmas on AITA, and design a multi-stage topic discovery process using both expert and crowd-sourced validation to map 94% of these scenarios into 47 interpretable topics. We posit that topics are an informative lens through which to study AITA. Our rigorous discovery and validation process is designed to eliminate ambiguities that will propagate to subsequent analysis. These topics need not be mutually exclusive, and indeed the richness of content of AITA dilemmas means that most are better characterized by a topic pair instead of a single topic. As Figure 2.1 shows, AITA topic pairs vary both in popularity and in the judgments they attract. Many AITA dilemmas involve traditionally non-moral domains, suggesting a more nuanced structure than those of traditional philosophical thought experiments.

The main contributions of this work include:

- Curating a large collection of everyday moral dilemmas, which is publicly released;[4]

[1] https://reddit.com/r/AmITheAsshole

[2] A member of a subreddit is a user who subscribes to the subreddit. Active users who post or comment are typically a proper subset of all members.

[3] According to https://subredditstats.com

[4] The dataset and code can be found at https://github.com/joshnguyen99/moral_dilemma_topics

- A novel data-driven topic discovery method with multiple stages of validation to map these dilemmas into five meta-categories spanning 47 meaningful topics;
- Demonstrating ways that an understanding of everyday moral dilemmas can produce new insights into philosophical discussions relating to moral theorizing; and
- Empirical insights showing how everyday moral dilemmas are generated by combinations of topic pairs, how certain topics attract or repel other topics, and how the moral valence of similar words can vary across different topic pairs.

2.2 Related Work

This work is related to the rich literature on moral dilemmas, topic modeling and discovery, and online collective judgment and decision making.

2.2.1 Moral dilemmas

Moral dilemmas (Sinnott-Armstrong, 1988) and morally charged situations (Driver, 1992) play a crucial role in philosophical theorizing. There is an empirical literature aimed at teasing out the mechanisms that drive individual judgments about classic dilemmas (Greene et al., 2001); this work has become increasingly important in informing moral domains like algorithmic decision-making systems such as driverless cars (Awad et al., 2018) or kidney exchange programs (Freedman et al., 2018).

We note three features that characterize much of the existing empirical work and set it apart from the current study. First, existing work tends to focus on stark dilemmas—like the so-called "trolley problems" (Thomson, 1976)—that require individuals to pass judgment on unfamiliar and unrealistic situations. Second, existing work tends to rely on survey data or laboratory experiments rather than conversations with peers. Moral judgment and justification are sensitive to perceived beliefs and intentions of one's audience, including the experimenters themselves (Tetlock, 1983). Hence such settings may not reveal the full range of the participants' reasoning. Observational posts of online social media represent the sort of "unobtrusive measure" (Webb et al., 1999) that can avoid experimenter effects. Third, existing work tends to give subjects pre-packaged, simple moral dilemmas. Yet figuring out how to frame a moral problem in the first place is often an important issue in its own right (Appiah, 2008). By contrast, AITA represents a rich source of moral dilemmas that are realistic and familiar, presented by an involved party as part of a conversation with peers, and in a forum that allows for dynamic probing and re-framing the issues at hand. The AITA dataset thus represents a valuable resource for studying moral dilemmas and crowd-sourced judgments, one that can complement existing hypothesis-driven work.

2.2.2 Topic modeling in text

The task of understanding large document collections is sometimes referred to as 'describing the haystack'. Data clustering approaches are widely used for such problems. Methods that are specifically designed for text data include Probabilistic Latent Semantic Indexing (PLSI)

(Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003). In particular, LDA has been widely applied to historical documents, scientific literature, and social media collections (Boyd-Graber et al., 2017), to name a few.

We categorize the evaluation of topic models into intrinsic and extrinsic methods. Intrinsic methods evaluate components of the topic models themselves. Hold-out data likelihood (Blei et al., 2003) has been the *de facto* choice when evaluating an entire LDA model intrinsically. For each topic, human-in-the-loop approaches with intruding words (Chang et al., 2009), or coherence metrics based on the probability of word co-occurrences (Mimno et al., 2011) have been shown to correlate with human judgments. Extrinsic methods evaluate topic models with respect to domain-specific tasks; examples of these are as diverse as the application domains. In the scientific literature, topic model outputs have been compared against surrogate ground truths such as author-assigned subject headings, and used for trend spotting over time (Griffiths and Steyvers, 2004). In analyzing historical newspapers, Newman and Block (2006) annotated a subset of topics of interest to history and journalism but were not concerned with either covering the whole dataset or ensuring most topics are meaningful. In literature, derived statistics from topics have been shown to evaluate specific conjectures about gender, anonymity, and literary themes (Jockers and Mimno, 2013). Recently Antoniak et al. (2019) used LDA to discover narrative paths and negotiation of power in birth stories. Their dataset is much smaller (2.8K) and more topically concentrated than the AITA dataset used in this work. Also, their topics are validated using an existing medical taxonomy, whereas there is no such resource for everyday moral conflicts.

What differentiates this work in the application of topic models are the striving for coverage of a large collection, and the goal of supporting both qualitative and quantitative tasks. To the best of our knowledge, the two-stage validation combining the opinions of experts and general users is new.

2.2.3 Moral judgments on social media

This topic area is quickly gaining momentum in computational social science. Two recent works are focused on analyzing language use in moral discussions. Zhou et al. (2021) profiled linguistic patterns in relation to moral judgments, showing that the use of first-person passive voice in a post correlates with receiving a not-at-fault judgment. Haworth et al. (2021) called the judgment on a post "reasonability" and built machine learning classifiers to predict the judgments using linguistic and behavioral features of a post. Other works are focused on automated prediction of moral judgements. Botzer et al. (2022) built a moral valence (`YTA` and `NTA`) classifier on AITA data and evaluated its utility on other relevant subreddits. Delphi (Jiang et al., 2022) is a research prototype that takes in a one-line natural language snippet and gives a moral judgment from a wider range of possibilities (e.g., *expected, understandable, wrong, bad, rude, disgusting*). Its large-scale neural model is trained on multiple data sources including parts of AITA. The related Social Chemistry project (Forbes et al., 2020) breaks down judgments of one-liner scenarios into rules of thumb, covering social judgments of good and bad, moral foundations, expected cultural pressure and assumed legality.

While recent work focuses on directly correlating the natural language content (of a post, a title snippet, or a comment) with moral judgments, we choose to focus on taxonomizing the

structure of moral discussions as a first step. We posit that there are diverse practices used by the online community in moral argument and reaching a verdict as a group. This hypothesis is supported in Section 2.6, showing that topics are an important covariate for the differences in the moral foundation to which posters appeal.

2.3 Dataset

2.3.1 Structure of `r/AmItheAsshole`

In a subreddit, discussions are organized into *threads*. Each thread starts with a *post*, followed by comments. Each post consists of a *title, author, posting time*, and *content*; and each comment contains an *author, timestamp, content*, and *reply-to* (the ID of a post or another comment). Community rules dictate that a post title must begin with the acronym 'AITA' or 'WIBTA' (Would I Be The Asshole?).

Collective judgments are reached via *tagging* and *voting*. Five types of judgments are defined in AITA: `YTA` (you are the asshole), `NTA` (not the asshole), `ESH` (everyone sucks here), `NAH` (no asshole here), and `INFO` (more information needed). Each comment can contain one of these tags. A user can cast an upvote (scoring $+1$) or a downvote (scoring -1) to a comment. The judgment of the top-scoring comment would become the community verdict, called *flair*, and be displayed as a tag for the post.[5] The *flair* of a post is assigned by a bot after 18 hours.[6] Figure A.1 in Appendix A (page 57) shows an example thread with the `YTA` flair, and another comment judging it as `NTA`.

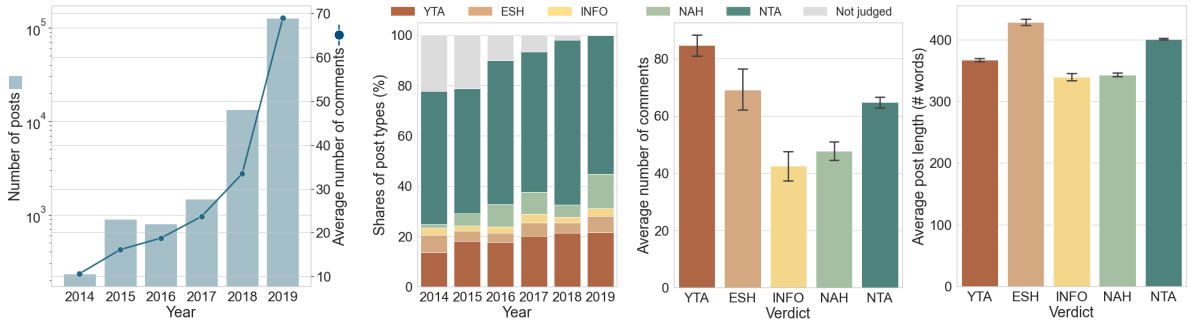
2.3.2 The AITA dataset

We use the Pushshift API (Baumgartner et al., 2020) to retrieve all posts and comments on AITA from 8 June 2013 to 30 April 2020, yielding 148,691 posts and 18,533,347 comments. When a post's flair maps to a judgment (such as `NTA`), we use it as the post's verdict. In the 946 posts without a valid flair, we reconstruct each post's verdict using the judgment contained in its highest-scoring comment. After this, 920 posts remain without flairs. To filter out moderation and meta posts, we keep posts whose titles start with "AITA" or "WIBTA," have at least 50 words, 10 comments, 1 vote, and 1 verdict. This yields 108,307 posts and 8,953,172 comments. Posts with fewer than 10 comments consist only of 20% of the dataset and are generally of lower quality. We use the 102,998 threads in or before 2019 as our training set, and 5,309 threads in the first four months of 2020 as the test set for the topics discovered (Section 2.5). When pair-wise comparison is called for, we group `NTA` and `NAH` into the `NA` judgment class with *positive valance* on the original poster. Similarly `YTA` and `ESH` are grouped into the `YA` class with *negative valance*.

We note that works using the Pushshift Reddit API that were published before 2018 may have involved missing data, which can lead to systematic biases in downstream analyses

[5] https://mods.reddithelp.com/hc/en-us/articles/360010513191-Post-Flair

[6] This timeframe was chosen by the community. The full process is documented in the AITA community FAQ https://www.reddit.com/r/AmItheAsshole/wiki/faq.



(a) (b) (c) (d)

Figure 2.2: Data statistics of AITA, 2014–2019. (a) Number of posts (bar) and average number of comments per post (line) by year. (b) Shares of judgment types (`YTA`, `ESH`, `INFO`, `NAH`, `NTA` or no verdict) by year. (c) Average number of comments (and 95% CI) per post, broken down by the posts' verdicts. (c) Average number of words per post (and 95% CI), broken down by the posts' verdicts.

(Gaffney and Matias, 2018). However, Baumgartner et al. (2020) have since recrawled the missing posts and thus our derived data is less likely to suffer from the same problem.

Figure 2.2 presents the number of posts, number of comments per post and breakdown of flairs. Over time, participation increased quickly as more members entered the subreddit. Both the number of posts and the average number of comments per post rose over the years, with 2018 and 2019 seeing the most significant increases (Figure 2.2a; note the y-axis in logarithmic scale). The flair shares (Figure 2.2b) remained consistent in 2018 and 2019, with `NTA` posts taking more than half of the posts (65.32% in 2018 and 55.14% in 2019). In terms of controversiality (Figure 2.2c), negatively judged posts (with flair `YTA` or `ESH`) tend to attract more comments than positively judged posts (with flair `NTA` or `NAH`). When looking at the post lengths (Figure 2.2d), `ESH` posts are the longest on average (mean = 433.2 words), reflecting the nuances required when describing situations with no clear winner. We also observe that `NTA` posts tend to be longer than `YTA` posts (`NTA`: mean = 400.6; `YTA`: mean = 370.6), while `YTA` posts attract more comments overall (`NTA`: mean = 79.4; `YTA`: mean = 107.6).

2.4 Discovering topics on AITA

We adopt a data-driven topic discovery process with two stages of manual validation, as outlined in Figure 2.3. An exploratory study that shaped our clustering choices is described in Appendix A.2 (page 56). Taking as input the 102,998 posts until the end of 2019 as the training set, we use text clustering algorithms to group the collection into clusters and describe their properties.

Clustering methods discover self-similar groups in data, called clusters. Given the goal of mapping different kinds of moral dilemmas on AITA, the ideal set of clusters should have a high *coverage* of the whole dataset, and the clusters (and posts within) should be *distinguishable* from each other as judged by human readers. Our choices of which clustering methods to use

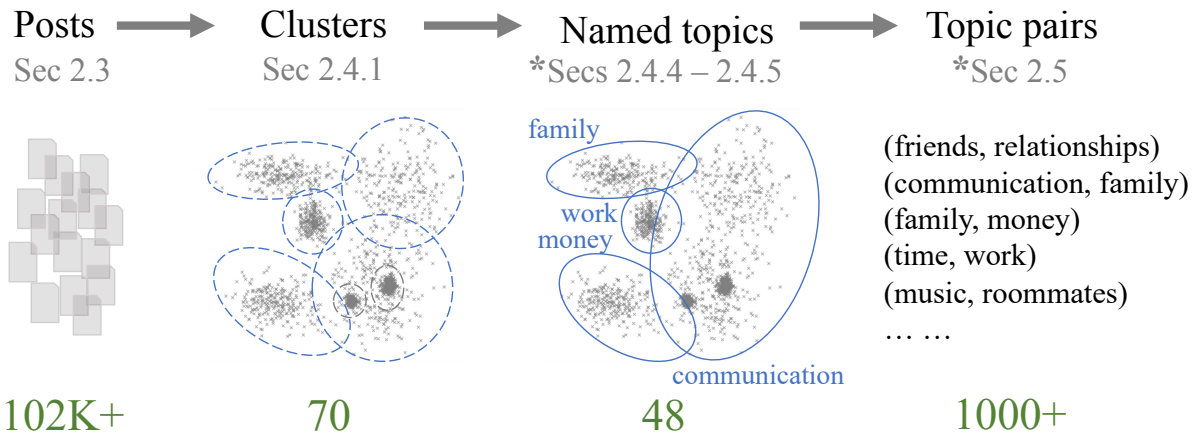


Figure 2.3: A high-level overview of the discovery process of AITA topics, with two stages of human validation indicated by *. Quantities at the bottom indicate the size (number of posts, clusters, topics and topic pairs) after each stage.

are informed by the desiderata from the work of von Luxburg et al. (2012). Firstly, our task is *exploratory* rather than confirmatory. Secondly, the use of the resulting clusters is both *qualitative* (in grounding the types of dilemmas to moral philosophy) and *quantitative* (for measuring behavioral and linguistic patterns of the resulting clusters). Moreover, we prefer clustering algorithms that allow clusters to overlap, since both the intersections and the gaps between two intuitive clusters (such as *family* and *money*) may be meaningful and interesting.

The rest of this section discusses the choices and trade-offs made in clustering posts (Section 2.4.1), the manual validation that turns clusters into named topics (Section 2.4.4), and observations of the resulting topics (Section 2.4.5).

2.4.1 Clustering posts

We perform probabilistic clustering using LDA. The input to LDA is a set of vectors containing word counts for each post. To create these vectors, we tokenize a post's body (excluding its title), lemmatize each token, remove stop words, and eliminate tokens which appear in fewer than 20 posts, all using `spacy` (Honnibal et al., 2020) and `scikit-learn` (Pedregosa et al., 2011). We keep $M = 10,463$ words across all training posts and denote these words as x_m, $m = 1, \ldots, M$. The outputs from LDA are two sets of probabilities. First, the representation for each of the K clusters $k \in \{1, \ldots, K\}$ is a multinomial word probability vector $p(x_m \mid k)$. This probability is sorted to produce the top words for each topic, which helps interpret the clusters. Second, the posterior probability of each cluster given each document d is $p(k \mid d)$, representing the salience of each topic within a document. They are sorted to produce the top cluster(s) for each document. Both probabilities will be used in topic evaluation and interpretation (Sections 2.4.4 and 2.5). While one main limitation of LDA is the use of unordered bag-of-words representations, the two probability representations lend themselves to direct human interpretation of the topics, which ensures that topics are *distinguishable* from each other. Moreover, representations generated from LDA support *overlapping* topics, both *qualitative* and *quantitative* analysis of topics, and discovery of trends and behavioral patterns.

Choosing the number of clusters is an important practical question for topic discovery, and greatly affects the *coverage* and *distinguishability* of the resulting clusters. We first examine the document perplexity on a hold-out dataset (defined in Equation (A.1) in Appendix A.3.1), which indicates that the optimal is around 40 clusters (see Figure A.2 in Appendix A.3.1). However, upon examining the sizes of the resulting clusters (by assigning documents to their top-scored cluster), we find that several clusters are too big in size (>15% of the dataset) and appear uninformative by their top keywords and top documents. We therefore increase the number of clusters to 70, which results in more balanced clusters ranging from 0.02% to 7.63% in size, all of which go through a subsequent vetting process by human experts (Section 2.4.4), resulting in 47 named topics after merging and pruning clusters. Note that it is not possible to set the number of clusters equal to 47 *a priori*, since clustering algorithms are influenced by random initialization and prone to producing a few clusters that are similar to each other (Boyd-Graber et al., 2017).

2.4.2 Alternatives in text representation and clustering.

Besides LDA on bag-of-words, we experiment with other models such as non-negative matrix factorization (Paatero and Tapper, 1994) and soft K-means (Dunn, 1973) and with other embedding methods such as TF-IDF (10,463 dimensions), Empath (Fast et al., 2016, 194 dimensions) and Sentence-RoBERTa (Reimers and Gurevych, 2019, 1,024 dimensions). While each method has its merits, we find that LDA described in this section is the most suitable. Detailed descriptions and comparisons can be found in Appendices A.3.2 to A.3.4 on page 58.

2.4.3 Cluster evaluation overview

LDA topics, just like outputs from other clustering algorithms, contain several sources of ambiguity and noise that make them difficult to use for downstream interpretation or moral reasoning tasks. First, clusters are defined by patterns of co-occurrence in data, but categories of stories need semantically recognizable *names* in order to support moral reasoning and generalization. Second, the correspondence between clusters and names is rarely one-to-one: there are often semantically similar clusters that share a name, or meaningless clusters defined by functional words for a domain, such as *edit*, *upvote*, *OP* (original poster) for Reddit. Such noise is well-known in practice, and a body of work has been devoted to topic model evaluation, stability and repair (Boyd-Graber et al., 2017, Section 3.4).

We design a rigorous two-stage evaluation for moral topics. The first stage is *naming* topics, covered in Section 2.4.4 below. This is driven by the need to having name topics in ways that are more succinct, semantically comprehensible, and free of the above noise. This process is called labeling in the topic model literature (Boyd-Graber et al., 2017). We opt to name topics manually, rather than automatically, which will not be able to prune meaningless clusters. Topic naming is done by a small number of *experts* (co-authors of this work, including both philosophers and computer scientists) because they need to be familiar with the LDA internal representation of ranked list of words, and also because of the need to deliberate (described in Section 2.4.4) when names are semantically similar but not identical.

The second stage is intended to validate the utility of the assigned names to a broad audience of online crowd workers. This is to ensure that the named topics are widely recognizable and that the names are appropriate for the posts in the corresponding clusters. See Section 2.5 on page 19 for more details.

2.4.4 From clusters to named topics

The unit for this annotation task is a cluster k, ($k \in \{1, \ldots, 70\}$). A screenshot of this web-based survey is shown in Figure A.4 in Appendix A. Each question starts with macroscopic information about the cluster—the 10 most probable keywords sorted by word probability $p(x \mid k)$. Showing 10 words is a common practice in LDA evaluation (Newman et al., 2010). This is followed by a microscopic view of the cluster—the content of three top posts, sorted by posterior probability $p(k \mid d)$, and three randomly chosen posts whose top-scoring cluster is k. By default, the list of posts is shown with the titles only, which can be expanded to show the first 100 words of the post by clicking on the title. For each task, the annotator is asked to provide a name for the cluster consisting of one or two words, or to indicate that a coherent name is not possible with *N/A*.

Six authors of this work participated in this annotation task. We collect three independent answers per question from three different annotators. Anonymized inputs are collated in a spreadsheet. Two of these annotators are then designated to resolve disagreements in naming. They review the results and make four types of decisions to name the 70 clusters: **unanimous**, **wording**, **deliberation** and **other**. There are 17 clusters with **unanimous** agreement, in which all three annotators agree on the exact wording, e.g., *shopping* and *pets*. Meanwhile, in 41 clusters, the names for the same cluster have very similar semantic meaning but exhibit **wording** variations such as synonyms. In this case, one of them is chosen based on brevity and specificity, e.g., *race* was chosen over *racism* and *babies* over *pregnancy*. A **deliberation** between the annotators is required for 9 clusters where different names are present. Here the annotators take into account whether there are two inputs that agree, the semantics of the top words, and the distinctiveness from other topics. For example, three annotators assign (*appearance*, *tatoos*, *appearance*) to a topic, and *appearance* is chosen after re-examining the keyword list and discussing the scope of the topic. Finally, there are 3 clusters with no agreement even after discussion. These are grouped into a placeholder topic **other**. Clusters with the same name are merged: 67 clusters are merged into 47 named topics in this process, with topic *family* having the most repetitions of 5. After merging, we end up with *47 named topics* (96,263 posts or 93.5%) and a placeholder topic *other* (6,735 posts or 6.5%). The topic *other* will be excluded from subsequent sections. See Appendix A.4 on page 61 for more details. Finally, as some topics are merged from several clusters, we aggregate the posteriors of clusters c with the same name into a *topic posterior* for topic k:

$$p(k \mid d) = \sum_{c\text{: name of } c=k} p(c \mid d), \tag{2.1}$$

where d is a post. Throughout the rest of this work, we refer to this definition when talking about the topic posterior. For example, the top-1 topic given document d is $\text{argmax}_k\ p(k \mid d)$.

This cluster annotation task is conducted by human experts as it requires an understand-

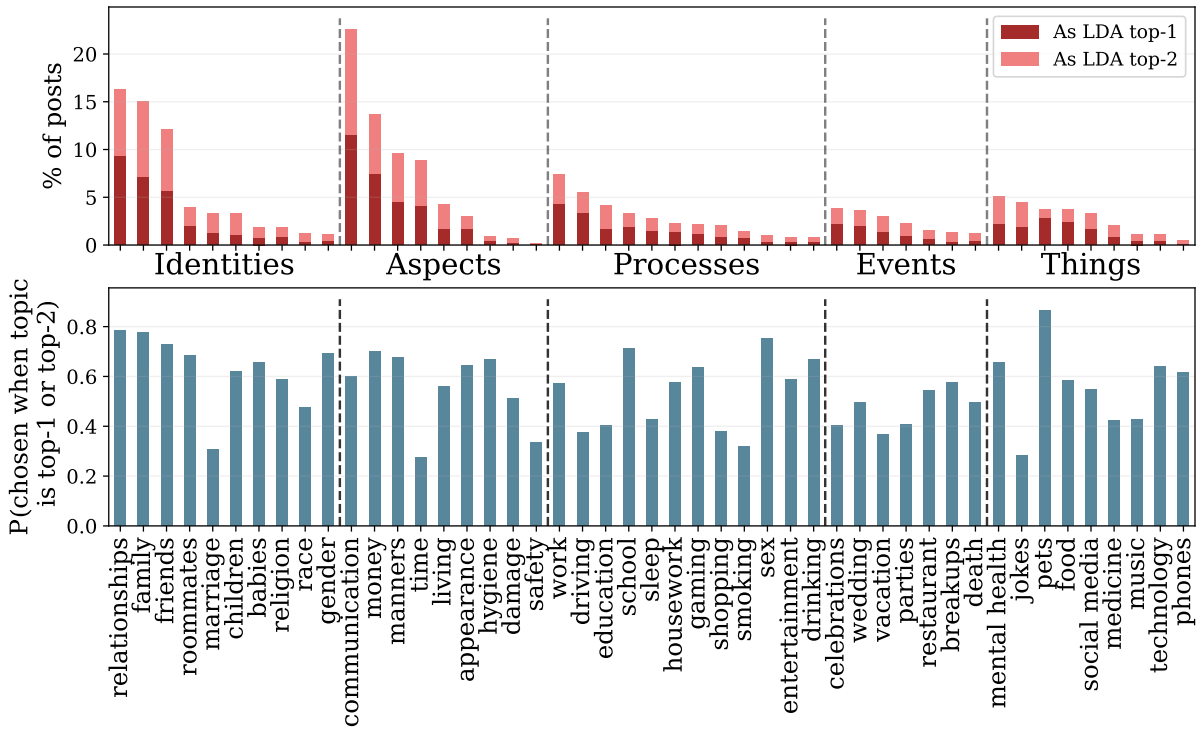


Figure 2.4: Topic level statistics on the *training* set, grouped by their meta-categories. The top bar chart shows the prevalence, as a percentage of all posts, of each topic as when it is the top-1 or top-2 highest-scoring topic (see Section 2.4). The bottom bar chart shows the topic-specific agreement rate for each topic (defined in Section 2.5.2).

ing of the AITA domain, the goal of topic mapping, and a high-level knowledge of what LDA keywords represent. The deliberation cannot easily be done in an online distributed setting. The apparently low fraction of unanimous agreement in this free-form naming task is consistent with what we observe in a topic discovery exploration (see Appendix A.2 on page 56). The named topics are then validated using crowd-sourcing by evaluating the match between topic names and post content in Section 2.5.

2.4.5 A summary of named topics

As an aid to navigate the set of topics, we further group the 47 named topics (less *other*) into five meta-categories. *Identities* (individuals and their social relationships to others) and *things* (other themes) broadly correspond to static narrative roles. Topics with a dynamic aspect are grouped into *processes* (things that happen indefinitely or regularly), *events* (specific one-off occasions that are individually important), and *aspects* (the manner in which a process or event occurs). The meta-categories, chosen by author consensus, are meant as a heuristic aid to interpretation; other carvings are possible, assignments might vary, and individual topics might cross boundaries. For example, we group *sex* as a process because many AITA posts are about the poster's sex life, which is an indefinite ongoing process, but individual instances

of sex might be better considered as events. Nevertheless, our rough grouping of topics aids interpretability. The list of topics along with their frequencies is shown in Figure 2.4 (top), grouped by meta-categories and sorted by their prevalence within. We can see that the most frequent topics are all within *identities* and *aspects*, likely due to the fact that AITA posts are often generated by social conflicts defined by relations to and manner of interactions with others.

We have five observations on the topic list. The first is that common scenarios in one's social life are covered—from family to professional relationships, from work to recreation. The second is that the topics are neither exhaustive nor fine-grained. For example, there is no topic on medical moral dilemmas common in TV dramas, likely due to their rarity in daily life. Some intuitive "topics" are absent but get coverage by their individual aspects. For example, there is no *travel* topic, but there are topics covering *vacation*, *work*, *money* and other individual aspects of travel. The third is that the prevalence of posts classified under topics such as *communication* and *manners* suggest that the way in which an action is performed is presented as morally salient. The fourth is that the relative prevalence of topics can change over time. Comparing to a validation set of 982 posts on the last three days of 2019, *family* and *celebrations* rose significantly, whereas *school* and *driving* dropped. Finally, it is surprising that the posterior probability of the top-ranked topic for each post tends to be fairly close to that of the second-ranked topic (mean difference is 0.141; see Figure A.5 in Appendix A for examples). This suggests that the top few topics for each post may be similarly relevant, rather than only the top topic being significantly relevant to a post.

2.5 Crowd-sourced topic survey

We design and conduct a set of crowd-sourced surveys to answer two key questions: how well do human annotators agree with the named topics, and how do users at large perceive topics of an AITA post? A complete description of the survey presented can be found in Appendix A.5 on page 67.

2.5.1 Survey setup

Each crowd-sourced survey consists of a number of questions, each of which is centered on an AITA post and starts with a fixed prompt: *"What topics below best describe the theme of the following post? Do not let your ethical judgement of the author affect your choices here."* We then present the post title and body text, and five topic choices. The first four choices are a randomized list of the top 4 topics according to the topic posterior, followed by a *None of the above* option. A participant can choose one or more non-conflicting options before moving on to the next post. An example question is shown in Figure 2.5. Free-form text boxes are also provided to collect participants' reflections at the end of each question, as well as at the end of the survey.

We use the Prolific crowd-sourcing platform to recruit participants.[7] Each individual can only enter once, and we collect answers from three different participants for each question.

[7] https://www.prolific.co

Question: What topic below best describes the theme of the following post? Do not let your ethical judgement of the author affect your choice here.

Title: WIBTA if I sprinkle in just a few lies into my grad school applications?

Context: I want to apply for a PhD program in Aerospace Engineering at the top names. But [I have] only 1 summer of research. [...] I was thinking of adding some more "fake" research experience by going through the aerospace engineering department at my school and listing some of the profs researches and saying I helped them during a certain timeframe. [...] I don't think they will contact every professor, since grad schools get lots of applications. [...] WIBTA if I fudge some research experience into my application?

- ❑ School
- ❑ Mental health
- ❑ Money
- ❑ Education
- ❑ None of the above

Figure 2.5: One example question in the topic validation survey. (See Section 2.5.1 and Appendix A.5 for more details.) Each question contains a post (title and body) and has four topic options. The participant can choose more than one option, or *None of the above* if no topic name matches the post. For this post, the 4 highest-scoring topics according to the LDA, in descending order of posterior probability, are *education*, *school*, *money*, and *mental health*.

To control the quality of results, we only allow fluent English speakers to participate. Before entering the actual survey, each participant is given one training question, containing one post clearly belonging to two of the given topics. Choosing the correct answers for this training question is a prerequisite for completing the rest of the survey.

Based on a pilot test among the authors, we set the length of each survey to 20 questions, with a time estimate between 12–20 minutes. A total of 285 participants (130 males, 151 females, 4 unspecified) completed the survey, each of whom was paid £2.5 for their work. Their average age is 28.2 (SD = 9.2), with 39.1% living in either the US or UK. This survey design is approved by the ethics committee of the authors' institution. More information about this survey and participation statistics can be found in Appendix A.5.4 on page 69.

We collect survey results in three settings. On the *training* split of AITA (see Section 2.3 above), we randomly select 20 posts for each topic, and call this setting **train**. The topic choices are the top 4 choices according to the LDA posteriors. We increase the size of the survey with

Table 2.1: Post-level agreement rates between survey participants and LDA topics.

Answer type	**Train**	**Test**	**Test+rand**
Top-1 only	65.1	59.2	68.0
Top-2 only	48.9	50.4	58.4
Top-3 only	36.3	39.3	8.2
Top-4 only	29.9	26.1	8.4
Top-1 or 2	83.2	81.9	88.4
Top-1 or 2 or 3	90.8	91.0	–
"None of the above"	4.8	5.4	9.5

5, 10 and 20 posts per topic gradually, and find that the statistics stabilize after 10 posts per topic. On the *test* split of AITA, which is not seen by either the LDA estimation or in topic naming, we randomly select 10 posts for each topic populated with its top 4 topics, and call this setting **test**. This gives us 450 posts in total. Note that for 5 topics with fewer than 10 posts, we simply include all the posts. Lastly, we use the same set of posts from the *test* set, but include the top-2 topics according to LDA, plus two other randomly selected distractor topics for each post. We call this setting **test+rand**, which is designed to observe whether or not the top 2 topics are significantly more descriptive than other randomly selected topics. These three settings are shown as column headings in Table 2.1.

2.5.2 Agreement rates for posts and topics

We report two metrics on the survey results: the post-level agreement rate and the topic-specific agreement rate.

Post-level agreement rate is the percentage of answers for which the participant agrees with at least one of the designated topics of a certain type, aggregated over different participants. Here the types of choices are *Top-k only* (with $k = 1, 2, 3, 4$), *Top 1 or 2*, *Top 1, 2 or 3*, or *None of the above*, presented as rows in Table 2.1. Agreements rates between the *train* and *test* settings are similar with a small decrease for answers in *test*, indicating that the topics generalize reasonably well to new posts. The decreasing trend from *top 1* to *top 4 only* is expected due to their decreasing LDA topic posteriors. In the *test+rand* setting, the presence of irrelevant (random) topics increases the probability that either the top-1 or top-2 topic being selected by 8%, and *None of the above* by 4%. This observation is consistent with well-known behavior patterns in choice-making (Simonson and Tversky, 1992), namely the tradeoff contrast that enhances options in the presence of unfavorable alternatives.

A topic-pair representation. The average number of topics chosen by participants is 1.70 (*train*: 1.80, *test*: 1.75, *test+rand*: 1.43). The frequencies for answer lengths can be found in Figure A.8 in Appendix A. Given that the survey leaves the number of topics chosen unconstrained, this observation reveals that participants often perceive more than one topic being relevant to the post. Moreover, the agreement rate for *top 1 or 2 topics* is 81.9% (+22.7% from *top 1 only* and 9.1% less than *top 1, 2 or 3*) for *test*, and *88.4%* on *test+rand*. This observation prompts us to define (unordered) ***topic pairs***, i.e., top-1 and top-2 topics for each post, as the

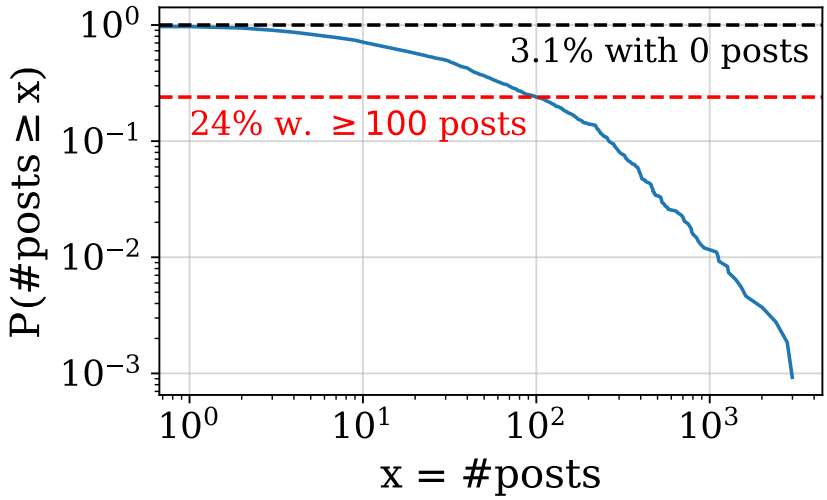


Figure 2.6: Distribution of the sizes of topic pairs. The x-axis is the size of a topic pair (in logarithmic scale) and the y-axis is the complementary cumulative distribution function (CCDF, also in logarithmic scale). For example, the CCDF at 100 posts represents the proportion of topic pairs that contain at least 100 posts. In our dataset, this is about 24% of the pairs (red dashed horizontal line). Thirty-three out of 1,081 pairs (3.1%) have no post at all (black dashed horizontal line).

automatically extracted relevant topics. The topic pairs are unordered, because the posterior probabilities of top-1 and top-2 topics are close in value (Section 2.4.5). Additionally, as surfaced in the deliberation process of topic naming task (Section 2.4.4), annotators could not distinguish which of the top two topics is more prevalent. We posit that the topic-pair representation makes the classification of moral dilemmas significantly more nuanced and richer. Further observations on topic pairs are presented in Sections 2.5.3 and 2.6.

Topic-specific agreement rate is defined as the percentage of times that a given topic is selected when presented as either *top*-1 *or top*-2 for a post, aggregated over different participants. Results for *train* are shown in Figure 2.4, and those for *test* and *test+rand* are in Figure A.6, which show the same patterns for topic prevalence and agreement rate. We observe that frequent topics such as *communication* and *friends* have relatively higher agreement rates ($\geq$ 60%). Topics belonging to *identities* generally have higher agreement rates than other meta-categories. A few infrequent topics have high agreement rates, such as *pets*, which may be explained by being defined by animal-related words. Topics such as *jokes* and *time* are among the least agreed upon; one explanation is that they may appear as the secondary topic or issue, together with another main issue. We note that the (weighted) average of the topic-specific agreement rates is lower than post-level agreement rate on the same setting, due to the latter requiring *either* the top-1 *or* top-2 topic being selected.

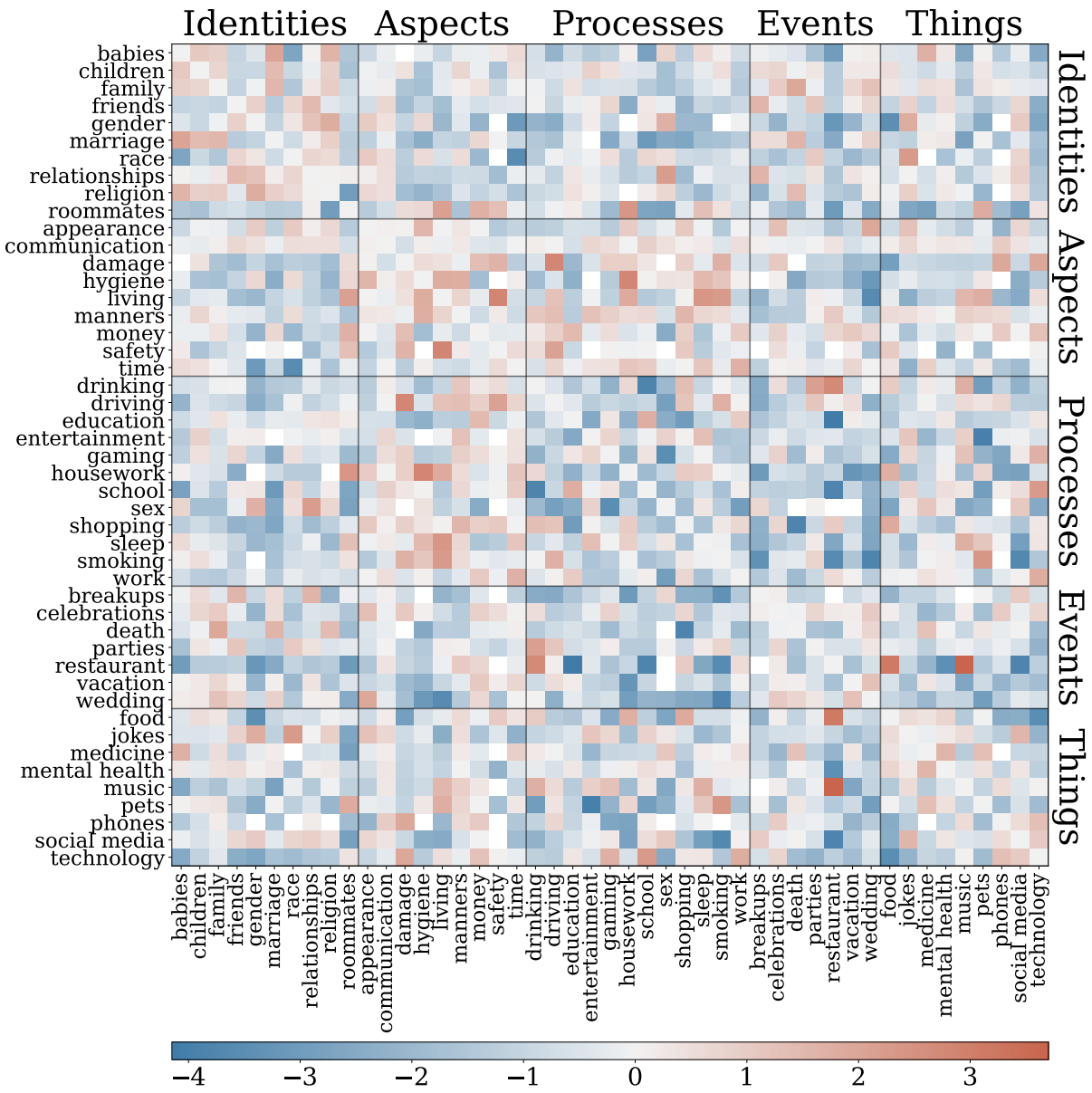


Figure 2.7: Point-wise mutual information, defined in Equation (2.2) on page 24, between every pair of topics found by LDA. A positive value (red) indicates that the two topics co-occur more often than expected by chance, and a negative value (blue) indicates that they co-occur less often than expected by chance. The topics are grouped by their five meta-categories.

2.5.3 A profile of topic pairs

From 47 named topics, there are $\binom{47}{2} = 1,081$ unordered topic pairs. Among these, 33 pairs (3.1%) have no posts, 396 pairs (36.6%) contain at least 50 posts, and 259 pairs (24.0%) contain at least 100 posts. The 10 largest topic pairs are shown in Figure 2.8 as row labels. Figure 2.1 on page 10 shows an overview of topic pairs, and Figure 2.6 shows the distribution of the

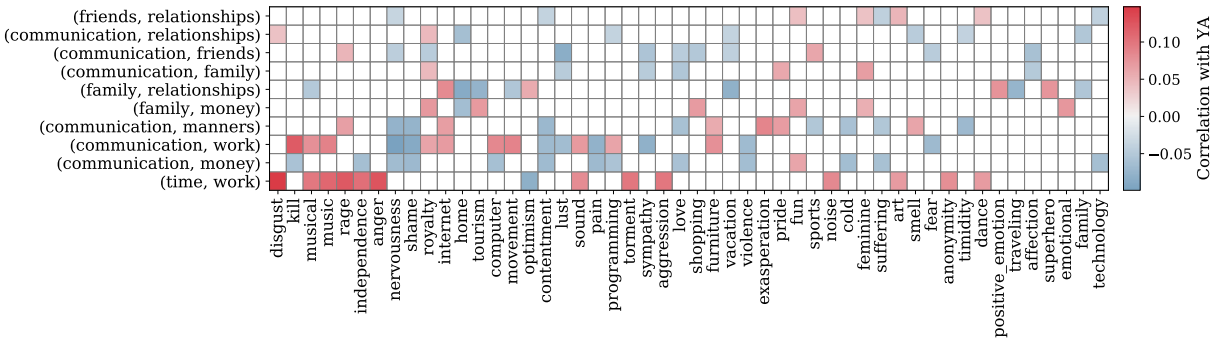


Figure 2.8: Pearson correlation between Empath emotional categories and the YA verdict in several topic pairs on AITA. The rows are the top 10 topic pairs in terms of size (number of posts). The columns are the top 50 Empath categories sorted by variance. A red cell indicates a positive correlation, i.e., the corresponding Empath category is associated with YA judgments, while a blue cell indicates the opposite. A white cell denotes a lack of statistical significance ($p > 0.05$; 388 in this plot).

sizes for all topic pairs.

How often do we observe topics k and k' together? We use the point-wise mutual information (PMI) to quantify how much two topics co-occur more than prior-calibrated chance (PMI > 0), or less than chance (PMI < 0):

$$\mathrm{PMI}(k, k') = \log_2 \frac{p(k, k')}{p(k)p(k')}. \tag{2.2}$$

The PMI matrix is shown in Figure 2.7. Among the meta-categories, topics in *identities* and *aspects* are likely to co-occur with another topic within the same meta-category, whereas those in *processes* do not. Topics in *aspects* tend to co-occur with those in *processes*, as one might expect.

Many topic pairs can be explained by semantic relatedness (or exclusion): *restaurant* tends to occur with *food* and *drinking* but not with *education*. On the other hand, some pairs appear to indicate conjunctions that are a frequent source of conflict and thus generate moral dilemmas. Some of these connections are obvious—witness the high PMI for *race* with *jokes*, or *children* with *religion*. Both express domains that generate moral conflict on their own; one might reasonably expect even more conflict at their intersection. On the other hand, some conjunctions suggest more subtle patterns of conflict, like *restaurant* with *music*, or *race* with *food*. One or both of the topics in these pairs do not seem particularly morally laden on their own. Some more complex interaction is likely at work. While the present work does not focus on particular mechanisms, we think this might be a rich topic for future work. We suspect that insofar as these pairs give rise to moral dilemmas, they might do so against a complex social background of expectations and norms. Additional profiles on the commenting and voting patterns across topic pairs can be found in Figure A.8 in Appendix A.

2.6 Linguistic patterns in topic (pairs)

We examine the variations in word use across topics and topic pairs.

2.6.1 Topic pair statistics via emotional categories

We first profile word use by Empath (Fast et al., 2016), a crowd-sourced collection of topical and subjective word lists, containing 194 categories (see the column labels of Figure 2.8 for examples) and 15 to 169 words in each category, totaling 7,643 words. We generate a 194-dimensional vector for each post, with elements corresponding to the fraction of words in each Empath category. For each topic pair, we compute the Pearson correlation between each Empath dimension and the binary indicators for `YA` judgments. Results are presented in Figure 2.8. Some categories, such as *independence*, negatively correlate with `YA` in (*communication*, *money*) but positively correlate with `YA` in (*time*, *work*). Categories such as *love*, *shame*, *nervousness* negatively correlate with `YA` in multiple topic pairs, whereas *fun* and *feminine* positively correlate with `YA` in multiple topic pairs. These correlation patterns indicate that the moral valence of similar words may differ across different topic pairs. It also emphasizes that topic pairs are a key covariate for further analyses.

2.6.2 Scoring moral foundation axes

To directly examine the topics' moral content, we appeal to an influential framework called moral foundations theory (MFT) which projects the space of moral problems into five moral "foundations": *care*, *fairness*, *loyalty*, *authority* and *sanctity* (Haidt, 2013). First, we rely on a word-count lexicon called Moral Foundations Dictionary 2.0 (MFD 2.0) (Frimer, 2019), which contains 2,041 unique words in total. For each post, we compute a 5-dimensional binary vector, with each dimension being 1 if the post contains at least one word in the corresponding foundation. We do the same for the top-scoring comment of each post (called *verdict*). These vectors are aggregated over the posts/verdicts of the same topic or topic pair, and normalized by the total number of posts/verdicts. This yields a five-dimensional vector with values between 0 and 1, representing the fraction of posts/verdicts with the corresponding foundation.

2.6.2.1 Moral foundation prevalence for topics and topic pairs

Figure 2.9 presents the proportions of posts (row a) and verdicts (row c) containing each foundation for five topics: *family*, *marriage*, *death*, *religion* and *money*. We display these statistics for all topics in Figures A.12 and A.13 in Appendix A. Also in Figure 2.9 (rows b and d), the same proportions are presented for topic pairs involving *family*. We observe some patterns consistent with the MFT. The foundation *care* appears significantly in most posts of any topic: for example, nearly every post within the topic *family* has the presence of *care* (radar plot a-1). In posts about *religion* (plot a-4), the authors tend to attach the foundations *sanctity* and *loyalty* in their narratives. These congruences provide a useful proof of concept. These observations are consistent when we look at the verdicts (plots c-1 and c-4, respectively). When topics are subdivided based on valence (`YA` and `NA`), the red and blue regions on row a mostly overlap, indicating there is little difference on what moral foundations positively and negatively judged

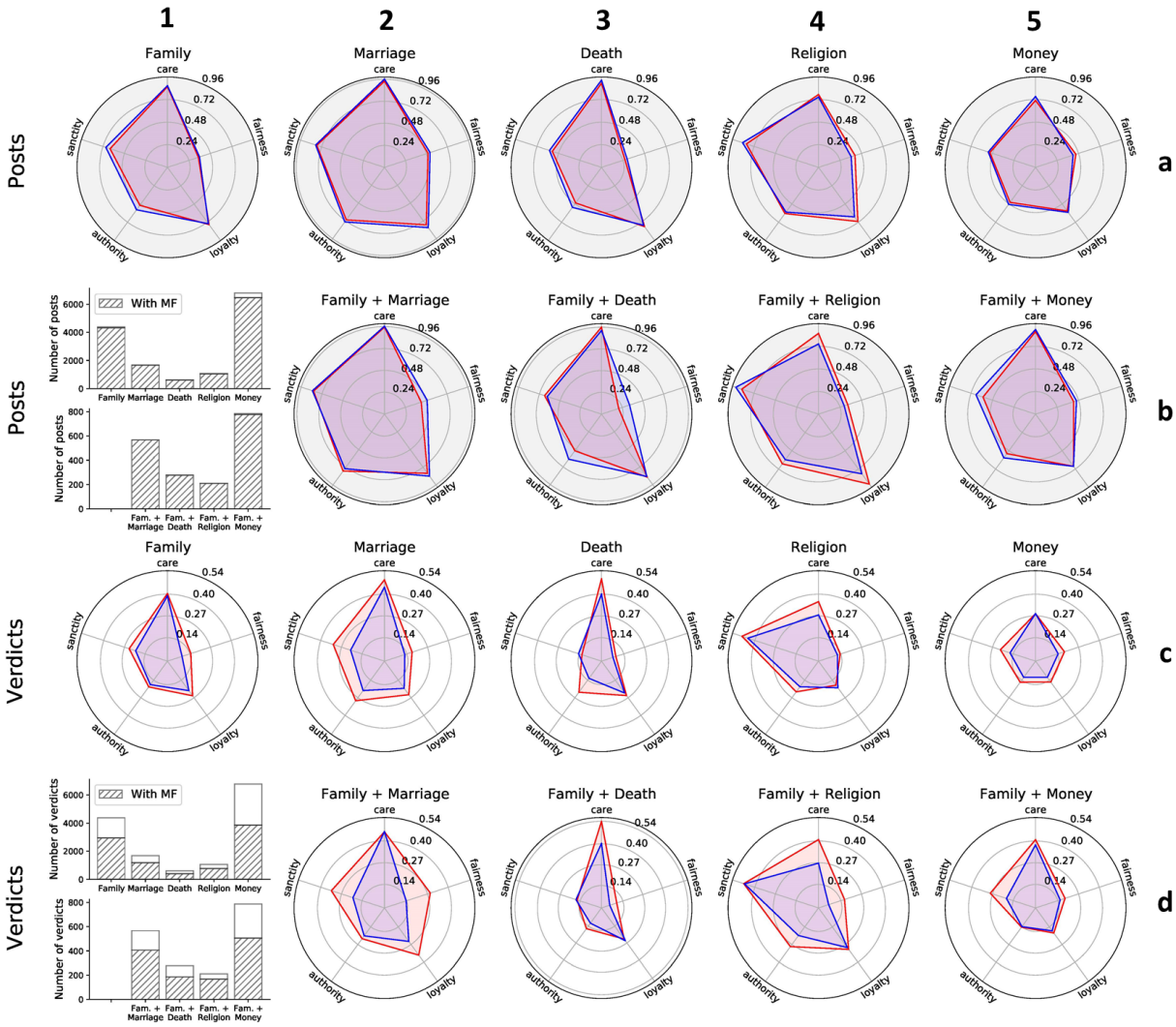


Figure 2.9: Strengths of moral foundations in selected topics and topic pairs. In each radar plot, each pentagon's vertex represents the proportion of posts (or verdicts) in a topic (pair) that have the presence of at least one moral foundation. Red (resp. blue) pentagons represent YA-judged (resp. NA-judged) posts (or verdicts). The bar plots at positions b-1 and d-1 represent the number of posts (or verdicts) in each topic (pair), and how many of them contain a moral foundation.

posts appeal to. When we look at verdicts (row c), YA verdicts typically adhere to every moral foundation more than NA verdicts. This could be explained by the fact that negatively-judged comments are longer than positively-judged comments on average, increasing the chance they include a moral word.

We also find evidence that secondary topics provide an interesting additional interpretive layer. Figure 2.9 (rows b and d) shows that the combination of topics often produces unexpected effects on the underlying moral foundations to which posts and verdicts appeal. For example, the combination of *family* and *money* produces YA judgments that appeal to *sanctity* more frequently than either topic does alone (plot d-5, compared with c-1 and c-5); a similar pattern is seen in *family* and *marriage* with *fairness* (d-2, compared with c-1 and c-2). Conversely, some MFD loadings are driven more by one topic or another. The mechanism behind these interactions remains a topic for future research. We suggest that this is good evidence that our topics provide a cross-cutting categorization (Dupré, 1993) of the moral domain, one that might reveal a more fine-grained structure that drives individual moral judgments.

There are also interesting dissociations between posts and their verdicts. Posts use a wide range of identifiable moral language across different MFD domains. This confirms that posters to AITA treat what they are saying as morally laden. The verdicts, on the other hand, tend to focus in on a smaller subset of moral considerations. For example, posters concerned with *family* and *religion* very often focus on both *sanctity* and *loyalty* (plot b-4), but verdicts tend to downplay that in favor of a strong focus on *sanctity* (plot d-4). Some reasons also seem to distinguish verdicts: YA judgments for *family* and *marriage* focus more on *loyalty* and *sanctity* than do NA judgments (plot d-2). These effects come apart from the original posts, where the radar plots largely overlap between NA and YA (plot b-2). These dissociations suggest that verdicts do not necessarily follow the original framing of the poster, and that the subsequent discussion plays an important role in focusing attention on details. They also show that dilemmas can have a non-additive structure (Kagan, 1988), in which the presence of one topic can affect the importance of reasons raised by a different one.

We note that all five studied moral foundations are often present, to varying degrees, within what is broadly the same online population. Even strong predictable associations (such as *religion* with *sanctity*) coexist alongside appeals to other types of reasoning. It is no surprise that real-world cases are often quite messy. This is part of the attraction of AITA. Part of that complexity comes from the interaction of different domains, which are revealed by our topics. Hence a bottom-up approach provides a valuable complement to experimental studies, which for good reason often focus on clear cases.

The latter part of this thesis (cf. Chapter 3) discusses in more detail other tools to measure moral foundations in text. In particular, we find that dictionary-based word count methods, such as what is used in this section, have a number of serious limitations, most notably the subjectivity of word choice in these lexicons and the sensitivity of foundation scores to an input's length. In the next chapter we propose a better alternative to these approaches and redo the same analysis using this new tool, which, in numerous cases, leads to qualitatively different findings. The reader should, hence, interpret the results in this section with caution.

2.6.2.2 Coverage of moral foundations dictionary

Of the 102,998 posts in the training data, there are only 5,425 (5.3%) posts without the presence of any foundation in its description. However, we find that the MFD 2.0 has relatively low coverage on the AITA verdicts. There are 44,282 (43%) posts for which MFD finds no presence of any foundation in their verdicts. Of these, verdicts in topics *phones*, *music*, *shopping*, *roommates*, *driving* and *celebrations* have the highest missing rates of above 50%, while verdicts in *religion* have the lowest rate of 28%. This is evidence that MFD 2.0 may miss important moral considerations, particularly on the comparatively shorter verdict posts. Below is an example verdict where the MFD 2.0 does not detect any foundation:

> **Post title**: "AITA For Firing An Employee After His Parents Died?"
>
> **Verdict**: "`YTA` for firing him without first going through the steps of describing his issues to him and giving him a chance to improve. He's been back for only 2-3 weeks. It's not about 'having heart', it's about making a dumb business decision for both you and him. So much smarter to work with this guy to get him back on track after a temporary setback than to push the eject button and have to find and start over with a new person. Dumb."

This verdict appeals to considerations of both *authority* and *fairness*. *Authority* is the power to issue commands and enact rules that are generally followed by the appropriate subject group; employer-employee relationships fall under this heading. *Fairness* involves adhering to a set of procedural safeguards, and the employer in this example plausibly violated these procedures.

We note that several versions of MFD have been introduced by different authors. Other types of lexicon are also available, such as the morality-as-cooperation vocabulary. Potentially combining different lexicons and validating our findings across different dictionaries are left as future work.

2.7 Conclusion

In this work, we analyze more than 100,000 interpersonal moral dilemmas on a Reddit forum called AITA. Using a multi-stage data-driven approach involving text clustering and human expert annotation, we group these posts into 47 high-quality topics with high coverage of 94% of the dataset. Through crowd-sourced validation, we find high agreement between human annotators and our topic model when describing the themes of an AITA post. Furthermore, we observe that topic pairs are better than individual topics at depicting a post's content, and therefore better serve as a thematic unit over AITA posts. We make several observations that suggest topic (pairs) is a key factor for thinking about daily moral situations. For instance, certain topics attract or repel other topics even when neither topic is particularly morally laden; the moral valence of similar words can vary across different topic pairs; and interaction effects in which final verdicts do not line up with the moral concerns in the original stories in any simple way.

2.7.1 Ethical considerations

We take steps to ensure that the study on moral dilemmas minimizes risk of harm. In both annotation tasks, we hide Reddit usernames and embedded URLs in posts to avoid identifying the original posters. We do not edit the names mentioned in posts since they are mostly initials or pseudonyms created by the poster. We present aggregated data that cannot be traced back to particular survey participants. Our survey design is approved by our institution's ethics committee.

2.7.2 Limitations

As with all observational datasets, our collection method cannot retain posts and comments which had been deleted before the retrieval time, possibly leading to missing or incomplete data. Furthermore, it is impossible to precisely trace the comment containing the winning verdict in a thread, because after 18 hours (the amount of time after which the Reddit bot determines the flair), comments' scores can change drastically. This is a drawback compared to other Reddit datasets such as `r/ChangeMyView` in Tan et al. (2016), where the original posters explicitly give the winning comment a special symbol. Despite AITA participants being self-selected, and cannot be considered a representative sample either of Reddit users or the population at large, this work assumes that the content in AITA reflect daily life in interesting ways. The resulting topics provide evidence of the diversity and nuance of the set of daily moral discussions, and does not provide measures of representativeness for each topic. Our data is limited to posts that follow the posting guidelines set up by AITA moderators. These guidelines prohibit posts about reproductive autonomy, revenge, violence, and conflicts with large social demographics. Conflicts within these prohibited topics could fall within the bounds of morality but are excluded from our dataset. Lastly, Reddit does not release demographic information about its registered members. This means that our analysis cannot determine the extent to which the everyday moral dilemmas posed on AITA (as well as the moral judgments expressed in the comment section) are the product of the specific social or institutional roles embodied by its registered members.

2.7.3 Future directions

The present study only looks at posts and verdicts on AITA. A natural extension would be to examine the content and structure of comments on each post. Our data also shows that posts often reflect a mixture of topics; it would be interesting to see whether the subsequent discussion preserves this mix or whether the search for reflective equilibrium (Rawls, 1971) implies focusing on specific topics. It is also known that moral judgments can depend on the way situations are framed (Sinnott-Armstrong, 2008); studying discussions might shed new light on these framing effects.

Part of the motivation for studying AITA was a philosophical interest in morally charged situations (Driver, 1992). We are interested in the degree to which debates on AITA might challenge the traditional distinction between moral norms and merely conventional norms like rules of etiquette (Foot, 1972; Southwood, 2011). There have been recent challenges to this sharp division (Martin and Stent, 1990). Our results are consistent with this challenge, with

an important contribution from topics like *manners* and *communication* suggesting that the way things are done can be as important as what is done. Further work may shed light on what distinction, if any, can be drawn between the two domains. Second, a core tenet of early Confucian philosophy is that the everyday challenges and exchanges that people experience are of profound importance to morality (Olberding, 2016). We note that the everyday challenges and exchanges that occupied early Confucian philosophers are similar to our real-world moral dilemmas. Future research could help identify links between the two. Finally, we note that a large number of topics concern particular kinds of relationships, like *children*, *family*, and *friends*. This may be of particular interest to care ethics, as well as some forms of virtue ethics and communitarianism, which emphasize the moral importance of meaningful relationships (Collins, 2015).

Chapter 3

Measuring Moral Dimensions on Social Media with Mformer

3.1 Introduction

Recent years have witnessed a growing interest in the study of moral content on social media. Many online discussions have a tendency to reflect aspects of morality, and researchers thus far have aimed to study how and to what extent moral dimensions vary throughout this vast domain. One particularly influential framework in the analysis of such content is moral foundations theory (MFT), which maps morality to five fundamental psychological dimensions called "moral foundations": *authority*, *care*, *fairness*, *loyalty* and *sanctity* (Haidt and Joseph, 2004; Haidt, 2013). Prior research suggests that the variation in moral sentiment within and across cultures can be attributed to differences in the way these cultures value each moral foundation. Notable works, including those on vaccine hesitancy (Weinzierl and Harabagiu, 2022), social norms (Forbes et al., 2020) and news story framing (Mokhberian et al., 2020), have extensively taken this dimension-mapping approach to uncover large-scale patterns of moral belief and judgment.

As human labeling does not scale to the size of modern corpora, MFT-based studies of online moral content must rely on tools to automatically detect moral foundations in text. However, existing methods, especially word count programs based on human-crafted lexicons, are surprisingly lacking in their consistency and ability to generalize to different domains (see Figure 3.1 for an illustration). Variations across these methods can lead to significant changes in downstream findings based on such measurements. In this work, we propose Mformer, a Moral foundations classifier based on transformers fine-tuned on datasets from diverse domains, which is released publicly.[1] Compared to a set of current approaches, we find that simply using diverse datasets for fine-tuning works surprisingly well—we observe that Mformer consistently achieves better accuracy on several datasets, with a relative AUC improvement of 4–17%. Through two case studies involving moral stories on Reddit and controversies on Twitter, we demonstrate the effectiveness of Mformer in explaining non-trivial variations in people's moral stances and judgments across many social issues. The main contributions of this work are as follows:

- We introduce Mformer, a moral foundations classifier based on a fine-tuned language

[1] https://github.com/joshnguyen99/moral_axes

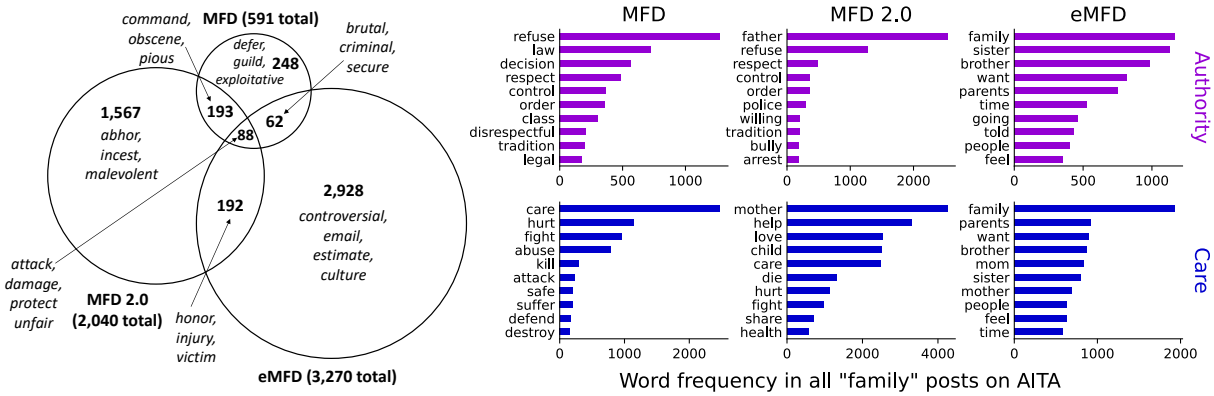


Figure 3.1: Three existing lexicons—MFD, MFD 2.0, and eMFD—used for word count in detecting moral foundations. *Left*: Venn diagram depicting the sizes of these lexicons with some example words. *Right*: the 10 most popular words for two moral foundations (*authority* and *care*) in each lexicon that are found in 6,800 `r/AmItheAsshole` posts of the topic *family*. See Section 3.3 on page 34 for a detailed discussion.

model on text data from diverse domains (Section 3.4).

- Through an in-depth analysis of word count programs, we show why and how they tend to fall short in labeling moral foundations in text (Section 3.3). On the other hand, Mformer consistently performs the best across several in- and out-of-domain datasets (Sections 3.4 and 3.5).
- We demonstrate the utility of moral foundations in text analysis through two case studies involving (i) moral stories and judgments and (ii) stance toward several controversial topics (Section 3.6). We highlight the difference between downstream conclusions resulting from word count and those from Mformer. This suggests that many prior findings relying on MFT measurements may warrant further scrutiny.

3.2 Related Work

Our work builds upon the literature on defining key moral dimensions, detecting them in text, and measuring a variety of patterns pertaining to morality within large corpora.

3.2.1 Moral foundations theory

Moral foundations theory, developed in psychology, proposes a taxonomy over what are called "moral intuitions" (Haidt and Joseph, 2004; Haidt, 2013). In an attempt to explain the similarities and differences in moral judgment across cultures, Haidt and colleagues postulated key categories of intuitions behind people's moral judgments. These so-called "moral foundations" fall into five spheres: *authority*, relating to traits such as deference to higher authorities to maintain group cohesion; *care*, upholding the virtues of nurturing and protection; *fairness*, involving equal treatment and reward; *loyalty*, relating to prioritizing one's group and alliances; and *sanctity*, including intuitions about maintaining the sacredness of the body and

avoiding moral contamination. Anthropological evidence suggests that these foundations are universal, although which traits constitute a virtue and which constitute a vice vary across cultures (Graham et al., 2013).

Most studies using MFT have focused on characterizing political ideology, especially in the U.S. context. For example, pertaining to the foundations *care* and *fairness*, liberals often support the commitment to justice and actions that uphold equality and minimize suffering. Conservatives, on the other hand, tend to value all five foundations somewhat equally, additionally upholding virtues such as loyalty to one's country and respecting authority and order (Graham et al., 2009; McAdams et al., 2008; Haidt and Graham, 2007; Van Leeuwen and Park, 2009). The theory has also been used to characterize the differences in moral sentiment surrounding socially significant topics such as stem cell research (Clifford and Jerit, 2013), vaccine hesitancy (Amin et al., 2017), euthanasia, abortion, animal cloning, and same-sex marriage (Koleva et al., 2012).

3.2.2 Automatically detecting moral foundations

A requirement for large-scale studies of MFT is the automatic detection of moral foundations. While the gold standard for moral foundation labeling is human annotation, this requires extensive training and evaluation, which does not feasibly scale to large text collections. We categorize the automatic scoring of moral foundations into word count, embedding similarity, and supervised classification methods.

Word count methods rely on a human-crafted lexicon, which is a mapping from words to foundations. When scoring a document, each token contained in the lexicon counts as one occurrence of the foundation that it maps to. Three lexicons, often called moral foundations dictionaries (MFDs), have been extensively used in the literature: MFD (Graham and Haidt, 2012), MFD 2.0 (Frimer, 2019), and extended MFD (Hopp et al., 2021); we further review these in Section 3.3. Word count methods bear resemblance to LIWC, a popular program that scores psychological and emotional features in text (Tausczik and Pennebaker, 2010).

Embedding similarity methods define the relevance between a document and a moral foundation as the cosine similarity between their word embeddings, such as word2vec (Mikolov et al., 2013). In particular, each foundation is represented by the average embedding of its "seed words," and similarly each document is the average of its tokens. Mokhberian et al. (2020), for example, used this method to score moral foundations as features of news articles. Similar to Rocchio classification (Manning et al., 2008), embedding similarity essentially relies on the centroid of a set of seed words and is well-known to have limited expressivity. Furthermore, choosing a good set of seed words remains a major practical challenge.

Supervised classification methods treat the detection of moral foundations as a text classification task. For example, Hoover et al. (2020) annotated a dataset of tweets and used it to train support vector machines for classification. Similarly, using annotated social media data, Trager et al. (2022) fine-tuned BERT (Devlin et al., 2019), a language model (LLM), for the same task and reported state-of-the-art performance. However, these models may not generalize well to new datasets (Liscio et al., 2022). Recently, Guo et al. (2023) trained moral foundation classifiers with a new loss function to incorporate domain variations. In comparison, our approach uses a standard LLM architecture trained on multiple data domains. This

approach is arguably more straightforward as it requires no additional hyperparameters, does not demand explicit one-hot encoding of the domains, and thus offers improved adaptability when a new domain enters into the training dataset.

Finally, some prior work made a distinction in the polarity of moral foundations, i.e., whether an example portrays a virtue or a vice of a foundation. As we explain later in Section 3.4.1, there are several conceptual and practical concerns for this approach. Thus, we decide to only score moral foundations, irrespective of polarity, in this work.

3.2.3 MFT-based analyses of text

A growing body of work has adopted MFT to the analysis of moral rhetoric from online media. For example, Mokhberian et al. (2020) studied the relationship between story framing and political leaning of several news sources. The study found that on topics such as immigration and elections, conservative-leaning sources tend to focus on the virtues of *sanctity* such as austerity and sacredness, while liberal-leaning sources emphasize the condemnation of the vices of *sanctity* like dirtiness and unholiness. Hopp et al. (2020) analyzed movie scripts and identified relevant foundations in movie scenes exhibiting moral conflicts. Using network-theoretic methods alongside moral foundations, the authors were able to construct communities of characters with specific shared moral characteristics. In argument mining, Kobbe et al. (2020) proposed to use moral foundations in the automatic assessment of arguments. The study found a significant correlation between moral sentiment (the presence of moral foundations) and audience approval of an argument. In another line of work, Forbes et al. (2020) and Ziems et al. (2022) used MFT to categorize collections of moral "rules of thumb" collected from large crowds and assessed LLMs' moral sentiment predictions.

Empirical results based on moral foundations scoring currently face two important limitations. First, there is no unified and highly accurate method on which researchers rely to label their text corpora. Second, and as a result, conclusions drawn from prior studies likely suffer from low-quality scoring and may vary based on what method is chosen. A solution sufficiently addressing these two challenges will allow MFT-based analyses to scale and facilitate their reproducibility, comparison, and generalization.

3.3 Limitations of Moral Foundations Dictionaries

Word count methods based on lexicons remain the most popular for automatically detecting moral foundations, often serving as the default choice among researchers. Here we detail how they work and several of their limitations when applied to text corpora.

3.3.1 Scoring via word count

Each lexicon, called a moral foundations dictionary (MFD), contains a list of words and the foundations they represent. For example, the word "deceive" can be mapped to the foundation *loyalty*. A document to be scored is first tokenized; then, for each token that is also in the MFD, the count for the foundation that it is mapped to is incremented. For instance, if the token is "deceive," the count for *loyalty* is increased. These methods are easy to implement,

involve no model training, and are interpretable since they directly show what words signify a foundation.

3.3.2 Available lexicons

Three versions of the MFD have been used extensively in the literature. The first MFD, released as a dictionary to be used in LIWC-like programs, contains nearly 600 words (Graham and Haidt, 2012). Later, a new version with over 2K words called MFD 2.0 (Frimer, 2019) was released in which the creators extended their expert-crafted word lists by querying a word embedding (Mikolov et al., 2013) for similar terms. Another recent variant, called the extended MFD (eMFD, 3.2K words) (Hopp et al., 2021), contains one-to-many mappings associated with moral foundation weights, between 0 and 1, for each word. We list some example words of these MFDs in Table B.1 in Appendix B (page 78) and show some overlap between them in Figure 3.1.

3.3.3 Limitations

We want to understand the different MFD versions and what they each capture, but have since identified other limitations that will affect downstream measurements and interpretations.

First, these lexicons have a fixed and limited vocabulary. See Figure 3.1 on page 32 (left) above for some example words and a visualization. Most of the words in these lexicons are supposedly morally relevant, but their overlap is surprisingly small. For instance, the 281 words common to both MFD and MFD 2.0 only account for 47.5% of MFD and 13.8% of MFD 2.0. In eMFD, 89.5% of the words do not appear in either MFD or MFD 2.0 at all. Such lack of consensus in expert- and machine-created lexicons is concerning, and one can rightfully question whether the resulting scores are reliable.

Second, when using the MFDs, it is unclear how word variations should be handled. For example, the MFD 2.0 explicitly contains the verb "desecrate" and its inflectional forms "desecrates," "desecrated" and "desecrating." However, this is not the case for the verb "deify," which is in the lexicon, while "deified" is not. A direct but possibly inexhaustive way to address this is to lemmatize all tokens in a document before performing a lexicon lookup. Another related issue is how to disambiguate the parts of speech of some words in these dictionaries. For example, given the word "bully" in the MFD 2.0, should it be counted when it is verb, noun, or even adjective (which, in this case, means "excellent")? Researchers have attempted at disambiguation (Rezapour et al., 2019), but accounting for word senses alone does not mitigate the drawbacks of lacking inflections or the incompleteness of the vocabulary.

Third, longer documents tend to have a higher chance of having a dictionary match. In Figure B.1 (top panel) of Appendix B.3, we show that the number of words found in each dictionary is highly correlated with how many words each text input has (at r=0.59, 0.72, and 0.98 respectively for MFD, MFD 2.0, and eMFD). When foundation scores are normalized by the input length, such strong positive relationships disappear (Figure B.1, bottom panel), suggesting that length-normalized scoring might be preferable if one has to use dictionary-based methods. On the other hand, there are examples for which the detected moral foundation is due to one matching word in a long post—Appendix B.3 shows an example that triggers the

foundation *authority* because the word "refuse" appears once in a 140-word long post.

Finally, the hand coding of words to moral foundations by experts is a source of personal subjectivity and social bias. In the same analysis using Reddit posts, we discover some problematic associations. Figure 3.1 (right) presents the most frequently matched words that are related to two foundations: *authority* and *care*. Using MFD 2.0, some associations such as "father" with *authority* and "mother" with *care* appear to reflect the bias of the lexicon creators when assigning moral intuitions to very general familial roles. Appendix B.3.1 on page 82 contains a systematic comparison of foundation scores for posts containing "father" and those with "mother," showing that posts with "mother" scores higher for *care* (MFD and MFD2.0) and *loyalty* (MFD), and that posts with "father" score higher for *authority* (MFD 2.0) and lower on *care* (MFD 2.0). The same figure also shows that Mformer does not suffer from the same bias.

The limitations discussed here are not restricted to moral foundations. Within sentiment analysis, where lexicons such as LIWC (Tausczik and Pennebaker, 2010) are used, the same pitfalls are especially illuminating. First, word count explicitly ignores the context-dependent nature of language by relying on (normalized) frequency as a direct proxy to sentiment (Pang and Lee, 2008; Puschmann and Powell, 2018), and by treating polysemous words equivalently (Schwartz and Ungar, 2015). Second, since lexicons are static, the validity and accuracy of this method highly depend on its input's domain (González-Bailón and Paltoglou, 2015). And third, word count can be shown to predict sentiment more poorly than simply word presence, suggesting that the length of an input may influence the overall score more substantially (Pang and Lee, 2008). Machine learning approaches, especially with the advent of LLMs, can overcome these challenges through their ability to learn contextualized patterns and superior cross-domain generalizability.

Overall, dictionary-based moral foundation scoring seems fragile, lacks consensus in what they capture, and has inherent biases in social stereotypes. We caution the use of manually curated lexicons without further scrutiny and recommend examining their coverage and accuracy before interpreting aggregate results. Recognizing that developing a good lexicon takes significant effort and is difficult to get right, we adopt a data-driven approach for moral foundation scoring which avoids these limitations and can account for the nuances in language that exist beyond individual words.

3.4 Constructing and Evaluating Mformer

In this section, we describe Mformer, a language model fine-tuned from a wide range of data to score moral foundations in text. We first introduce the datasets on which Mformer is trained (Section 3.4.1). Then we describe the training procedure along with some baselines (Section 3.4.2). Finally, we present evaluation details that highlight Mformer's efficacy (Section 3.4.3).

3.4.1 Datasets

We first describe the dataset used to train and evaluate moral foundation classifiers. We combine three publicly released, high-quality data sources labeled with moral foundations.

Table 3.1: Three moral foundations datasets used to develop Mformer.

Source	Twitter	News	Reddit	Total
Data period	'10–'17	'12–'17	'20–'21	–
Number of examples	34,987	34,262	17,886	87,135
Average number of tokens	19.3	28.0	41.7	27.3
Number of Annotators	854	13	27	–
% of *authority* examples	33.4	24.9	19.2	27.1
% of *care* examples	40.6	24.8	26.5	31.5
% of *fairness* examples	35.9	24.2	29.5	30.0
% of *loyalty* examples	31.1	24.4	11.1	24.4
% of *sanctity* examples	22.3	19.9	9.8	18.8

3.4.1.1 Twitter

This dataset contains 34,987 tweets encompassing seven "socially relevant discourse topics": All Lives Matter, Black Lives Matter, 2016 U.S. Presidential election, hate speech, Hurricane Sandy, and #MeToo (Hoover et al., 2020). Annotators were trained to label the tweets with moral foundations and their sentiments (virtue and vice), with at least three annotations per tweet. We keep all tweets in this dataset for our use and determine that a tweet contains a foundation f if at least one annotator labeled it with f. Further, for each foundation, we merge the labels for its virtue and vice into one: e.g., the raw labels "purity" and "degradation" are mapped into the same foundation *sanctity*.

3.4.1.2 News

This dataset was used to construct the eMFD lexicon described in Section 3.2.2. Hopp et al. (2021) pulled 1,010 news articles, most of which on politics, from the GDELT dataset (Leetaru and Schrodt, 2013) and employed online workers to label these articles with moral foundations. Specifically, each annotator was assigned a foundation-article pair and then asked to highlight all sections in the article that contain this foundation. We segment every article into sentences and assign a moral foundation f to a sentence if any part of it is contained within a highlighted section labeled with f. This yields 32,262 instances in total.

3.4.1.3 Reddit

This dataset contains 17,886 comments on 12 different subreddits roughly organized into three topics: U.S. politics, French politics, and everyday moral life (Trager et al., 2022). In annotation, the authors separated the foundation *fairness* into two classes: *equality* (concerns about equal outcome for all individuals and groups) and *proportionality* (concerns about getting rewarded in proportion to one's merit). Another label, *thin morality*, was defined for cases in which moral concern is involved but no clear moral foundation is in place. We merge both *equality* and *proportionality* into their common class *fairness* and consider *thin morality* as the binary class 0 for all foundations, which results in the same five moral foundation labels. Finally, for

each comment, we assign a binary label 1 for foundation f if at least one annotator labeled this comment with f.

3.4.1.4 A profile of the datasets

We combine the three sources—Twitter, News, and Reddit—into one dataset, yielding 87,135 instances with 2.4M tokens. Table 3.1 on page 37 presents some summary statistics. Each example has on average 27.3 tokens, with Reddit comments the longest (41.7 tokens) and tweets the shortest (19.3 tokens). The foundations *care* and *fairness* have the most positive instances in total, each with at least 30% of the dataset. Among the three sources, tweets tended to contain more foundations than Reddit comments. For example, over 31% of tweets contain *loyalty* while only 11.1% of Reddit comments do. Finally, for each foundation, we split this dataset into a training and test set with ratio 9:1, stratified by that foundation. In Appendix B.4, we describe the datasets in more detail, including their annotation scheme and agreement rate, and how label disagreement and train-test splitting are handled. Finally, there exist other datasets labeled with foundations, such as *covid* and *congress* used by Guo et al. (2023). We choose not to include them due to their significantly smaller size—in the 1–2,000 range rather than 15,000+.

3.4.1.5 Capturing moral foundation polarity

Some prior work has additionally considered *polarity*, i.e., whether a text instance conveys a *virtue* or a *vice* of a moral foundation, resulting in ten classes (two for each foundation). In this work, we decide against this approach—instead only aiming to score the relevance of a foundation regardless of polarity—for three reasons. First and conceptually, virtues and vices are very loosely-defined terms whose perception is subject to cultural differences (Graham et al., 2013, see §2.4.4 for an example of *authority*). Second, while some previous work has treated the virtue/vice distinction as a sentiment analysis task (Hopp et al., 2021), we believe this is somewhat naïve since it lacks a theoretical justification. Third and operationally, the assignment of virtues or vices by human annotators is another source of noise on top of the noise in moral foundation labels. This is coupled with the fact that not all available datasets/lexicons capture this polar distinction. We do not argue that virtues and vices are irrelevant; rather, we believe they deserve a more thorough theoretical and practical treatment, which is beyond this work's scope.

3.4.2 Moral foundations classifiers

3.4.2.1 Mformer

LLMs have achieved state-of-the-art performance across a range of NLP benchmarks. Our work is not the first to use LLMs for this task; for example, Trager et al. (2022) fine-tuned BERT (Devlin et al., 2019) to create moral foundation classifiers. However, we note that prior work primarily focused on setting up a baseline for future work. As such, a careful treatment of the fine-tuning process and evaluation is necessary to substantiate the adoption of such methods.

We choose the RoBERTa-base architecture (Liu et al., 2019) with 12 self-attention layers for this task. Each document is tokenized and then two special tokens, <s> and </s>, are added to the beginning and end of the document, respectively. A classifier module follows the final attention layer, where the 768-dimensional embedding of the <s> token goes through a fully connected layer with 768 neurons followed by tanh activation. Finally, this is linearly mapped to a two-dimensional output vector and then converted to probabilities via a softmax layer. In fine-tuning RoBERTa, we find the optimal learning rate and the number of training epochs by performing a grid search. We end up with five binary classifiers, each of which outputs a score between 0 and 1 for every input text. We call the final fine-tuned models Mformer, for Moral foundations using transformers. More training details are found in Appendix B.5.2 on page 88.

3.4.2.2 Baselines

For comparison we consider as baselines all methods described in Section 3.2.2: *word count*, *embedding similarity*, and *supervised classifiers*.

For word count, we score a document based on the description in Section 3.3. We experiment with three lexicons: MFD, MFD 2.0, and eMFD. For MFD and MFD 2.0, we increment the foundation count by one every time its example word is encountered and then divide the count by the total number of tokens. This represents the frequency with which the foundation is found among the tokens. For eMFD, since the lexicon contains soft counts between 0 and 1, every time a word in the dictionary is found we add all scores to their corresponding foundations. Then, the five-dimensional vector of foundation scores for the document is normalized by the number of tokens that match the eMFD's entries. For all three lexicons, the foundation scores are in $[0, 1]$. More detail is found in Appendix B.1.

For embedding similarity, with each foundation f and a document d, the score for d is defined as the cosine similarity between the embedding vectors for f and d. To encode f and d, we use the GloVe embedding (Pennington et al., 2014), specifically the "Twitter" 200-dimensional version. The vector representation for f is defined as the average of the vectors for the "seed words" that represent f. Similarly, the vector for d is the average of the vector representations of all of its tokens. The range for foundation scores is $[-1, 1]$. For more detail, including the seed words describing each foundation, see Appendix B.2 on page 80.

Finally, for supervised classifiers, we train a simple logistic regression model on a range of sparse and dense document embeddings. We find that, unsurprisingly, the embedding with the best performance is Sentence-RoBERTa, which is based on RoBERTa fine-tuned for sentence similarity (Reimers and Gurevych, 2019). In Appendix B.5.1, we provide more details of logistic regression and compare it with support vector machine as used in previous work (Hoover et al., 2020).

3.4.2.3 Alternative to binary classifiers

Mformer is a collection of five binary classifiers each for one moral foundation and the corresponding set of RoBERTa weights. We also consider a weight-shared variant in which only one model is used but the final classification layer contains five neurons, each followed by

Table 3.2: Highest-scoring test examples for each foundation. Each bar chart on the right-hand column displays the scores predicted by Mformer for the five moral foundations (from left to right): *authority* (A), *care* (C), *fairness* (F), *loyalty* (L) and *sanctity* (S). The red bars represent the scores predicted for the corresponding ground-truth labels.

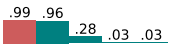
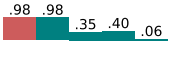
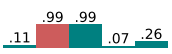
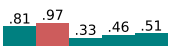
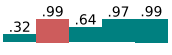
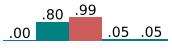
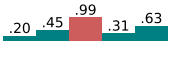
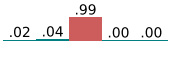
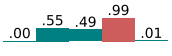
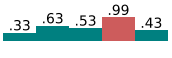
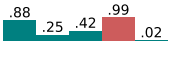
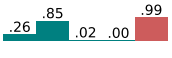
Authority	Twitter	I am a proponent of civil disobedience and logic driven protest only; not non/ irrational violence, pillage & mayhem! #AllLivesMatter	
	News	Earlier Monday evening, Pahlavi addressed a private audience and urged 'civil disobedience by means of non-violence.'	
	Reddit	Our politicians are openly encouraging rioters to harm and even kill police now	
Care	Twitter	#BlackLivesMatter SHAME ON YOU! @SenSanders is the best hope for social justice and you hurt him, you hurt me, you hurt us all. SHAME!	
	News	Just 10 days later, a gunman shot and killed three police officers in Baton Rouge, Louisiana, in what authorities called another 'ambush-style' attack.	
	Reddit	Wage slavery is indeed disgusting. Stay strong and safe comrade I wish you the best of luck. Educate agitate organize	
Fairness	Twitter	It's about humanity & equality #ChapelHillShooting #AllLivesMatter	
	News	Victims of despotism are entitled to fairness and justice, and this is the message we are conveying to the whole world.	
	Reddit	This money was taxed when it was earned, taxed when it was given and will be taxed when it is spent. Yikes	
Loyalty	Twitter	Storify. Solidarity and support #Ferguson #Blacklivesmatter [URL]...	
	News	The small rally was aimed at offering unity and solidarity for all regardless of race, ethnicity, gender, religion, sexual orientation or identity.	
	Reddit	It's never unpatriotic to criticize the POTUS. They work for us.	
Sanctity	Twitter	we value the sacred human dignity of every single life! ~@realDonaldTrump #votervaluessummit #ProLife #MAGA	
	News	He said that sexuality was to adhere to 'its God-given purpose.'	
	Reddit	Incest is disgusting and he should seek help. What a gross f***. Sorry you're dealing with this but this is sick.	
			A C F L S

Table 3.3: AUC for moral foundation classifiers (Section 3.4.2) evaluated on the hold-out test sets (Section 3.4.1).

Foundation	Word count			Embedding	Logistic regression				Mformer
	MFD	MFD 2.0	eMFD	similarity	tf-idf	spaCy	GloVe	S-RoBERTa	
Authority	0.64	0.63	0.64	0.52	0.75	0.72	0.72	0.78	0.85
Care	0.62	0.66	0.69	0.55	0.78	0.77	0.77	0.81	0.85
Fairness	0.56	0.64	0.66	0.58	0.77	0.76	0.76	0.79	0.84
Loyalty	0.57	0.59	0.60	0.51	0.76	0.74	0.74	0.77	0.83
Sanctity	0.54	0.60	0.59	0.59	0.71	0.73	0.71	0.76	0.83

a sigmoid activation. In other words, this multi-label model outputs five binary probabilities simultaneously predicting each foundation. Compared to Mformer, multi-label RoBERTa requires less storage and training resources. However, we find that this model performs uniformly worse than Mformer, achieving 10.7–19.3% lower test AUC than its binary classification counterparts (see Appendix B.5.3 on page 90).

3.4.3 Evaluation

We evaluate classification methods presented in Section 3.4.2 using the hold-out test set described in Section 3.4.1. The results show that Mformer outperforms all existing methods in scoring all foundations, often by a considerable margin.

3.4.4 Evaluation metric

It is worth noting that this dataset is multi-label: each instance can contain between zero and five foundations. Our goal is to build five classifiers each predicting the binary label of each foundation given an input. Two considerations are taken into account. First, all classifiers described in this section output a "score" representing the likelihood that a foundation exists in an input. Second, as shown in Table 3.1 on page 37, the dataset is unbalanced for all foundations with the percentage of positives being as low as 18.8%. A suitable metric should be *threshold-free* (it considers the likelihood scores and not just binary predictions), *scale-invariant* (it considers prediction scores ranked on any scale), and take into account *unbalanced class prior*. We therefore choose the area under the receiver operating characteristic curve (AUC) for evaluation. Traditionally used in signal detection theory, this metric has a useful statistical property: the AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. While the range of the AUC is $[0, 1]$, with higher values suggesting better classifiers, a realistic lower bound for this metric is 0.5, which represents a classifier that randomly guesses the positive class half of the time (Fawcett, 2006).

3.4.5 In-domain evaluation

We score all test examples using the methods described in Section 3.4.2 and report the test AUC in Table 3.3. We find that embedding similarity shows the worst performance with

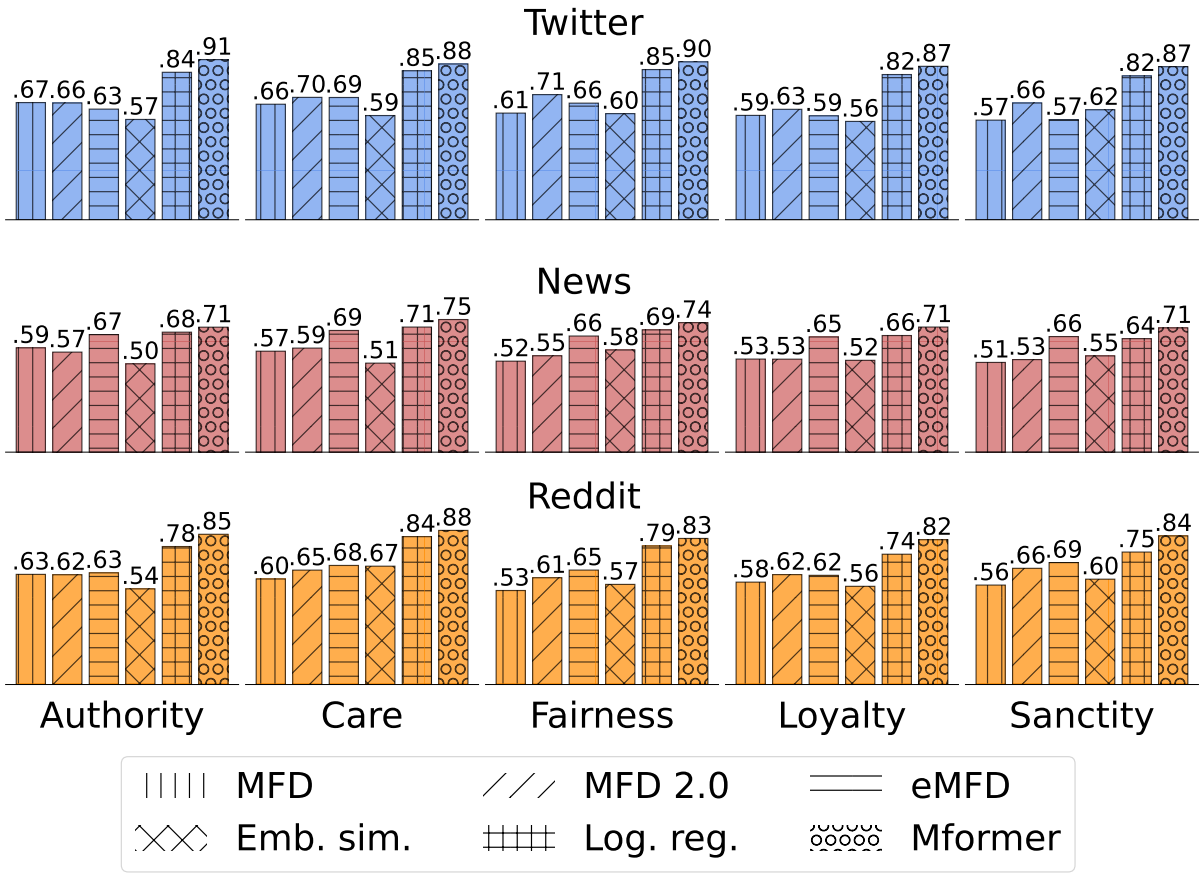


Figure 3.2: Area under the ROC curve (AUC) on the Twitter (top row), news (middle row) and Reddit (bottom row) portions of the test set for six moral foundation scoring methods: MFD, MFD 2.0, eMFD, embedding similarity, logistic regression, and Mformer.

AUC between 0.51 and 0.59 for all foundations, only slightly better than random guessing. Simple word count methods surprisingly perform better than embedding similarity, with each updated version of the MFD tending to improve from the previous. Logistic regression models further improve from word count, achieving the AUC between 0.76 and 0.81 for the Sentence-BERT embedding. The highest recorded test AUC for all foundations is by Mformer, where all foundations achieve an AUC between 0.83 and 0.85, a relative improvement of 4–12% from logistic regression. When comparing the performance across foundations, we find that *care* and *fairness* are the easiest to score (both AUC = 0.85), while *loyalty* and *sanctity* are the more difficult but not sizably (both AUC = 0.83).

Since the test sets are merged from three sources—Twitter, News, and Reddit—we examine the performance of these classifiers on each subset in Figure 3.2. The same findings hold: embedding similarity performs the worst, followed by word count (which gets better each newer version of the lexicon), then by logistic regression, and Mformer remains the best. Of all domains, tweets are the easiest to score: the AUC for *authority* and *fairness* goes up to 0.91 and 0.9, respectively. In contrast, sentences taken from news articles are the hardest: the AUC

for all foundations only ranges from 0.71 to 0.75, with the lowest for *loyalty*. We suspect that the relatively low performance on news sentences is because of the way they were labeled: In Hopp et al. (2021), annotators highlighted the *sections* of an article containing a foundation, while we subsequently process these sections using *sentence segmentation*. On the other hand, every tweet and Reddit comment was independently labeled for every foundation, which explains the quality of their labels.

Given its superior performance, we adopt Mformer as our final classifier. In Table 3.2, we show the highest-scoring examples for every foundation. The scores displayed in the right-hand column suggest that a lot of examples contain more than one moral foundation. For example, according to Mformer, the Reddit comment "Wage slavery is indeed disgusting. Stay strong and safe comrade I wish you the best of luck. Educate agitate organize" conveys *care*, *loyalty*, and *sanctity* with very high confidence. The coexistence and interplay of these foundations suggest complex moral constructs, which we will examine in Section 3.6 below.

3.5 Mformer: Out-of-Domain Evaluation

As shown in Section 3.4, Mformer demonstrates superior predictive performance to existing methods on the test set. Here, we further highlight that Mformer also generalizes well to other data domains—one in the psychology literature and three in NLP—without any further fine-tuning. Specifically, in this section, we describe each dataset in detail and discuss Mformer's performance presented in Figure 3.3. Cross-domain evaluation of moral foundations classifiers has been studied in Liscio et al. (2022); however, the "domains" in their work are only restricted to the seven topics in the Twitter dataset (described in Section 3.4.1). Given the positive results recorded, we emphasize the potential of Mformer to be adopted for many analyses of moral rhetoric based on MFT without the costly training of a new model.

3.5.1 Moral foundation vignettes (VIG)

This dataset contains 115 vignettes designed by the authors to assess humans' classification of moral foundations (Clifford et al., 2015). Each vignette is a short description of a behavior that violates a foundation. An example for *fairness* is "You see a politician using federal tax dollars to build an extension on his home." As presented in Figure 3.3, Mformer performs very well on this dataset, achieving an AUC of 0.95 for *authority* and higher than the second-best method, logistic regression, by 7–15%. The only surprising exception is *loyalty*, on which Mformer achieves an AUC of 0.75, slightly lower than logistic regression of 0.76 and equal to embedding similarity. Upon inspection, we find that some examples of *loyalty* tend to be misclassified as *authority* like the following vignette: "You see a head cheerleader booing her high school's team during a homecoming game."

3.5.2 Moral arguments (ARG)

This dataset contains 320 arguments taken from two online debate platforms (Wachsmuth et al., 2017). Kobbe et al. (2020) subsequently labeled each argument with moral foundations. On this dataset, we also observe very good results for Mformer with all AUC between 0.81

and 0.86, the highest among all methods and up to 17% higher than the AUC for logistic regression. We find *authority* and *sanctity* relatively more difficult to classify. Some instances of *authority* tend to be confused with *care*; e.g., "Some kids don't learn by spanking them.So why waste your time on that, when you can always take something valuable away from them." This is also observed for arguments containing *sanctity*—see Appendix B.7.1 for an example.

3.5.3 Social chemistry (SC)

This dataset contains 292K moral rules-of-thumb (RoTs) labeled with moral foundations, social judgment and others (Forbes et al., 2020). We use the test set and score all of its 29K instances. For Mformer, the AUC ranges between 0.70 and 0.80—highest among all methods—with specifically high AUC for *loyalty*. We also find that logistic regression comes close to Mformer, and is much better than word count methods which often perform marginally better than chance. As we explain in more detail in Appendix B.7, the relatively low performance of Mformer, compared to that observed in VIG or ARG, may be attributed to this dataset's high level of label noise. As an example, the following RoT is predicted with a very high score for *care* but does not contain this ground-truth label: "People should temper honesty with compassion, especially when it comes to family."

3.5.4 Moral Integrity Corpus (MIC)

This dataset contains 99K annotated RoTs derived from 38K responses to questions on Reddit (Ziems et al., 2022). The responses were generated by chatbots in order to facilitate the study of their moral biases. We use the test set with 11K examples for evaluation. Similar to SC, the AUC for Mformer on this dataset, ranging from 0.65 to 0.75, is relatively lower than that in VIG or ARG, but remains the highest among all methods. We also suspect that this is largely due to label noise in the dataset as the RoTs were labeled in a similar fashion to those in SC. For instance, this RoT is predicted with a high score for *fairness* but does not contain the ground truth: "It's wrong to fight in an unjust war."

3.6 Studying Moral Dilemmas and Controversies using Mformer

So far, we have established that current methods used to score moral foundation may actually perform not much better than chance, and we have advocated for the adoption of Mformer based on its good performance across a number of domains without any further fine-tuning. In this section, we present some case studies, partly replicating some previous work, that highlight Mformer's efficacy in discovering patterns of morality in several socio-political domains.

3.6.1 Moral dimensions in everyday conflicts

Recently Nguyen et al. (2022, cf. Chapter 2) analyzed over 100,000 moral discussions on `r/AmItheAsshole` (AITA), where users post an interpersonal conflict they have experienced and ask the community to judge if they are in the wrong. The authors used topic modeling to

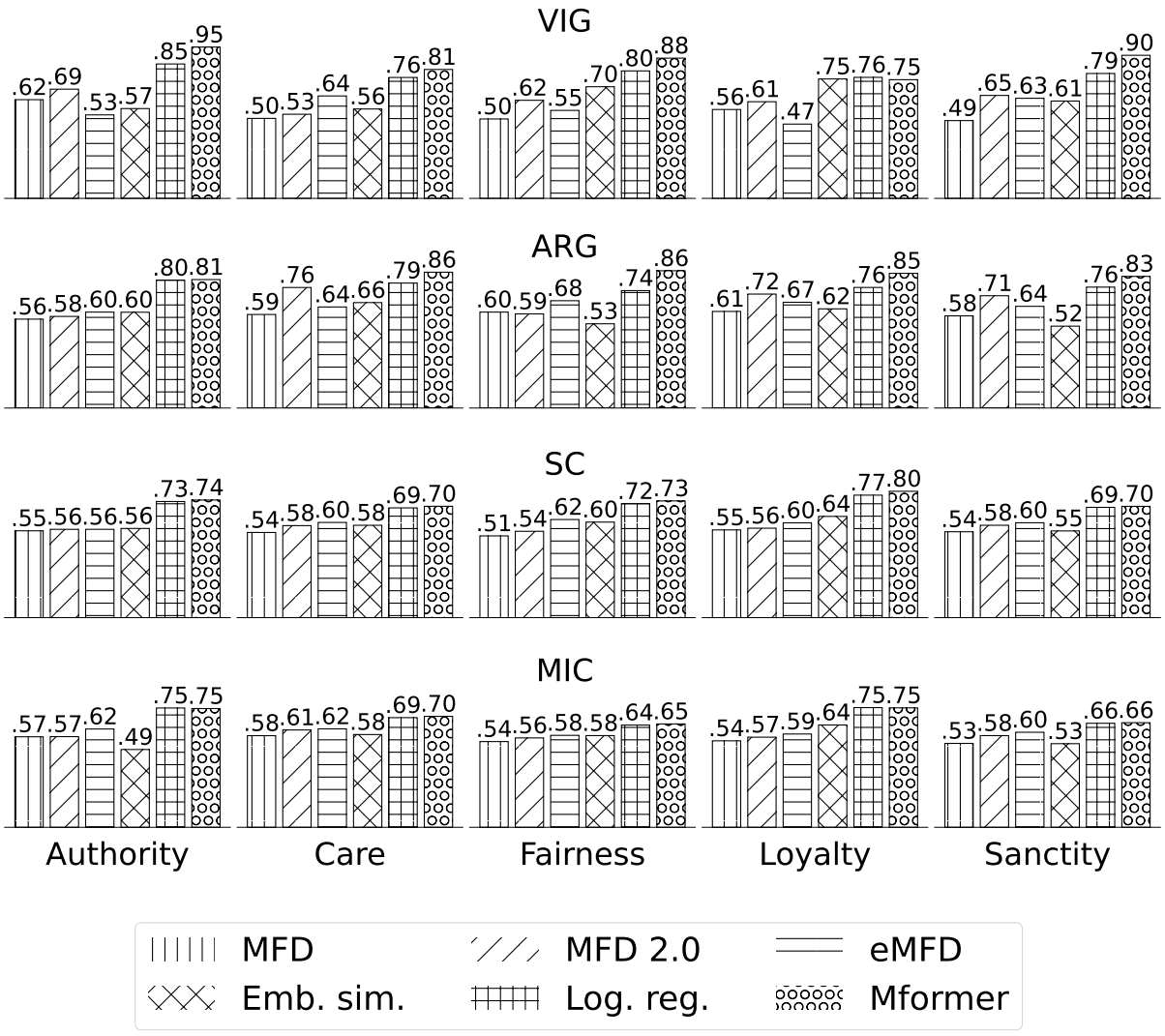


Figure 3.3: Area under the ROC curve (AUC) on four external datasets: moral foundation vignettes (VIG, Section 3.5.1), moral arguments (ARG, Section 3.5.2), social chemistry (SC, Section 3.5.3) and moral integrity corpus (MIC, Section 3.5.4). We use six different moral foundations scoring methods for prediction: MFD, MFD 2.0, eMFD, embedding similarity, logistic regression, and Mformer.

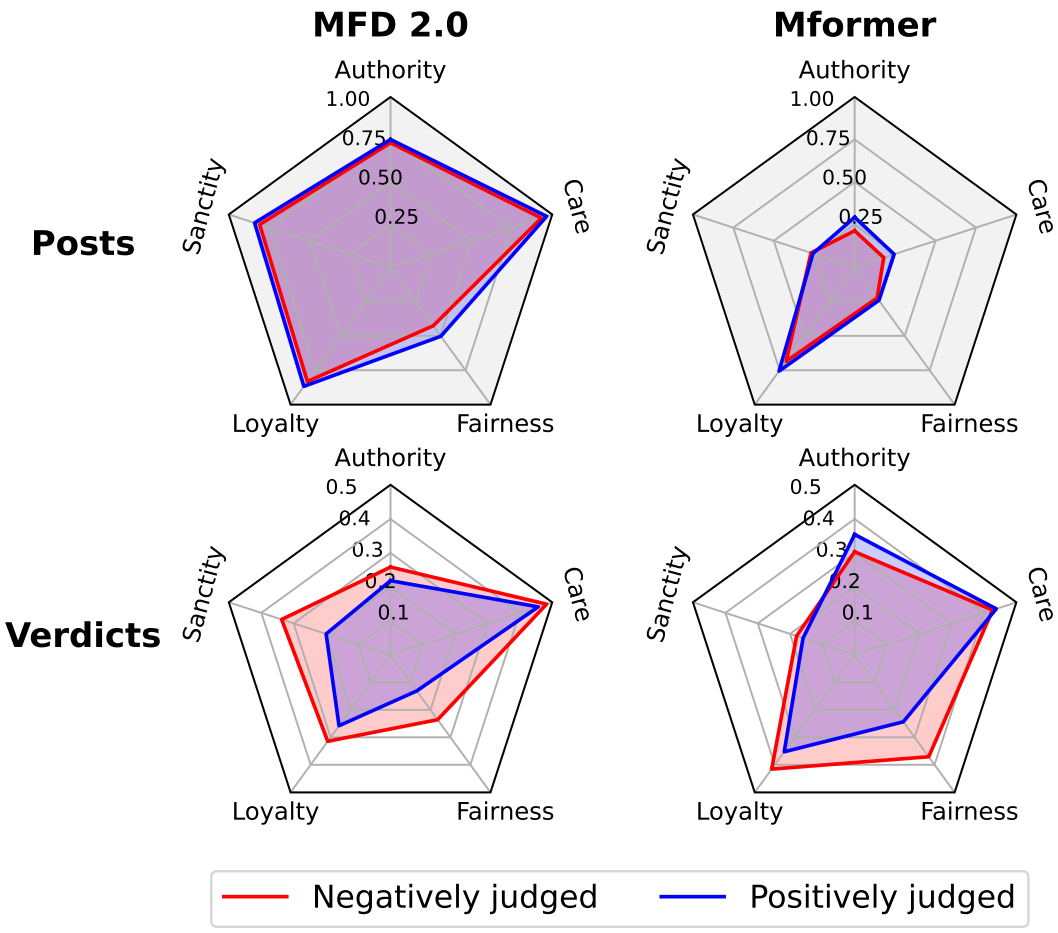


Figure 3.4: Prevalence of moral foundations in posts (top) and verdicts (bottom) in the (*family*, *marriage*) topic pair on AITA. Each number in a radar plot indicates the proportion of posts (or verdicts) that contain the corresponding moral foundation. The moral foundations are detected by two methods: MFD 2.0 (which was employed in Chapter 2) and Mformer. Red (resp. blue) indicates negative (resp. positive) verdicts.

find the most salient topics of discussion in this community and found that topics and topic pairs are a robust thematic unit over AITA content. They then used the MFD 2.0 to label all posts and verdict comments to examine the patterns of framing and judgment pertaining to moral foundations across all topics and topic pairs.

Here we replicate the same study, this time using Mformer as a moral foundation labeling method instead of MFD 2.0. Our aim is to examine whether the findings in the previous study still hold when a better classifier is used. For more detail on the setting, see Appendix B.8.1 on page 97. Similar to Chapter 2, we calculate the *foundation prevalence*—defined as the proportion of posts/verdicts that contain each moral foundation—in each topic to examine the relative importance of these five moral foundations within every sphere of moral discussion.

Figure 3.4 presents the radar plots for foundation prevalence among all posts and verdicts in the (*family*, *marriage*) topic pair. Using MFD 2.0, we find that all foundations except *fairness*

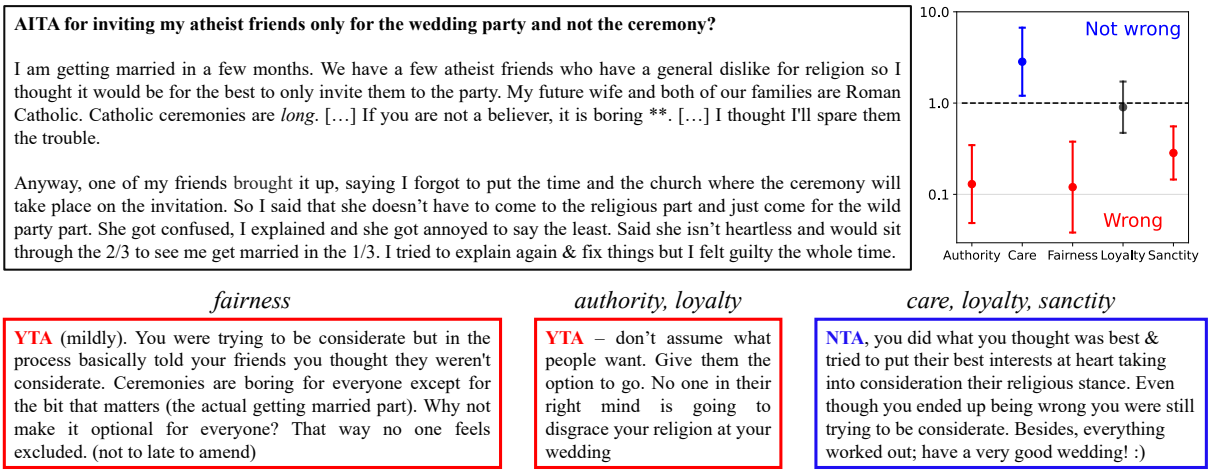


Figure 3.5: An example controversial thread on AITA. *Top left*: the post, including its title in bold and body text. *Top right*: odds ratios and 95% CIs between the presence of a moral foundation in a judgment and the judgment's valence. Values above the dashed horizontal line indicate that a foundation is associated with positive valence (i.e., "NTA" or "NAH"), while values below the dashed line indicate an association with a negative (i.e., "YTA" or "ESH") judgment. *Bottom*: three judgments for this post. The foundations contained in each judgment are annotated at the top.

are salient among posts in this topic; however, the results by Mformer indicate that only *loyalty* is dominant. For verdicts, while the results by Mformer agree with the previous finding that the foundation *care* is significant, we also find that it is *loyalty*, not *sanctity*, that is of major concern among these judgments. Differences in prevalence also hold for most of the 47 topics found in that study, which we present in Figures B.9 and B.10 in Appendix B.

These results highlight that important findings are subject to change when researchers use different scoring methods. We believe that Mformer is a good candidate for adoption as it performs better than all existing methods across numerous benchmarks (Sections 3.4 and 3.5) and is robust to its internal binary thresholds (see Figure B.8 and the discussion in Appendix B for an examination).

3.6.2 Moral dimensions in opposing judgments

Here we present another study using AITA content. In contrast to Chapter 2, where we only looked at posts and their verdicts (i.e., highest-scoring judgments), we aim to analyze *all judgments* within *one post* to find any systematic differences in conflicting judgments—those that claim the author is in the wrong and those who think otherwise. To do so, we filter the dataset to contain only "controversial" posts with at least 50 judgments, which are split somewhat equally between the positive and negative valence. This yields 2,135 posts accompanied by 466,485 judgments. A detailed setting can be found in Appendix B.8.2 on page 99.

We use Mformer to score every post and judgment on five moral foundations. To convert the scores to binary labels, we set the highest-scoring 20% of the posts to contain that founda-

Table 3.4: Results of the chi-square test for the independence of moral foundations and stance (in favor or against) toward a controversial topic. The columns denoted by "MFD" give the results presented in Rezapour et al. (2021, Table 5). The columns denoted by "Mformer" are the results based on the binary labels predicted by our Mformer models. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. All insignificant results at the 0.05 level are replaced by the "–" symbol.

Topic	Authority		Care		Fairness		Loyalty		Sanctity	
	MFD	Mformer	MFD	Mformer	MFD	Mformer	MFD	Mformer	MFD	Mformer
Atheism	–	–	–	–	–	11.82***	–	47.58***	–	46.22***
Climate	–	–	–	–	–	16.55***	–	–	–	–
Trump	–	51.77***	–	5.90*	–	–	–	34.05***	–	–
Feminist	–	–	–	–	–	–	–	–	–	–
Clinton	–	–	–	–	–	17.55***	–	–	13.49**	–
Abortion	–	–	–	12.21***	6.84*	11.05***	–	5.72*	–	15.13***

tion; the same applies to judgments. In other words, a post (or judgment) is said to contain foundation *loyalty* if it scores higher than at least 80% of all posts (or all comments) on *loyalty*. This rather high threshold is motivated by our striving for high precision at the expense of recall. For each post and a foundation f, we calculate the odds that a positive ("not wrong") judgment contains f, compared to the odds that such a positive judgment does not contain f.

Figure 3.5 displays an example controversial thread on AITA. The post received 397 positive (the author is "not wrong") and 471 negative ("wrong") judgments. The odds ratio plot at the top right suggests some distinct patterns: those that think the author is not wrong are 1.6 times more likely to focus on *care* (OR=1.57, 95% CI=[1.08, 2.28]) by arguing that the author is simply looking out for their atheist friends and helping them avoid a religious event they might be uncomfortable with. On the other hand, those that judge the author to be in the wrong are 2.5 times more likely to emphasize the foundation *fairness* (OR=0.40, 95% CI=[0.24, 0.66]) by stating that the author ought to be fair to all guests by inviting them to the ceremony as well. Negative judgments are also more likely to underlie *authority* (OR=0.41, 95% CI=[0.27, 0.63]) and *sanctity* (OR=0.58, 95% CI=[0.43, 0.77]); these judgments often argue that the author should not assume, on behalf of his friends, that they would not want to be at the ceremony just because it is religious and they are not. Finally, we find that the foundation *loyalty* is not significantly associated with positive or negative judgments (OR=0.96, 95% CI=[0.72, 1.27]); it seems that within this situation, the author's loyalty to their friends is not being questioned as much as other moral values.

Not all moral dilemmas give significant findings. Even after filtering less controversial posts, we only find, among the 2,135 controversial threads, 1,136 (53.2%) of them to have significant results where at least one moral foundation is associated with a clear valence of wrong/not wrong. Nevertheless, the results suggest that conflicting judgments for moral dilemmas can be explained by their appeal to different moral foundations, which can be robustly detected by Mformer.

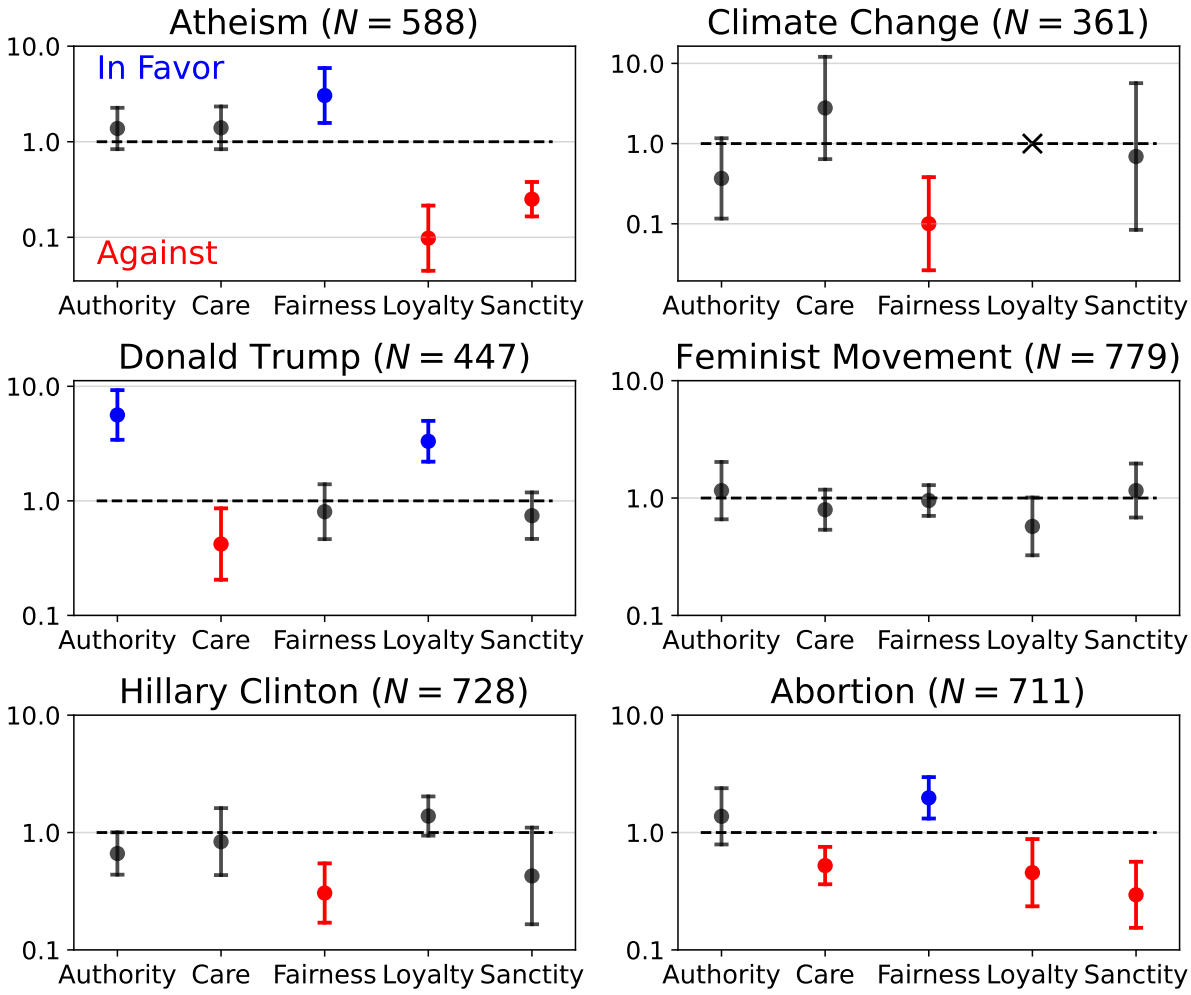


Figure 3.6: Odds ratios and 95% CIs between the presence of a moral foundation in a tweet and the tweet's stance toward each topic. An odds ratio above the dashed horizontal line indicates that a foundation is associated with the "in favor" stance. The "X" mark for *loyalty* in climate change is due to no tweets containing this foundation.

3.6.3 Moral dimensions of different stances

Stance classification is concerned with determining whether a person is in favor of or against a proposition or a topic (Mohammad et al., 2017). Here we explore the endorsement of moral foundations and its relationship with people's stance toward several controversial topics, as expressed through tweets. This study partially reproduces the analysis in Rezapour et al. (2021). We use a dataset of 4,870 tweets across six political topics: *atheism*, *climate change is a real concern*, *Donald Trump*, *feminist movement*, *Hillary Clinton* and *legalization of abortion* (Mohammad et al., 2016). Since we are only interested in polar stances (in favor or against), we remove all instances where the stance was labeled as "none" (i.e., either neutral or irrelevant to the topic). This results in 3,614 tweets in total, with the number of tweets per topic between 361 and 779.

We score all tweets on every moral foundation using Mformer and convert the raw predicted scores to binary labels by setting the highest-scoring 20% of the tweets to contain each foundation. We then replicate the chi-square analysis by Rezapour et al. (2021, Section 5.3), the results of which are presented in Table 3.4. Apart from some similar findings, such as that no association is found between moral foundations and stance toward *feminist movement*, we find many differences after using Mformer for detecting foundations instead of the MFD in the previous study. For instance, Rezapour et al. (2021) found no significant correlation within the topic *Donald Trump*, but our results show that this happens for three foundations: *authority* ($p < 0.001$), *care* ($p < 0.05$) and *loyalty* ($p < 0.001$). We suspect that these differences are largely due to the aforementioned shortcomings of word count methods: Most tweets are short, and so can easily fail to contain lexical entries, leading to zero counts for some foundations and hence false negatives.

Finally, we estimate the effect size of these associations by calculating the odds ratio (OR) between two binary variables: whether a foundation is present in a tweet and whether the tweet's author is in favor of a topic. Figure 3.6 shows the ORs for each topic over all five foundations. Almost all topics show a clear distinction in moral foundations between tweets in favor of and those against it. For example, Twitter users supporting the *legalization of abortion* are 2 times more likely to appeal to *fairness* (OR=1.98, 95% CI=[1.32, 2.97]), such as in the tweet "Good morning @JustinTrudeau. Do you plan to tell @WadePEILiberal that women on PEI deserve the right to choose? #cdnpoli #peipoli." On the other hand, those against this view are more likely to underlie the foundations *care* (OR=0.52, 95% CI=[0.36, 0.75]), *loyalty* (OR=0.45, 95% CI=[0.24, 0.88]), and, especially, *sanctity* (OR=0.29, 95% CI=[0.15, 0.56]) as in the tweet "U don't have to be a religious to be pro-life. All u have to believe is that every life is sacred." The strong association between the disapproval of abortion and the foundations *sanctity* and *care* is consistent with the finding in Koleva et al. (2012), although this prior work found mixed results when it comes to *fairness* and *loyalty*.

Another topic that allows us to compare with prior findings is *climate change*. Previously, Feinberg and Willer (2013) showed that, on social media and news outlets, the moral rhetoric surrounding the environmental discourse primarily focuses on the *harm/care* foundation. While this may be true, we do not find significant evidence that an appeal to *care* signifies a positive or negative attitude toward climate change (OR=2.78, 95% CI=[0.64, 12.07]). This could be due to the fact that *care* is salient in arguments on both sides of the discourse, but

each side portrays different framing patterns around the foundation, which is an interesting topic of examination for future work. Nevertheless, our results show that *fairness* is highly associated with the negative stance toward this topic (OR=0.10, 95% CI=[0.03, 0.38]). Tweets against climate change often focus on the unfair treatment of those who hold "alternative" viewpoints, such as in "Climate deniers is a term used to silence those pointing out the hypocrisy in the fanatical zeal on #climatetruth."

3.7 Conclusion

In this work, we examine tools for characterizing moral foundations in social media content. We show that empirical findings based on MFT are specifically dependent on the scoring methods with which researchers label their data. Furthermore, we find that these methods, especially word count programs, often perform poorly and are biased in several ways. We instead propose Mformer, a language model fine-tuned on diverse datasets to recognize moral foundations, as an alternative classifier. We highlight the superior performance of Mformer compared to existing methods across several benchmarks spanning different domains. Using Mformer to analyze two datasets on Reddit and Twitter, we demonstrate its utility in detecting important patterns of moral rhetoric, such as conflicting judgments for the same moral dilemma that depend on the specific foundations upon which participants rely.

3.7.1 Limitations

Labeling moral foundations is an inherently subjective task, as evidenced by the low inter-annotator agreement rates found in previous studies (Hoover et al., 2020; Trager et al., 2022) and by the label noise in some datasets presented in Section 3.5. This has a direct effect on models trained on these datasets, and we believe that an effort to correct label noise will be beneficial. With regard to the findings in Section 3.6, we acknowledge that the scope of morality goes beyond what is observable on social media, and internet samples may not be representative of moral life as a whole. However, we think that social media datasets are large enough in size that one is able to draw important conclusions about how certain communities describe their moral concerns and judgments. Potential misuse of the proposed model and method could include content targeting and spreading misinformation.

3.7.2 Ethical considerations

All datasets used in this work are publicly available from prior work. In all analyses, we ensure that personal or potentially self-identifying information, such as usernames or URLs, is removed. The findings we present are descriptive insofar as morality is relevant to social issues debated online and we do not make normative claims within any of the examined domains.

3.7.3 Broader perspectives

We primarily focus on the MFT because it is popular, which is in turn due to the availability of large annotated datasets. Even within this framework, the focus on how exactly each moral foundation is portrayed (e.g., as a vice or a virtue) is potentially important and can yield more fine-grained, novel results. In addition, alternative categorizations of moral beliefs based on competing theories exist and are worthy of examination. Among these, morality-as-cooperation is a rising candidate (Curry, 2016; Curry et al., 2019), providing another set of dimensions with a different theoretical foundation.

Recognizing the prevalence of word count methods in detecting moral foundations, we believe a thorough evaluation is warranted. This work establishes that, similar to other contexts such as sentiment analysis, lexicon-based word count programs often ignore contextual information, do not generalize well due to domain dependency, and may reveal social biases due to the way words are hand-chosen. Through a careful treatment of Mformer from training to (cross-domain) evaluation, we show that machine learning-based methods have the potential to overcome these challenges.

Social media is a prolific resource for studying many aspects of morality, such as what moral dimensions are emphasized on both sides of a controversial issue. Findings from these studies can inform us about important social norms that guide debate on these issues, and can have practical implications for automated content moderation within online discussion forums as well as for the understanding of moral conflicts by machines. This work aims to caution researchers interested in this direction about the limitations of the available tools for measuring moral dimensions, and provide a more robust and reproducible alternative that has been evaluated across a range of benchmarks.

Chapter 4

Conclusion

The subject of investigation in this thesis is *moral dilemmas that arise in daily life*. We adopt Driver's (1992) concept of morally charged situations to study these moral stories, and argue that such situations are as important as classical dilemmas for understanding moral life. In a sense, these ethical conflicts and corresponding actions resemble the sort of situations studied by moral philosophers: whatever action an agent commits, it will elicit both praise and blame from the community, and there does not seem to be a sufficiently satisfying resolution. As we have seen throughout the chapters, however, these daily conflicts are different in several non-trivial ways.

First, they are significantly more pervasive. The `r/AmItheAsshole` dataset studied in this thesis already captures over 100,000 posts and 8 million comments in its first six years of existence, and these numbers are growing rapidly. This observation suggests that ethical concerns among laypeople attract a lot of online attention, which in turn becomes a great source of observational data for interested researchers.

Second, despite their mere quantity, everyday moral dilemmas are often messy and complicated in their own way. AITA stories, as we have seen, involve a great deal of detail that is seldom found in classical dilemmas. They represent interpersonal conflicts among a diverse population. Told by online users, they lack the clarity of moral vignettes typically found in laboratory settings. They arise organically, at any place and time, from online discussions. Their participants, often anonymous, tell stories, make judgments and display agreement in very different ways. And they are usually much more low-stakes, compared to matters of life and death often studied in philosophical ethics.

Systematically studying these moral dilemmas requires considerable automation often in the form of a bottom-up, data-driven approach, while human involvement is generally called for to achieve consistency. In response to these two challenges, we present in Chapter 2 a novel two-stage approach involving topic modeling and human validation. Using this approach, we map the collection of over 100,000 moral stories to a set of 47 high-quality, interpretable discussion topics. Surprisingly, the "domains" from which daily moral conflicts arise tend to be nominally neutral, such as *work*, *money* and *appearance*. This suggests a nuanced view of morality among internet users, compared to stark moral dilemmas in the traditional sense such as the kidney exchange program. These topics—and, moreover, "topic pairs," the thematic unit perceived by humans—are general enough to cover almost all recorded moral stories, but fine-grained enough to show non-trivial variations in the patterns of judgments within each topic. We conclude that topics and topic pairs can serve as a useful thematic unit within this

domain and an important covariate in assessing features of everyday moral conflicts.

There exist several limitations the study in Chapter 2. Pertaining to moral dilemmas and judgments, it is unclear whether the relevant data collected from social media and discussion forums is a representative sample of "moral life" as a whole. The likely answer is no. While the collection of moral debates on AITA is the largest so far,[1] it is possible that debates about other topics are missing. As we remark in Section 2.7, AITA rules forbid the discussion of reproductive autonomy, revenge and violence, but moral dilemmas are about as likely to arise from these domains as they do from any of the discovered topics. The analyses conducted in this thesis have not considered the *temporal* aspect of moral dilemmas and judgments. For example, due to the organization of online discussions, the "first-mover advantage" is a commonly observed phenomenon where early comments tend to receive significantly more votes and hence become more influential in forming the verdict (see Tan et al. (2016) for an example). In addition, the dynamics of moral discussions, including whether or under what condition an equilibrium can be reached, have not been studied despite the feasibility of our data. Regarding the analysis of moral content on AITA or related sources, it will also be useful to account for the demographic background of the users, which may reveal the effect of social or institutional roles on the types of moral dilemmas posted online.

Finally, data-driven studies of moral content on social media can benefit from theoretical developments in different disciplines such as social psychology. They provide a basis on which observations on human behavior are grounded. Further, in light of new data, their validity may be reinforced or revised appropriately. Approaches to operationalizing these theories—that is, making them amenable to computer-based measurement and analysis—exist, but they often face several conceptual limitations. In Chapter 3, we examine moral foundations theory for analyzing moral content. Despite the popularity of its theoretical framework, we find that existing computational tools are insufficient in several important ways. In particular, (i) word count methods have very little overlap in their lexicons (see the Venn diagram in Figure 3.1) and naïvely treat the prevalence of moral foundations as the count of subjectively constructed words with a lot of limiting linguistic assumptions; (ii) embedding similarity methods do not perform well empirically and are often inconsistent due to the high dimensionality of the embedding space; and (iii) machine learning-based methods remain relatively underexplored.

In response, we introduce Mformer, a moral foundations classifier fine-tuned from a language model using about 87,000 labeled examples taken from three different sources. Through a series of evaluations including both in- and out-of-domain data, we show that Mformer predicts moral foundations consistently better than existing baselines. Finally, we use Mformer to study a range of moral stories and judgments across two social media platforms, the result of which suggests that many prior findings relying on MFT measurements may warrant revision due to the quality of their labeling model.

One of the main challenges to measuring moral foundations in text is the considerable level of noise in the training data. Although substantial care has been taken in prior work to avoid label noise, the subjective nature of this task still makes it a significant issue. Classifiers trained on these datasets are negatively affected, accordingly. While Mformer has been tested using a wide range of existing benchmarks, suggesting its remarkable generalizability, its

[1] Up to the time of the publication of Chapter 2.

adoption in analyses of moral content can always benefit from hold-out evaluation in novel domains.

With regard to theory-driven taxonomies of moral intuitions and judgments, we acknowledge that other scales exist and are worthy of exploration. As remarked in Section 3.7, morality-as-cooperation (Curry, 2016; Curry et al., 2019) is an alternative with a growing interest both from a philosophical and practical point of view. An analysis of moral content based on morality-as-cooperation may shed light on areas where moral foundations theory's explanatory power is insufficient. For example, this theoretical alternative may be able to characterize the dimensions on which opposing stances on the feminist movement diverge (see Figure 3.6).

Appendix A

Supplemental Material for Chapter 2

A.1 AITA: structure and winning verdicts

In Section 2.3, the subreddit AITA is organized into *threads*. Figure A.1 gives an example thread in the dataset. On the left of the figure, the thread starts with a *post* made by an *original poster (OP)* or *author*. The post contains a *title*, its *author's username*, *posting time* and *content* (or *body text*). We only show the title and body text here. On AITA, titles should start with "AITA" (Am I the asshole) or "WIBTA" (Would I be the asshole), and the body text further describes the author's situation.

Below each post (shown on the right of the Figure A.1) are *comments* made by the author or other people. Comments can also reply to other comments. A comment includes its *author's username* and *posting time*. In our dataset, a comment also has an ID, its parent's ID and the ID of the post it is replying to. To make a judgment, a comment's author must include one of the following five *tags*: `YTA` (the OP is at fault), `NTA` (the OP is not at fault), `ESH` (everyone is at fault), `NAH` (no one is at fault) and `INFO` (more information on the situation is needed to judge). In Figure A.1, we show two top-level comments (replying directly to the post) with different tags, and one lower-level comment (replying to another comment) without a tag.

Posts and comments can be upvoted and downvoted by community members. The *score* is the difference between upvotes and downvotes. To determine a post's verdict, AITA uses a Reddit bot that tallies all comments below that post and uses the tag within the highest-scored comment as the winning judgment. That judgment is attached to the post as a *flair*. In Figure A.1, the winning comment is shown with a gold badge, and the post is given an `YTA` flair.

A.2 Initial topic exploration

In our first attempt to find categories in moral dilemmas, we performed a bottom-up discovery, inspired by card-sorting (Spencer, 2009). The six authors of this work each read the content of 20 randomly selected posts (with each sample post read by at least 2 people), and assigned preliminary category labels (of one or a few words) according to their understanding. We were somewhat surprised to find that rather than falling into mutually exclusive categories as we initially anticipated (e.g., *money, relationship*), each post tends to have more than one label, which covers a wide range of aspects in daily life, such as *identities* (e.g., *friends, room-*

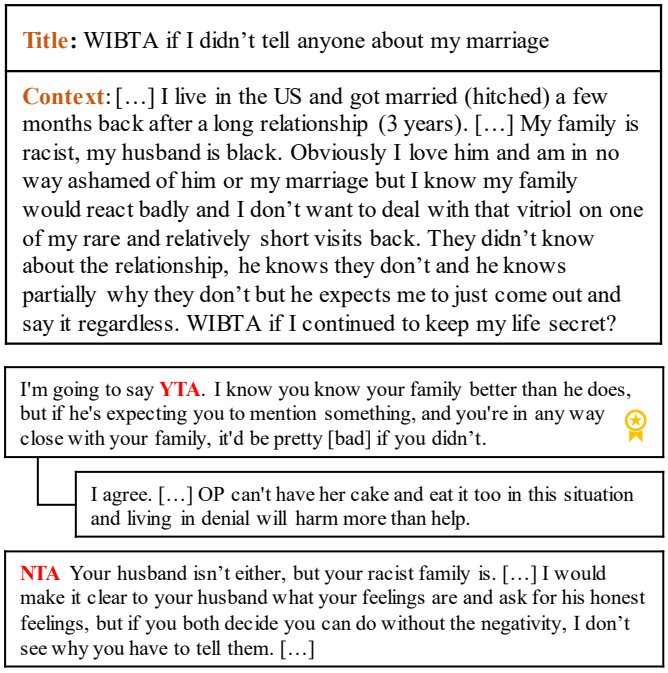


Figure A.1: An example thread on AITA containing a post (left) and comments (right). The gold badge indicates the most up-voted comment, which also becomes the winning verdict, `YTA`.

mates), *events* (e.g., *birthday, wedding*), and *themes* (e.g., *jealousy, dishonesty*). Moreover, on this small sample, the inter-annotator agreement was poor (around 30%), and the labeling practice among annotators varies from assigning a few broad labels to assigning a large number of detailed labels. One author performed a *fill-in-missing-categories* exercise. It soon became clear that the number of categories that are not present in the small 60-post sample but *could* be present can be vast, with the prevalence of it hard to estimate. For example, considering the topic of locations, the *school* label is present, but what about *gym*, *shops*, *restaurant*, *church*; and how many of each are there?

This exploration led us to conclude that manual discovery is not enough. In particular, the lack of definitions and a vocabulary for categories of moral dilemmas and assigning posts to categories prevented the progress on each other.

A.3 Topic modeling and text clustering

A.3.1 Perplexity for LDA clusters

Choosing the number of topics is a hyperparameter tuning task for LDA. To do so, we randomly split the 102,998 posts into training and validation sets of ratio 80:20. We train LDA on the training set, and rely on the perplexity of the validation set to assess the number of topics.

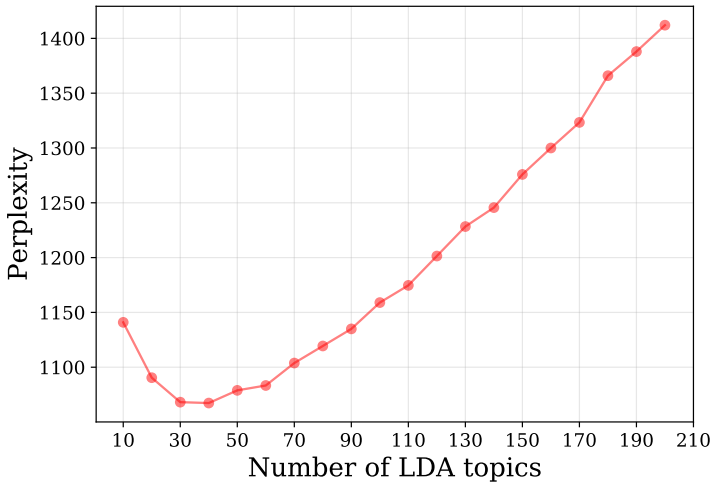


Figure A.2: Perplexity, defined in Equation (A.1), calculated on the hold-out validation set against the number of LDA topics. A lower perplexity indicates a higher hold-out likelihood, and hence a better-fit model.

The perplexity is defined as

$$\text{Perplexity} = \exp\{-1 \times \text{log-likelihood per word}\}, \tag{A.1}$$

where the log-likelihood per word is estimated using its variational bound.

We use the scikit-learn implementation of perplexity for the LDA model trained on the bag-of-words embedding (Pedregosa et al., 2011). We refer to this model as `LW`. Figure A.2 plots the perplexity against a range of LDA topic numbers. The minimum perplexity is achieved at 40 topics.

A.3.2 Other text clustering methods

Aside from `LW`, we perform clustering using non-negative matrix factorization (NMF) (Paatero and Tapper, 1994) and soft K-means clustering (Dunn, 1973).

NMF performs a decomposition of a text embedding into a product two matrices, one for the document-topic distribution and one for the topic-word distribution. The former matrix is used to find the most salient topics for each document. We employ NMF on two embeddings: the 10,463-dimensional TF-IDF (with the same vocabulary as bag-of-words, but word counts are weighted by their inverse document frequency) and the 194-dimensional Empath. Informed by the LDA outputs, we set the number of topics to 70 topics. We refer to the NMF model trained on TF-IDF as `NW`, and that trained on Empath as `NE`.

We also use soft K-means, with stiffness parameter $\beta = 1$, on the Sentence-RoBERTa (Reimers and Gurevych, 2019) embedding to mimic topic modeling: each cluster is considered a topic, and the probabilities assigned to the topic for each document represent the likelihoods of the document belonging to the topics. Informed by LDA outputs, we also use 70 clusters in training. We denote this model `KB`.

Table A.1: Most and least coherent topics, along with their top word lists, for each model. The coherence score is the UMass metric, defined in Equation (A.2).

Model	Word list	Coherence
Most coherent clusters		
LW	friend, good, group, talk, hang, like, time	-1.25
NW	thing, time, like, love, try, year, month	-1.20
NE	family, children, home, domestic work, wedding, party, friends	-0.44
Least coherent clusters		
LW	bf, event, volunteer, bc, military, join, anime	-8.47
NW	wife, law, marry, brother, marriage	-4.03
NE	strength, power, healing, masculine, violence, exercise, legend	-2.09

A.3.3 Topic coherence

When displaying topics to users, each topic is generally represented as a list of 5 to 20 words, in descending order of their topic-specific probabilities. Suppose that cluster k is described by its word list V_k, of size M. For words x_m and $x_l \in V_k$, let $D(x_m)$ be the number of documents in k with at least one appearance of word x_m, and $D(x_m, x_l)$ be the number of documents containing one or more appearances of both x_m and x_l). The topic coherence (Mimno et al., 2011) of cluster k is defined as

$$C(k; V_k) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(x_m, x_l) + 1}{D(x_m)}. \tag{A.2}$$

The addition of 1 on the numerator is to avoid logarithm of zero. Intuitively, a coherent set of words should have word pairs appearing in the same document. The coherence score compares the pair appearances $D(x_m, x_l)$ against the individual appearances $D(x_m)$ over all possible pairs of words: the higher this ratio, the more *coherent* this set is. Without adding 1 to the numerator, the coherence $C(k; V_k)$ is always non-positive, and the closer it is to zero, the more coherent k is.

We use an implementation of topic coherence called `tmtoolkit`.[1] The models for which coherence is calculated are `LW`, `NW` and `NE`. (There is no importance-based ordering of the vocabulary in each `KB` cluster.) For each `LW` cluster, we simply use the 25 words with highest posteriors $p(x_m|k)$ to represent a cluster. For `NE` and `NW`, we use the topic-word matrix after factorization and choose 25 words with highest weights for each cluster.

Table A.1 presents the most and least coherent clusters, along with their 7 most salient words and topic coherence. We note that the least coherent cluster for `LW` also is a cluster we named *other* in Section 2.4.

[1] https://tmtoolkit.readthedocs.io

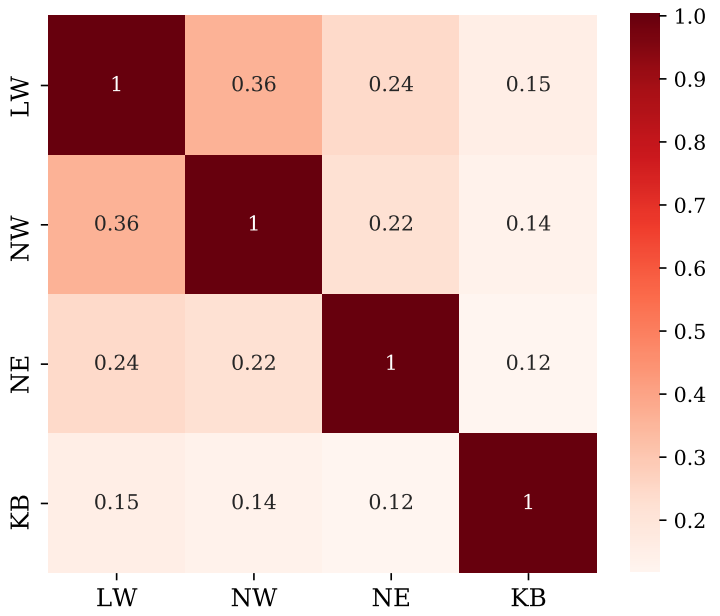


Figure A.3: Adjusted mutual information (AMI), defined in Equation (A.3), between four different clusterings: `LW`, `NW`, `NE` and `KB`. AMI ranges from 0 (no matching) to 1 (perfect matching).

A.3.4 Observations on clusters

The `LW`, `NW` and `NE` methods all create uneven clusters, from those with fewer than 50 posts to those with nearly 10,000 posts. On the other hand, `KB` clusters are mostly even in size, with each encompassing about 1.5% of the posts.

For `KB`, while most clusters have similar sizes, the overlapping of clusters' content is common. Also, while each `LW` cluster tends to co-occur with one other `NW` cluster, there are many `KB` clusters which correspond to multiple `LW` clusters. For example, the following `KB` cluster with the 10 most probable keywords {*friend, like, date, talk, relationship, guy, year, time, feeling, girl*} tends to go with multiple other `LW` and `NW` clusters, including those about communication, relationships and dating.

We find that the Empath embedding used in `NE` is useful in discovering some clusters, such as {*work, business, occupation, white_collar_job, office, college, meeting, giving, school, blue_collar_job*}. However, it is not informative enough, compared to bag-of-words and TF-IDF, in finding finer-grained topics. For example, an `NE` cluster about communication co-occurs with `NW` and `LW` clusters on relationships, communication and social activities, making it too broad.

Finally, we observe that `LW` tends to give the most coherent clusters on our dataset. `LW` clusters are informative enough, in that their word lists are consistently related. They tend to correlate well with `NW` clusters, but have more meaningful keywords.

Considering each algorithm as a hard clustering by assigning each training example its cluster with the highest posterior probability, we compare the clusterings using the *adjusted mutual information* (AMI) metric. Specifically, let U and V be two clusterings, the AMI between

them is

$$\mathrm{AMI}(U, V) = \frac{I(U, V) - \mathbb{E}[I(U, V)]}{\frac{1}{2}[H(U) + H(V)] - \mathbb{E}[I(U, V)]}, \tag{A.3}$$

where $H(U)$ is the entropy of U, $I(U, V)$ is the mutual information between U and V, and $\mathbb{E}[I(U, V)]$ is the expected mutual information. AMI ranges from 0 (no matching) to 1 (perfect matching). See Vinh et al. (2010) for more detail. Figure A.3 presents the AMI between all pairs of clusterings. The largest AMI between two clusterings is that between `LW` and `NW`, corroborating our previous observation that `LW` clusters tend to co-occur with `NW` clusters. `KB` clusters tend to match the least with other clusterings.

A.4 Survey for topic naming

In line with the literature on topic modeling (Boyd-Graber et al., 2017), especially on LDA, we find that naming a cluster should not simply be based on the cluster's most salient keywords. As pointed out in the following survey, we find some clusters having somewhat related words to humans but, after carefully reading some of their representative posts, we decided they should not be considered topics. As a result, we set up some criteria for a cluster to be given a topic name.

To be considered a topic, a cluster should have the following properties:

- its keyword list should be unambiguous in suggesting the topic's theme;
- its posts should be about the same topic when read by humans; and
- ideally, several human readers should agree on the topic name, given its keyword list and some posts.

To this end, we organized an annotation task among six authors of this work, asking ourselves to give a short (1- to 2-word long) name for each cluster. As described in Section 2.4, we used the 70 clusters found by LDA; therefore, there are 70 questions in total. Below we describe the components of this survey in more detail.

Table A.2: LDA clusters and their 10 most salient keywords. The topic names are from Section 2.4.

Topic name	Cluster size	10 most probable keywords in ascending order of $p(x \mid k)$
appearance	1,664	wear, hair, dress, look, like, shirt, clothe, shoe, makeup, color
–	137	trash, tattoo, review, bug, garbage, bin, product, design, star, throw
babies	893	baby, child, pregnant, kid, month, birth, pregnancy, week, husband, time
breakups	317	ex, break, partner, year, relationship, month, cheat, date, contact, new
celebrations	2,484	birthday, gift, buy, like, year, present, thing, christma, day, card
children	1,257	kid, wife, son, child, old, year, young, boy, parent, nephew
communication	7,855	talk, like, try, thing, time, start, upset, way, apologize, come
–	1,448	text, message, send, texte, day, respond, reply, week, time, talk
–	448	speak, language, english, customer, sorry, coffee, mistake, mobile, write, country
–	52	jane, behavior, session, john, anna, discuss, kate, conversation, issue, therapist
damage	291	book, fix, damage, replace, pool, repair, read, new, paint, accident
death	508	grandma, die, grandmother, pass, grandparent, family, mom, funeral, year, fiance
drinking	338	drink, bar, water, beer, bottle, alcohol, night, drunk, glass, wine
driving	1,494	car, drive, ride, pick, gas, driver, hour, way, minute, uber
–	1,047	road, light, turn, stop, pull, lane, way, walk, car, run
–	666	park, spot, car, parking, street, lot, space, driveway, garage, people
education	1,829	school, parent, year, college, high, graduate, job, live, university, work
entertainment	348	watch, movie, tv, coworker, film, office, worker, like, theater, work
family	1,328	sister, dad, mom, parent, year, old, family, sibling, live, young
–	830	family, people, member, parent, like, holiday, come, time, year, thanksgiving
–	717	mum, drug, year, life, dad, mother, abuse, father, die, donate
–	478	mom, brother, parent, mother, old, young, come, little, family, live
–	232	daughter, cousin, aunt, uncle, family, old, year, parent, come, young
food	1,631	eat, food, lunch, like, bring, weight, buy, try, ice, pizza
–	635	cook, meal, dinner, meat, food, like, vegan, eat, chicken, dish
friends	5,959	friend, good, group, talk, hang, like, time, year, close, people
gaming	1,270	play, game, video, team, time, player, like, start, character, sport
gender	419	woman, man, gay, male, female, mark, come, people, gender, straight
housework	1,466	clean, dish, leave, wash, laundry, thing, clothe, kitchen, mess, chore
hygiene	423	bathroom, shower, walk, smell, toilet, tooth, pee, nail, paper, brush
jokes	2,037	joke, like, people, laugh, comment, funny, thing, mean, fun, look
living	782	door, open, building, window, close, knock, leave, come, apartment, neighbor
–	415	room, living, space, bedroom, share, stuff, come, guest, small, kitchen
–	138	house, live, home, stay, buy, parent, rule, housemate, leave, let
manners	2,229	start, shit, yell, fuck, like, scream, fucking, try, throw, come

–	2,131	sit, seat, people, walk, look, lady, bus, gym, stand, wait
marriage	1,403	husband, mother, father, year, child, law, divorce, marry, family, mil
medicine	921	doctor, hospital, pain, sick, surgery, appointment, medical, need, help, cancer
mental health	2,132	issue, like, anxiety, mental, health, try, people, thing, time, help
money	4,505	pay, money, buy, month, job, work, help, need, spend, save
–	1,455	sign, pay, agree, charge, edit, check, state, situation, month, claim
–	528	ticket, sell, buy, card, concert, band, price, win, free, sale
music	592	music, tip, listen, song, play, loud, hear, server, like, headphone
other	6,694	time, try, thing, life, help, year, like, good, talk, bad
–	25	bf, event, volunteer, bc, military, join, anime, army, organization, attend
–	16	girlfriend, christma, secret, christmas, gf, year, eve, max, pen, santa
parties	1,071	party, friend, night, drunk, leave, people, come, police, happen, invite
pets	1,947	dog, puppy, walk, time, bark, owner, let, like, care, leave
–	964	cat, pet, animal, adopt, care, kitten, shelter, vet, love, feed
phones	201	phone, number, jack, app, answer, look, check, cell, screen, iphone
race	424	guy, girl, white, dude, black, look, people, racist, race, boy
relationships	6,817	date, relationship, like, guy, meet, thing, talk, girl, love, year
–	1,661	boyfriend, time, come, stay, spend, visit, live, night, like, week
–	58	club, dance, red, flag, camp, practice, beach, bff, strip, competition
religion	990	church, believe, religious, people, opinion, religion, god, view, christian, belief
restaurant	859	order, restaurant, dinner, table, food, place, come, leave, pay, cake
roommates	2,263	roommate, live, apartment, rent, month, place, pay, lease, stay, new
safety	78	lock, steal, key, bike, door, milk, unlock, leave, locker, dh
school	2,063	class, teacher, school, student, study, grade, test, project, exam, group
sex	346	sex, kiss, sexual, porn, time, condom, like, start, stop, try
shopping	1,023	store, bag, stuff, leave, pick, box, item, shop, grocery, buy
sleep	1,678	sleep, bed, night, wake, morning, asleep, couch, room, time, stay
smoking	796	smoke, neighbor, weed, yard, cigarette, property, fence, smoking, tree, neighborhood
social media	1,778	post, picture, photo, facebook, social, medium, people, send, like, friend
technology	490	email, computer, laptop, report, internet, work, code, need, meeting, camera
time	4,570	work, day, time, hour, week, come, weekend, leave, night, plan
vacation	1,620	trip, vacation, plan, travel, hotel, day, time, week, year, flight
wedding	1,979	wedding, invite, marry, plan, party, friend, attend, fiancé, year, day
–	213	account, ring, throwaway, propose, mate, bet, fiancée, password, engagement, boat
work	4,642	work, job, company, boss, manager, new, coworker, time, office, people

Question: What is the name of this topic?

Topic words: **car, drive, ride, pick, gas, driver, hour, way, minute, uber**

1. AITA for not letting dad drive my car?

> My dad is working abroad, like 400km from my city (he can go by the train) and he wants me to give him my car. The problem is couple months ago he fucking crashed my car 10 minutes from home on a straight highway without no cars, because he was sleepy. Now he is mad I won't lend him my car. Am I really the ass hole? […]

2. AITA for leaving someone with no gas?

3. AITA For not allowing my parents to drive my car?

4. WIBTA if I didn't go to the airport?

5. AITA for not helping a man whose car broke down?

6. WIBTA if I report my roommates broken down car as an abandoned vehicle?

Figure A.4: An example question in the cluster naming survey. The annotator is given a list of topic words (in bold) and six posts. Each post consists of its title (in blue) and body text which can be revealed by clicking the title. Due to space constraints, only one post is shown here. The annotator is asked to give a name to this topic in one or two words by typing in the shaded box at the bottom.

A.4.1 Choosing the keywords for each cluster

In this topic model, each cluster is described by a probability distribution $p(x|k)$ over the entire vocabulary, where x is a word and k is a topic. More salient words are given higher probabilities. The vocabulary is sorted by this probability to find the top words for each cluster. We choose the top 10 words to describe a cluster, which is in line with other work in the literature (Boyd-Graber et al., 2017, and references therein). In addition, we find that on average, only 11 of each cluster's top words have $p(x|k) > 10^{-2}$. Table A.2 in Appendix A lists all 70 clusters along with their 10 most probable keywords. The keywords on each row are presented in decreasing order of topic-word probability.

A.4.2 Choosing the example posts for each cluster

As described in Section 2.4, the cluster sizes range from 16 to 7,855 posts, over the entire 102,998 posts in the training set. A very small number of posts might not sufficiently describe a cluster, while a very large number is not feasible for a human annotation task. We decide to use six posts to describe each cluster. Three of the posts were chosen randomly from the posts with highest topic-post posterior probabilities; we call them *clear* posts. The other three were chosen randomly from posts with lowest posteriors, and are called *mixed* posts. This task was done to ensure that the post lists contain both posts which may seem clear to annotators about

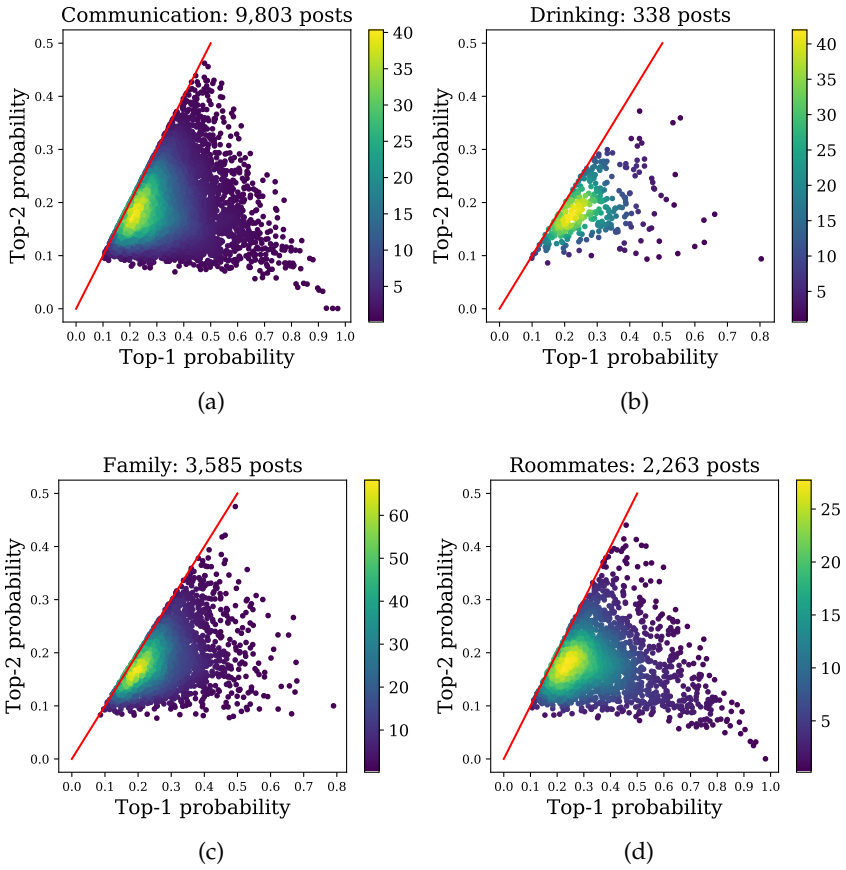


Figure A.5: Scatter plot of the top-1 and top-2 LDA probabilities for four topics. The red diagonal line represents all points whose top-1 and top-2 probabilities are equal. The density of the points is estimated using a Gaussian kernel with bandwith determined by Scott's rule (Scott, 2015).

their topic name, and posts which may confuse the annotators.

A.4.3 Question format

An example question is given in Figure A.4. First, we present the cluster's 10 most probable keywords as a list on top. Then each of the cluster's six posts is given below it. In each post, we present its title first, then followed by its body text. We only show the first 100 words of the post's body, to ensure the six posts do not take over too much space in each question. We made one observation that the first 100 words are sufficient in describing the post's content. If the text exceeds this threshold, we simply use "[...]" to denote the rest. To save space, we only show the body text of the first post in Figure A.4. Finally, the text box where the annotators give their answers at the bottom of the question is omitted in the figure.

A.4.4 Organizing annotation among authors

Six authors of this work participated in this annotation task. We designed the survey using Qualtrics.[2] To avoid long survey time, we allowed each author to participate multiple times, each with 10 randomly chosen questions. If an author gives more than one answer to the same question (due to randomly chosen questions), only the earliest answer is kept. We kept participating until 3 answers from 3 different authors were recorded for each question, resulting in 210 answers in total.

A.4.5 Post-annotation discussion of results

We determine that consensus is reached if at least 2 of the 3 answers for each question agree. Two authors met to discuss the survey results and categorized the consensus of the 3 answers in each question into four types:

- *Unanimous*: all 3 answers are identical. The topic name is set to the answer. There are 17 clusters of this type.
- *Wording*: all 3 answers are synonymous or are very close in meaning. An example is (*celebration*, *gifts*, *celebrations*), after which the topic is named *celebrations*. There are 41 clusters of this type.
- *Deliberate*: there is agreement between at least 2 answers, but after carefully looking at the topic, we decided to rename it. An example is (*family*, *family*, *family*), which initially was named *family* but later changed to *death*, because most of the posts in this cluster are about the passing of family members. There are 9 clusters of this type.
- *Other*: there is disagreement among the 3 topics. For instance, one question received (*entertain*, *relationships*, *army*). We decided to name all these clusters *other*. There are 3 clusters of this type.

A.4.6 Results and discussions

After revising the clusters' names, we resulted in 47 named topics and one *other* topic. The total number of posts with a named topic is 96,263, accounting for 93.5% of posts in the training set. The number of topics reduced from 70 to 48 because of overlapping names. We merge clusters with the same name together into a *topic*. The highest number of overlapping names is 5, for topic *family*. Table A.2 lists all LDA clusters, described by their keyword lists, and their sizes along with the names found above.

We note that because of a considerable number of clusters falling into the *wording* and *deliberate* categories, it is easy that one annotator alone comes up with a different set of topic names, or considers a cluster as a meaningful topic when it should not be. The post-annotation step, conducted by more than one annotation, was an important part of this task.

[2] https://www.qualtrics.com/

A.5 Crowd-sourced validation of topics

To assess how well the topics found in Section 2.4 describe a post's content, we design a crowd-sourced study to verify the names for a large number of posts in the dataset. This section provides additional information to that described in Section 2.5.

A.5.1 Question format

Each question provides participants with one post and five answers for the topic of the post. An example question in the survey is found in Figure 2.5. Each question contains a prompt: *"What topics below best describe the theme of the following post? Do not let your ethical judgment of the author affect your choices here."* Below the prompt is the question, starting with its title and body text (which is called "context" in the figure). For brevity, we omit some details of the example post's body text. Finally, the five options appear at the end of the survey. The first four options are topic names found in Section 2.4. A participant can choose one or more topics in the first four options. The final option is *None of the above*, which the participant can choose when the no topic satisfactorily describes the given post's theme.

A.5.2 Choosing the posts for the questions

We focus on the 47 named topics in Section 2.4, omitting the topic *other*. To ensure each topic has posts in the survey, we randomly sample posts in each topic. We initially conducted two surveys of the same format but with posts from the training set and test set:

- Training set: These are posts in the training set (posts before 2020) for LDA in Section 2.4, of size 102,998. For each topic k of the 47 topics, we randomly selected 20 posts whose most probable topic is k (based on posterior probability). The result is $47 \times 20 = 940$ posts, or 940 questions in the question bank. We call this survey **train** in Table 2.1.
- Test set: These are posts in the test set (posts in the first four months of 2020), of size 5,294. For each topic k, we randomly selected 10 posts. Since some topics do not have enough 10 posts, the result is only 450 posts, as opposed to $47 \times 10 = 470$ posts. We call this survey **test** in Table 2.1.

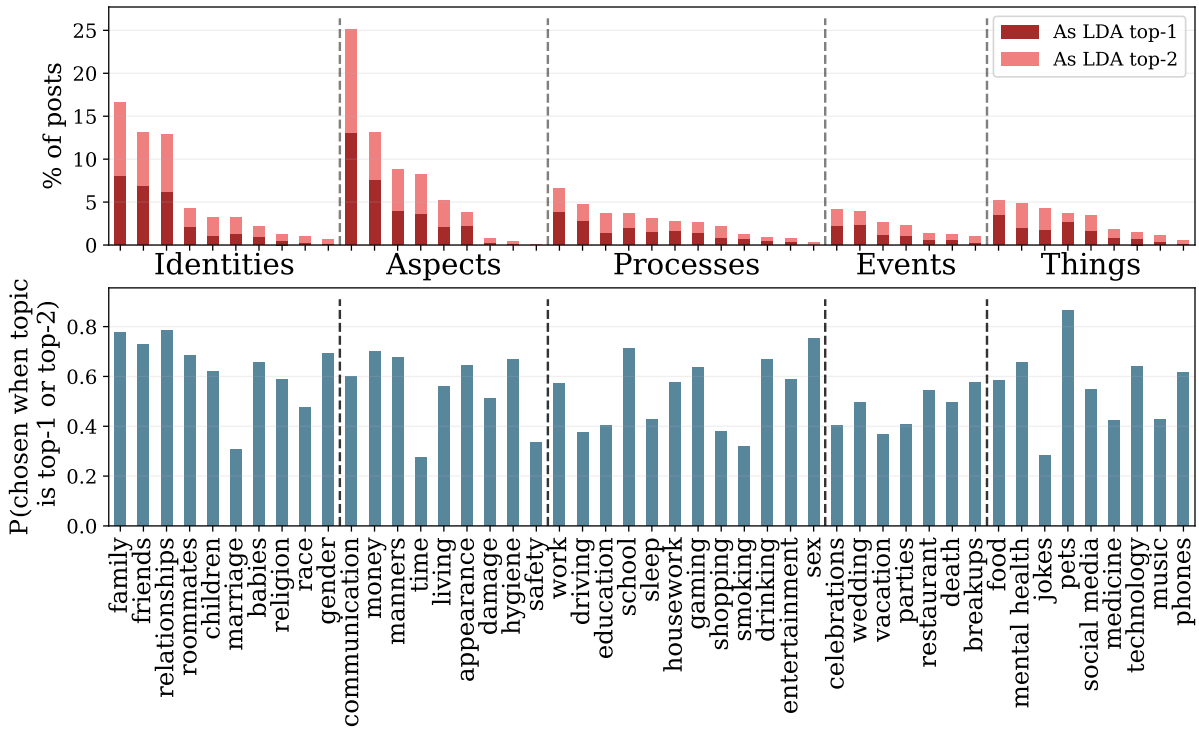
Doing so guarantees that every topic has posts in the survey, and allows us to compare the agreement rates among topics later.

A.5.3 Choosing answers for each question

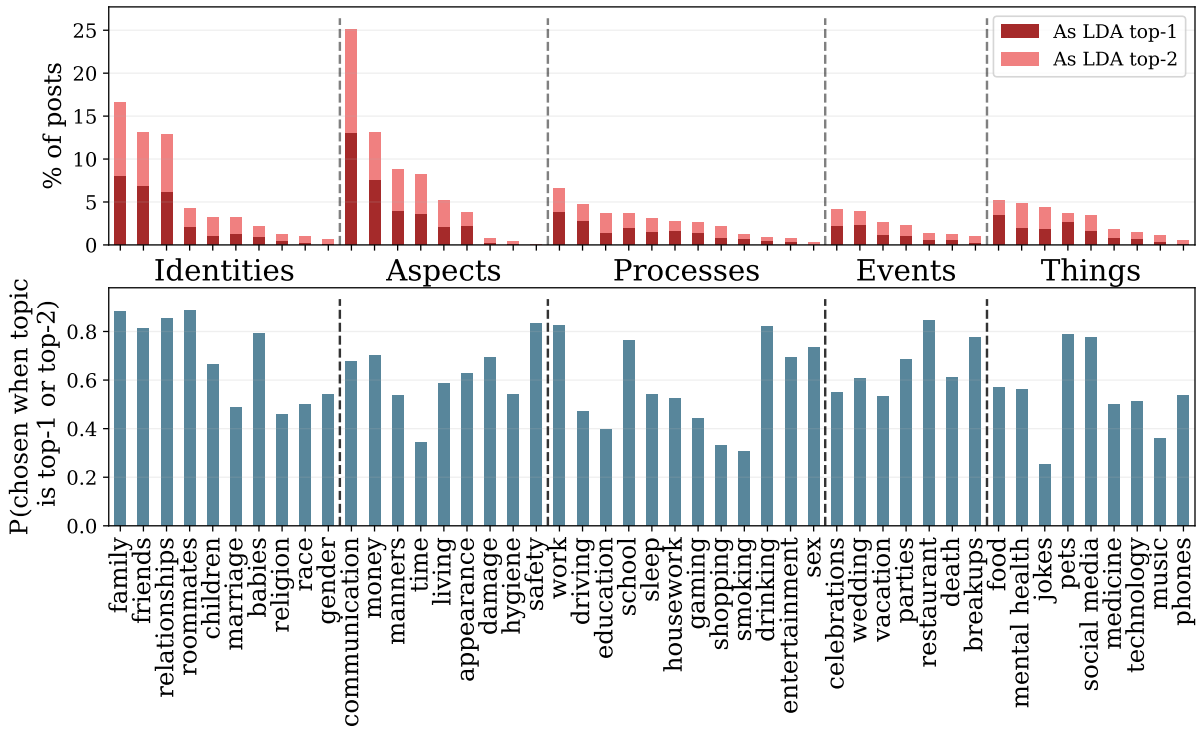
For each post in the **train** and **test** surveys, we chose the four most probable topics, according to the LDA posterior, for that post. If any of these four topics happens to be the topic *other*, we replace it with the fifth most probable topic.

In Section 2.5, we also examine the effect of replacing the top-3 and top-4 LDA topics with randomly chosen topics. To do so, we used the 450 posts in the **test** survey, and for each post, the top-2 LDA topics were kept, while the other two were randomly chosen from the rest. This created a new survey version, which we call **test+rand** in Table 2.1.

Finally, in each question, the four topics are randomly ordered in the answers.



(a)



(b)

Figure A.6: Prevalence (as a percentage of posts) and topic-specific agreement rate of topics in the *test* set, comprising the first four months of 2020. (a) The agreement rate is from the **test** setting. (b) The agreement rate is from the **test+rand** setting.

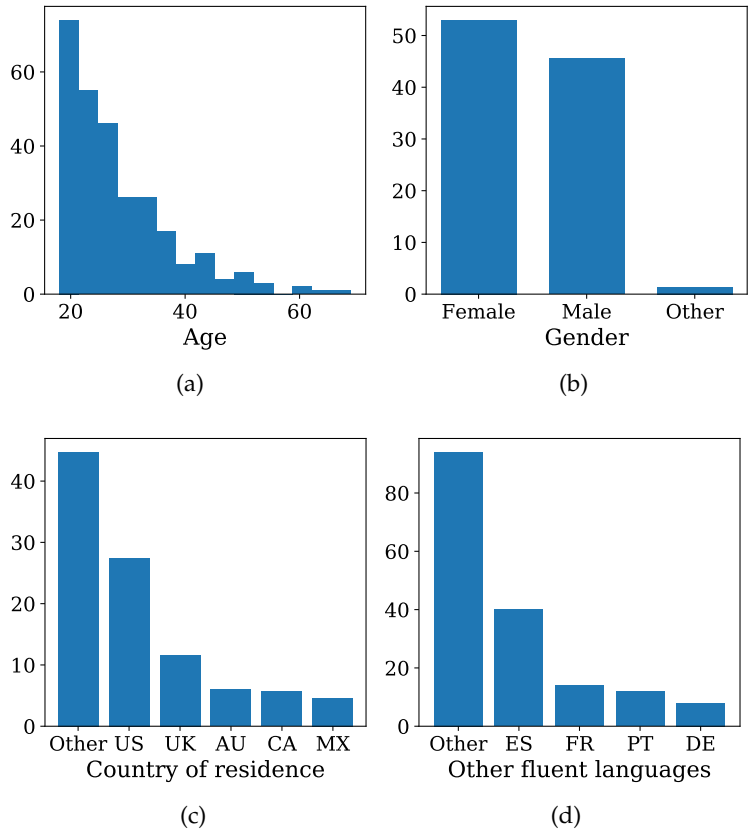


Figure A.7: Demographic information of the 285 participants on Prolific. (a) Distribution of participants' ages. (b) Gender shares. (c) Most popular countries of residence. (d) Languages they are fluent in, other than English.

A.5.4 Survey setup and implementation.

We custom-designed the survey using the SurveyJS library[3] and hosted the website on Heroku.[4] To recruit participants, we used the Prolific platform[5]. We recorded participants' demographic information, including their age, gender, residence, and fluent languages. We enforced one entry requirement that each participant must be fluent in English. A summary of this information is given in Figure A.7.

Thanks to a large number of participants on Profilic, we could allow each participant to enter the survey once. Each participant was given 20 randomly chosen questions from the question bank. As a reminder, there were three question banks, **train**, **test**, and **test+rand**, as described in the previous subsections. We recruited participants until all questions were answered 3 times by 3 different individuals. The remuneration was £2.5 for each participant, which averages to £0.125 per question. We expected participants to take 20 minutes each time,

[3]https://surveyjs.io
[4]https://heroku.com
[5]https://www.prolific.co

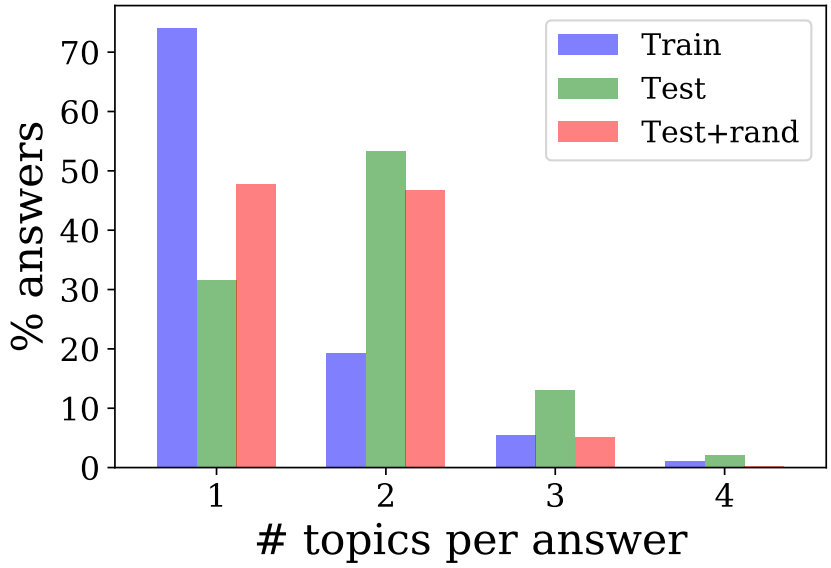


Figure A.8: Numbers of topics chosen in each answer and their shares in *train*, *test* and *test+rand* settings.

resulting in the average pay of £7.5/hour. This amount is recommended by Prolific, and is higher than the minimum pay of £5/hour. Finally, this survey had received ethics approval from the authors' institution before being rolled out.

Figure A.6 shows some demographic information about participants. In total, we recruited 285 participants, with an average age of 28.2 (SD = 9.2). 130 of the participants are male and 151 are female. The most popular countries of residence for participants are the US (27.5%) and the UK (11.6%). Of all other languages participants are fluent in, Spanish and French are the most popular.

A.5.5 Agreement rates

After collecting responses, we had 3 answers per question. We report two types of agreement rates, described below.

A.5.5.1 Post-level agreement rates.

The quantities reported in Table 2.1 are called post-level agreement rates, defined as the number of times an answer type appears in an answer, divided by the total number of answers for all questions. This rate is multiplied by 100 in the table. For example, the agreement rate on the **top-1** row is the number of times the LDA top-1 topic appears in an answer, divided the number of answers. So, for the **test** survey, which has $450 \times 3 = 1,350$ answers in total, 59.2% of the answers contain the LDA top-1 topic. On the **top-1 or 2** row, this refers to the number of answers which contain either the LDA top-1 or top-2 topic, or both.

A.5.5.2 Topic-specific agreement rate.

The quantities reported in the lower bar plots of Figures 2.4 and A.6 are called topic-specific agreement rates, which is defined for each topic. It is defined as the number times topic k appears in an answer, over the answers whose questions contain k as either the LDA top-1 or top-2 topic. For example, suppose we wish to find the agreement rate for topic *education*. There are X number of questions with *education* being the top-1 or top-2 topic, totaling $X \times 3$ answers recorded. Out of these $X \times 3$ answers, there are Y answers which contain *education*. The agreement rate is $Y/(X \times 3)$.

A.6 Topic pairs

In Section 2.5.2, we see that most dilemmas cover more than one topic. Particularly, we observe that the top two topics for each post do a much better job of describing the post's content, compared to its top-1 topic alone. In this section, we provide more information about topic pairs.

We focus on *unordered* topic pairs. This means that a post with (top-1, top-2) topics of (k, k') is in treated in the same group as a post with (top-1, top-2) topics of (k', k). While the posterior probabilities indicate the order of salience between these top 2 topics, human experts find it difficult to find the correct order. In other words, posts are generally described by their two highest-scoring topics, but their order of salience is not easily recognized.

A.6.1 Distribution of topic pair sizes

Similar to individual topics, topic pairs vary significantly in size. We use the (log) complementary cumulative distribution (CCDF) function to display the distribution of these sizes. Out of $\binom{47}{2}$ unordered topic pairs, we find 33 pairs (3.1%) without a post, and only 259 pairs (24%) with at least 100 posts (Figure 2.6). Finally, Figure A.9 shows all named topics and the other topics they are most frequently associated with.

A.6.2 Topic co-occurrence frequencies

To assess how often a pair of topics k and k', we use two methods described below.

A.6.2.1 Point-wise mutual information

Let $p(k)$ be the proportion of posts in the dataset whose top-1 topic is k, and $p(k, k')$ be the proportion of posts whose top-1 and top-2 topics are k and k', in either order of salience. If topics k and k' independently occur together, their joint distribution should be $p(k)p(k')$. To assess how more or less often k and k' co-occur than if they are independent, we compare their joint distribution $p(k, k')$ against the product of their marginals $p(k)p(k')$, in a metric called *point-wise mutual information* (PMI) (Section 2.5.3):

$$\text{PMI}(k, k') = \log_2 \frac{p(k, k')}{p(k)p(k')}.$$

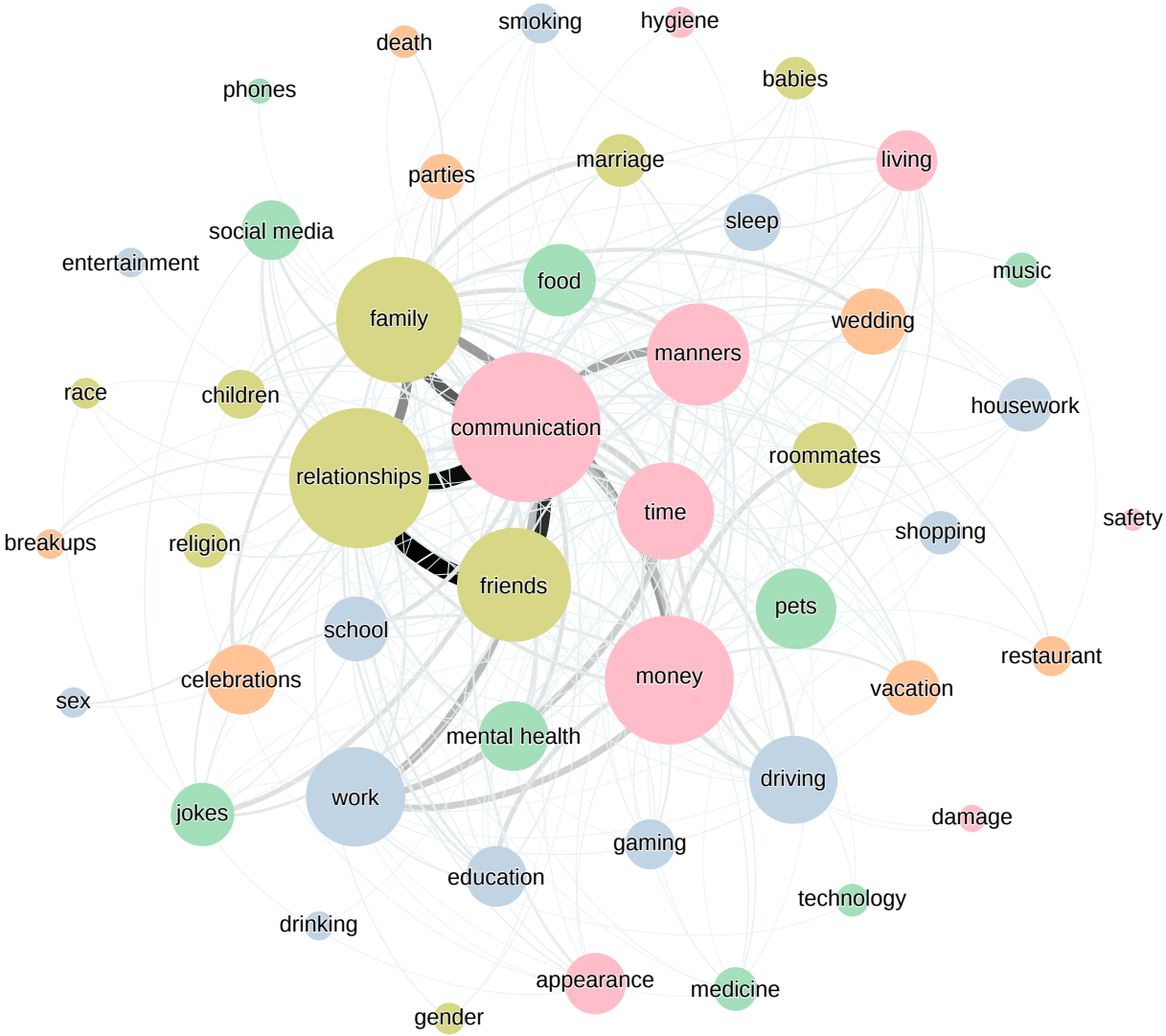


Figure A.9: A network of co-occurring topics on AITA. Topics are discovered and validated in Section 2.4. Node size denotes the number of posts and color represents its meta-category (yellow: *identities*, pink: *aspects*, blue: *processes*, orange: *events* and green: *things*). Edge width is proportional to the number of posts in each topic pair. Only topic pairs with more than 100 posts are shown.

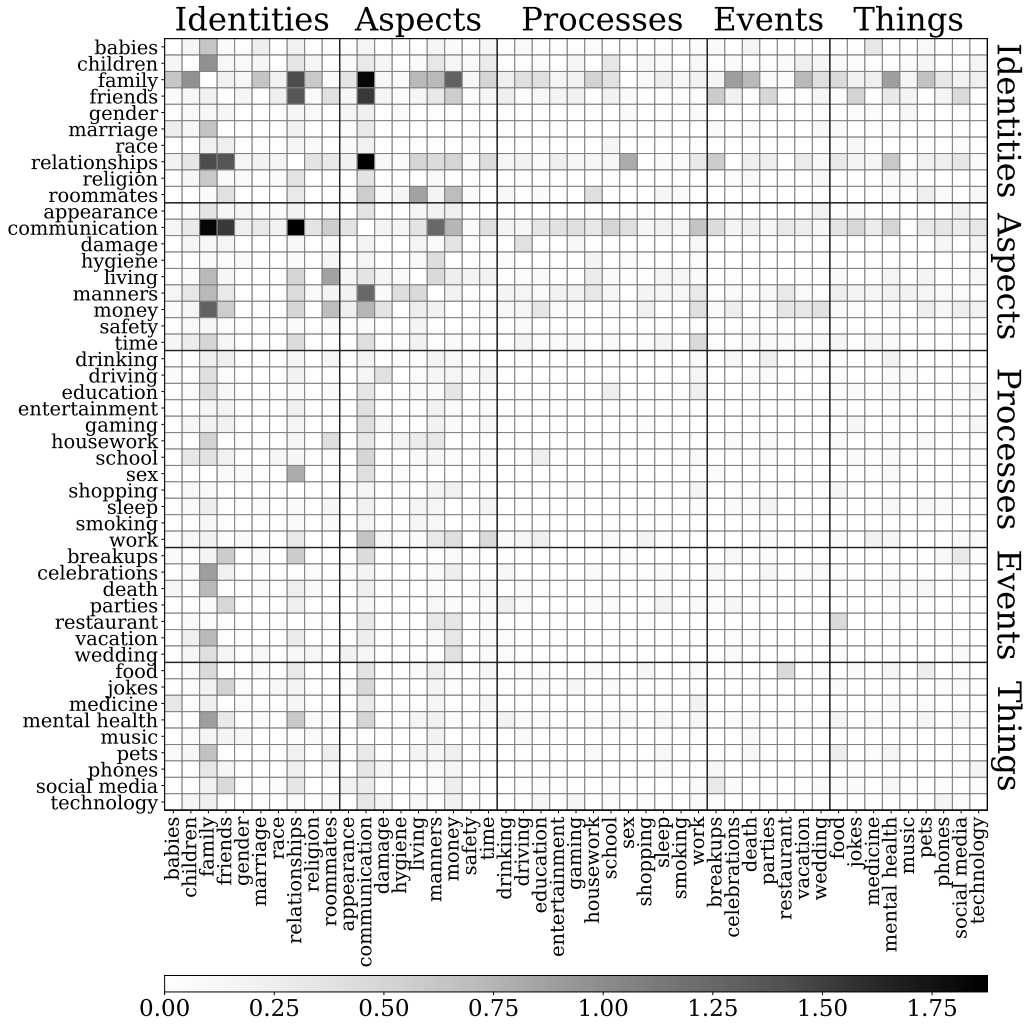


Figure A.10: Topic co-occurrence in human answers. Each cell represents the frequency (as a percentage of all pairs) of a topic pair which appears in answers in the survey in Section 2.5. Rows and columns are organized into meta-categories.

Figure 2.7 shows the PMI between all pairs of topics. Positive PMIs (red cells in the figure) correspond to pairs that are more likely to co-occur when assuming independence, whereas negative PMIs (blue cells) indicate less likely pairs.

A.6.2.2 Topic co-occurrence in human answers

The PMI is used to see topic co-occurrence within LDA. Using human input from the survey described in Section 2.5, we can measure which topic pairs tend to be chosen together by crowd-sourced workers.

Specifically, we look at answers containing at least two topics. For each answer, we extract every topic pair (so an answer of length 3 gives $\binom{3}{2} = 3$ pairs). Figure A.10 shows the frequency of each topic pair (as a percentage of all recorded pairs). As seen from the figure, most pairs

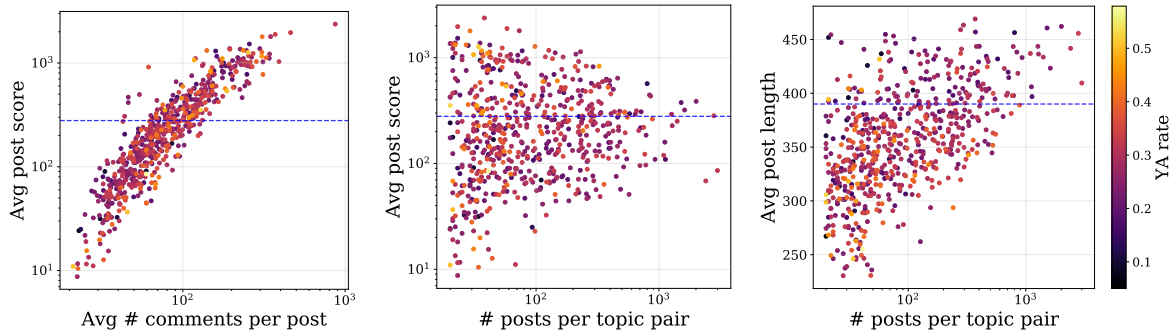


Figure A.11: Scatter plots for topic pairs. (a) Average number of comments per post versus average post score. (b) Number of posts versus average post score. (c) Number of posts versus average post length (in words). Topic pairs are colored by their YA rate. A blue horizontal line indicates the *y*-axis mean over all posts.

chosen by human annotators are in the *identities* and *aspects* meta-categories. On the other hand, topics in *processes*, *events* and *things* do not tend to co-occur with other topics in the same meta-category. Of all topics, *communication*, *family* and *relationships* co-occur the most with other topics. This is expected, as these topics are relatively large in size.

A.6.3 Voting and commenting patterns of topic pairs

Figure A.11 examines the judgment and voting statistics for topic pairs. We observe that the average post score (by voting) is positively correlated with the average number of comments (Figure A.11a). This is unsurprising in hindsight as both may be driven by the level of attention on a thread. Figure A.11b shows that smaller topic pairs have a wide spread of average topic scores, from around ten to several thousand, whereas large topics have scores around the global mean. Figure A.11c shows that smaller topic pairs have a wide spread of average post length, from 250 to 450 words, whereas large topics have lengths around 400. We do not see a salient trend regarding the average YA rate for topic pairs.

A.6.4 Additional statistics on topics and topic pairs

Table A.2 lists the 70 LDA clusters, their mappings to the 47 named topics, their size (in the number of posts) and coherence scores.

Figure A.9 visualizes the (undirected) network of topics, by connecting two topics that have more than 100 posts in the corresponding topic pair.

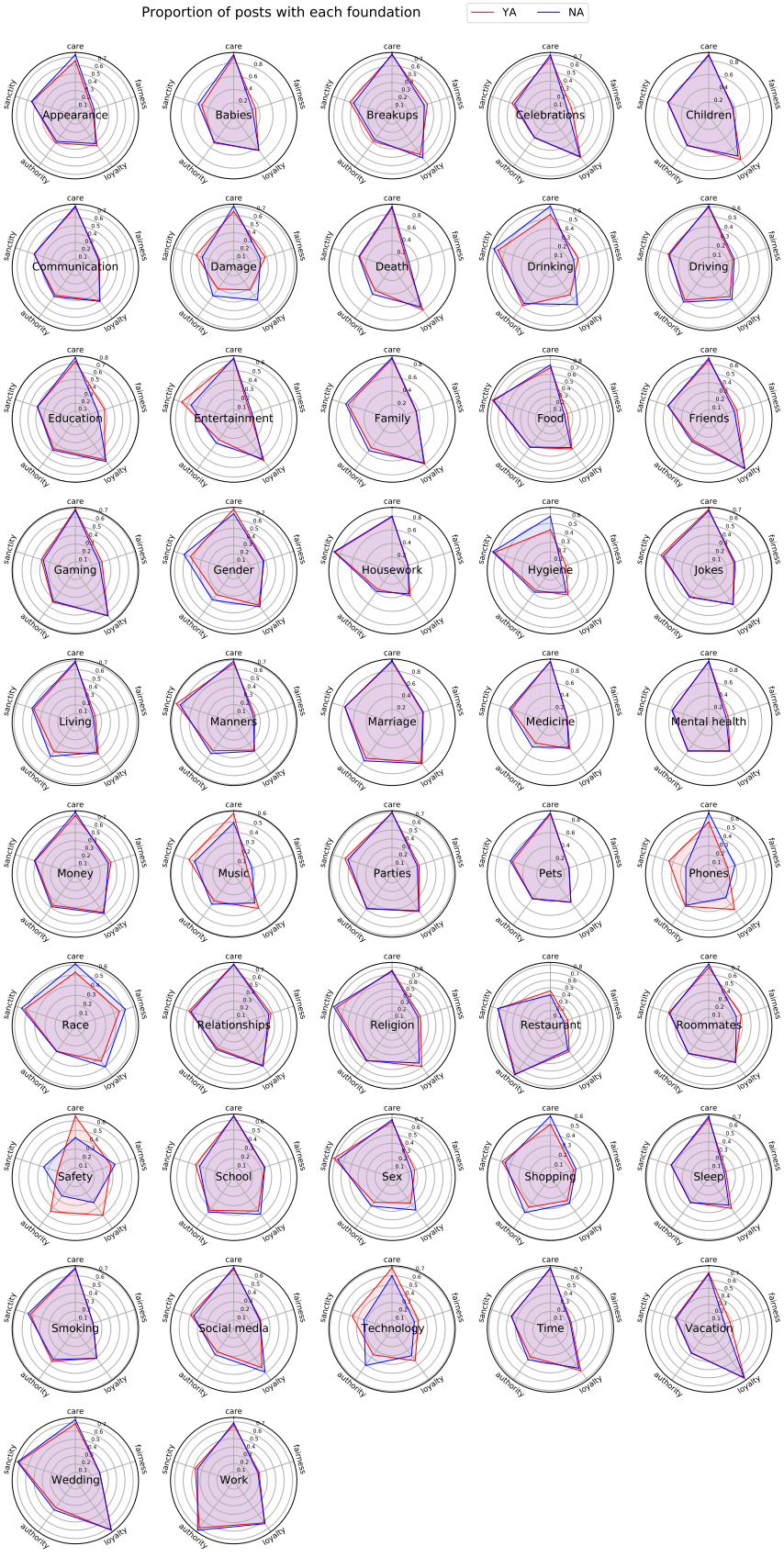


Figure A.12: Strengths of moral foundations in each topic's post. In each radar plot, each pentagon's vertex represents the proportion of posts in a topic that have the presence of at least that foundation. Red pentagons represent YA-judged posts; blue pentagons represent NA-judged posts.

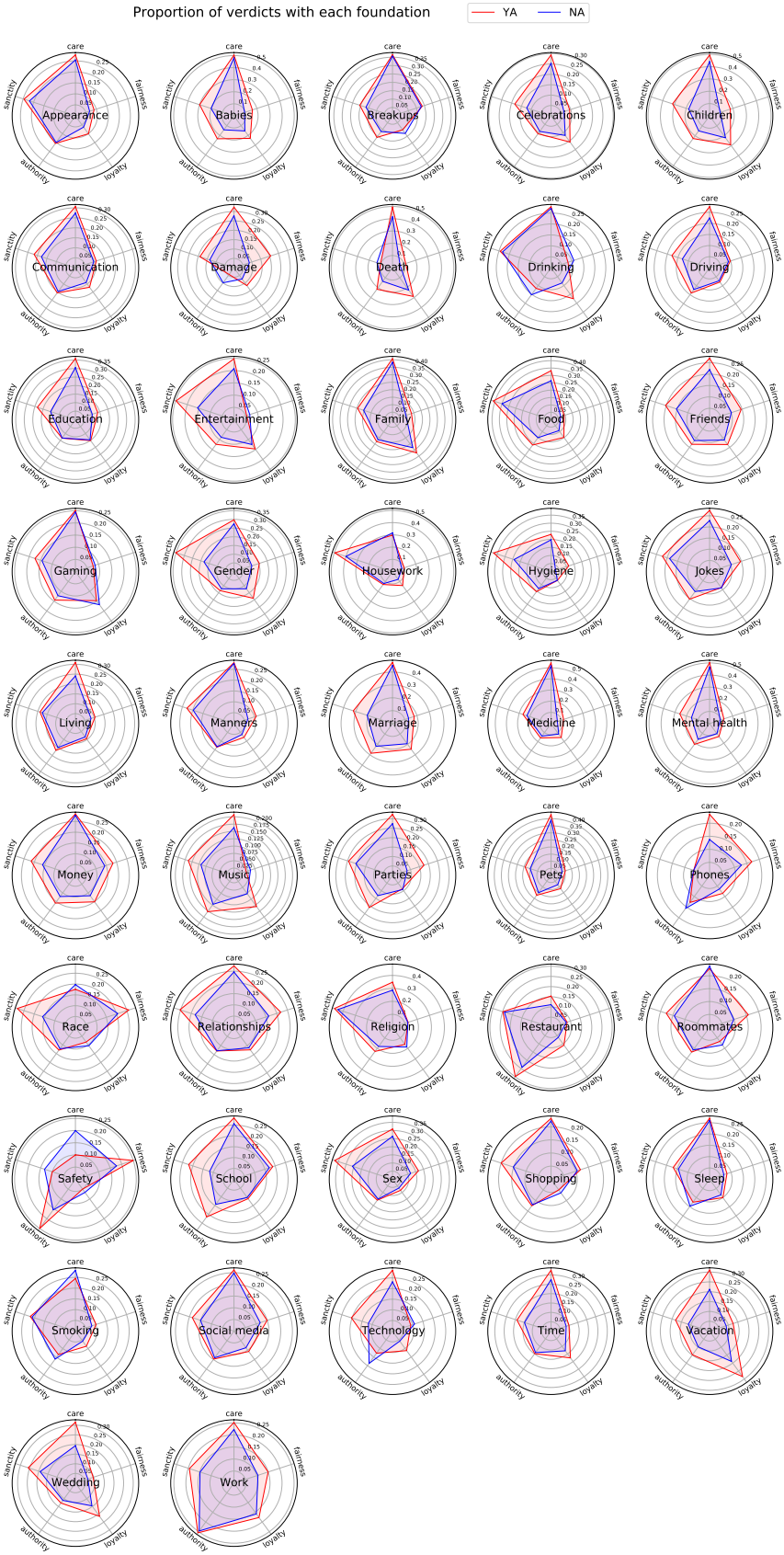


Figure A.13: Strengths of moral foundations in each topic's verdict. In each radar plot, each pentagon's vertex represents the proportion of verdicts in a topic that have the presence of at least that foundation. Red pentagons represent `YA` verdicts; blue pentagons represent `NA` verdicts.

Appendix B

Supplemental Material for Chapter 3

B.1 Moral Foundations Dictionaries

Here we provide more details on the three moral foundations dictionaries (MFDs) described in Section 3.2.2. Some example words in each dictionary are presented in Table B.1. We also describe in detail how a document is scored on five moral foundation dimensions using these dictionaries.

B.1.1 Moral foundations dictionary (MFD)

Graham and Haidt (2012) released the original version of the MFD to accompany their research on moral foundations (Haidt, 2013; Graham et al., 2013). The MFD contains author-compiled 325 entries mapping a word prefix to a foundation and a sentiment. An example mapping is `venerat* -> AuthorityVirtue`, indicating that any word beginning with "venerat"—such as "venerating," "venerated" or "veneration"—will be mapped to the virtue of *authority*. The authors also included another category called "morality general," which we exclude from our method. Finally, as Mokhberian et al. (2020) have previously converted all prefixes to eligible words, so we use their version of the MFD for word count.[1] This contains 591 words in total.

Since we do not model moral sentiments (virtue and vice), we merge the virtue and vice for each foundation together, giving us five original foundations. To score each document, we first perform tokenization using `spaCy` (Honnibal et al., 2020, model `en_core_web_md`, version 3.1.0). Then we go over each token to check whether it or its lemma is contained in the MFD. If there is a match, we increment the count of the corresponding foundation by one. For example, if the document contains the word "venerated," the count *authority* is incremented by one. The result is five counts where each count represents the number of times we encounter a foundation in this document. Finally, these counts are divided by the total number of tokens to give frequencies, or *scores*. For instance, a score of 0.05 for *loyalty* means that 5% of the tokens in this document are mapped to *loyalty*.

Note that a prefix/word can be mapped to more than one foundation. For example, "exploit" is mapped to *care* and *sanctity*. In cases like this, we increment the count of all eligible foundations.

[1] The dictionary is available online at https://github.com/negar-mokhberian/Moral_Foundation_FrameAxis.

Table B.1: Some example words in three lexicons used for scoring moral foundations: MFD, MFD 2.0 and eMFD. For eMFD, the words are sorted by their corresponding foundation weights so that the highest-scoring words appear first. For the intersection of these lexicons, see Figure 3.1.

Foundation	Lexicon	Example words
Authority	MFD	abide, command, defiance, enforce, hierarchical, lawful, nonconformist, obey, obstruct, protest, rebel, subvert, tradition, urge, violate
	MFD 2.0	allegiance, boss, chaotic, dictate, emperor, father, hierarchy, illegal, lionize, matriarch, overthrow, police, revere, subvert, uprising
	eMFD	elect, throwing, ambitions, sponsored, rebellion, protested, voluntarily, denounced, immigrant, separatist, banning, counter, interfere, nationals, protesting
Care	MFD	abuse, benefit, compassion, cruel, defend, fight, guard, harmonic, impair, killer, peace, protect, safeguard, violence, war
	MFD 2.0	afflict, brutalize, charitable, discomfort, empathized, genocide, harshness, inflict, mother, nurse, pity, ravage, torture, victim, warmhearted
	eMFD	tortured, pocket, cruel, harsh, sexually, raping, hostility, income, persecution, stranded, knife, drivers, imprisonment, killed, punishments
Fairness	MFD	balance, bigot, constant, discriminate, egalitarian, equity, favoritism, honesty, impartial, justifiable, prejudge, reasonable, segregate, tolerant, unscrupulous
	MFD 2.0	avenge, bamboozle, conniving, defraud, equity, freeload, hypocrisy, impartial, liar, mislead, proportional, reciprocal, swindle, trust, vengeance
	eMFD	rigged, undermining, disproportionately, compensation, discrimination, punished, steal, throwing, flawed, instances, excessive, restrictive, inability, discriminatory, tariffs
Loyalty	MFD	ally, clique, comrade, deceive, enemy, familial, guild, immigrate, joint, membership, nationalism, patriotic, spying, terrorism, united
	MFD 2.0	allegiance, backstab, coalition, fellowship, group, homeland, infidelity, kinship, nation, organization, pledge, sacrifice, tribalism, unpatriotic, wife
	eMFD	disrupt, wing, dictatorship, betrayal, loyalty, legislators, outsiders, renegotiate, fearful, credibility, pocket, couples, exploit, rage, retribution
Sanctity	MFD	austerity, chaste, decency, exploitation, filthy, germ, holiness, immaculate, lewd, obscene, pious, repulsive, sickening, taint, virgin
	MFD 2.0	abhor, befoul, catholic, exalt, fornicate, hedonism, impurity, leper, marry, nunnery, organic, pandemic, repulsive, trashy, waste
	eMFD	raping, sexually, swamp, rigged, exploit, tissue, objects, rape, infections, sex, uphold, wing, victimized, smoking, sacred

B.1.2 Moral foundations dictionary 2.0 (MFD 2.0)

In an attempt to extend the word lists in the original MFD, Frimer (2019) started with a list of prototypical (author-compiled) words for each foundation. Then, each list was extended by adding words that are close to the prototypical words; here, "closeness" refers to the cosine distance between word embeddings in the word2vec space (Mikolov et al., 2013). The newly found words then went under manual validation by the authors who decided whether to keep or remove each word for every foundation. The result is a collection of 2,104 words in the lexicon. We also merge the virtue and vice of each foundation into one, resulting in five unique labels. The scoring of documents is done identically to MFD.

The scoring of a document is done identically to MFD. Specifically, we merge both sentiments for each foundation, resulting in five total categories. We tokenize the document and check for each token whether it or its lemma is contained in the lexicon. Each time a match occurs, the count for the foundation that the token maps to is incremented by one. Similarly, in cases where a token maps to more than one foundation, all foundations are incremented. Finally, the five counts are normalized by the number of tokens to give foundation frequencies.

B.1.3 Extended moral foundations dictionary (eMFD)

Hopp et al. (2021) aimed to construct a dictionary in a data-driven fashion. The authors pulled 1,010 news articles from the GDELT dataset (Leetaru and Schrodt, 2013) and employed over 854 online annotators to label these articles with moral foundations. In particular, each annotator was assigned an article and a specific foundation. The annotator was asked to highlight the parts of the article that signify that foundation. This yielded a total of 73,001 raw notations.

Instead of mapping a word to some specific foundation, the authors consider a mapping to *all* foundations, each with a specific *weight*. The higher the weight, the more "relevant" the word is to a foundation. Each weight is between 0 and 1 and represents the frequency with which a word is highlighted in the context of that foundation. For example, the score associated with *care* for the word "suffer" is 0.32, indicating that when annotators were assigned to label *care* (among all eligible articles) and saw the word "suffer," they highlighted this word 32% of the time. This is a more nuanced way of associating words with MFs, allowing for flexibility in terms of association strength. Overall, the lexicon contains 3,270 words, each of which contains five scores for the foundations.

To score a document, we follow the procedure detailed in (Hopp et al., 2021). We start with a 5-dimensional vector of zeros for the foundations. After tokenizing a document, for every token that is in the dictionary, we add the 5-dimensional score vector for that token to the document's scores. Finally, we divide the document vector by the number of tokens that matched the lexicon's entries. The result is five scores for the document, each one between 0 and 1 and equal to the average foundation score contributed by all matching tokens.

Table B.2: Keywords used to create their concept vectors for each foundation using the embedding similarity method.

Foundation	Keywords
Authority	authority, obey, respect, tradition, subversion, disobey, disrespect, chaos
Care	kindness, compassion, nurture, empathy, suffer, cruel, hurt, harm
Fairness	equality, egalitarian, justice, nondiscriminatory, prejudice, inequality, discrimination, biased, proportional, merit, deserving, reciprocal, disproportionate, cheating, favoritism, recognition
Loyalty	loyal, solidarity, patriot, fidelity, betray, treason, disloyal, traitor
Sanctity	purity, sanctity, sacred, wholesome, impurity, depravity, degradation, unnatural

B.2 Scoring moral foundations using embedding similarity

Embedding similarity, or distributed dictionary representations (DDR) (Garten et al., 2018), is another method used in prior work to score moral foundations in text. Basically, moral foundations and documents are represented by vectors in high-dimensional Euclidean spaces (embeddings). The foundation score for each pair of document and foundation is defined as the cosine similarity between their vector representations.

We start with a *word embedding*, which maps every word in the dictionary to a dense vector in k dimensions. Examples of widely used embeddings are word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). For a foundation f represented by a set of n_f keywords with their word embeddings $\{w_i \in \mathbb{R}^k : i = 1, \ldots, n_f\}$. The vector representation for f is the average of all word vectors

$$c_f = \frac{1}{n_f} \sum_{i=1}^{n_f} w_i. \tag{B.1}$$

Documents can be encoded similarly. Given a document d as a sequence of n_d tokens $(w_i \in \mathbb{R}^k)_{i=1}^{n_d}$, the vector representation of the document is

$$c_d = \frac{1}{n_d} \sum_{i=1}^{n_d} w_i. \tag{B.2}$$

To score document d with respect to foundation f, we take the cosine similarity between their vector representations:

$$s_{d,f} = \frac{\sum_{j=1}^{k} [c_d]_j [c_f]_j}{\sqrt{\sum_{j=1}^{k} [c_d]_j^2} \sqrt{\sum_{j=1}^{k} [c_f]_j^2}}, \tag{B.3}$$

where $[\cdot]_j$ denotes the jth component of a vector. The score $s_{d,f}$ is in the range $[-1, 1]$, and the higher the score, the more "similar" c_d and c_f are.

In Section 3.4.2.2, we use the GloVe word embedding (Pennington et al., 2014), specifically the "Twitter" version which was trained using a corpus of 27 billion tokens and contains 1.2 million unique vectors in $k = 200$ dimensions.[2] Documents are tokenized similarly to that in

[2]The embedding can be found at https://nlp.stanford.edu/projects/glove.

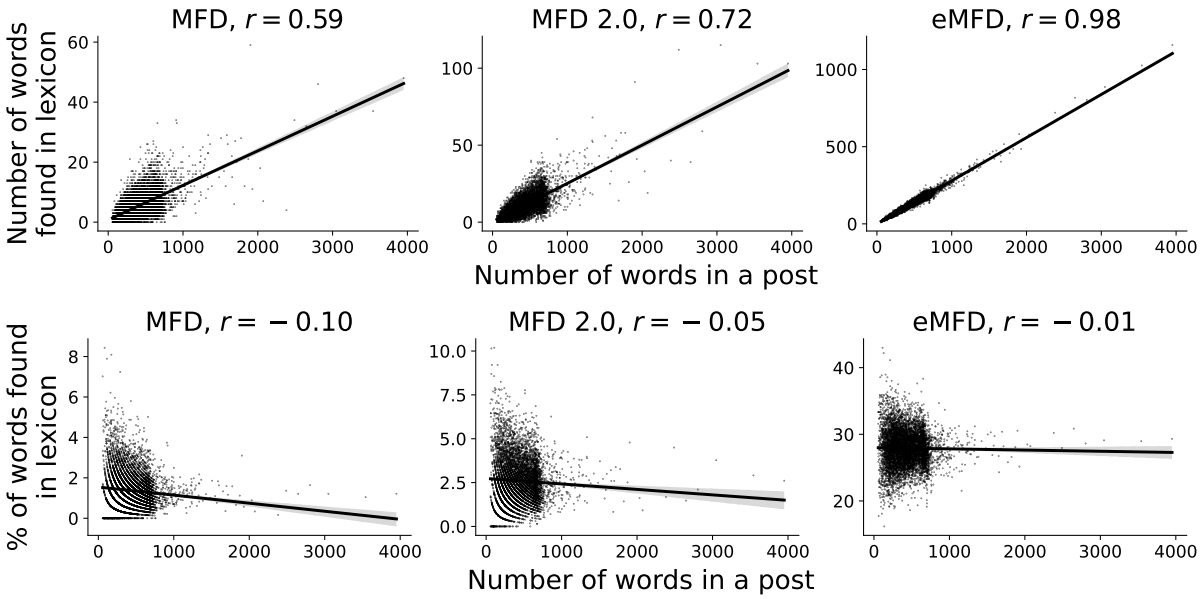


Figure B.1: Correlation between a post's length and the number of words in that post that are found in an MFD lexicon. We use the MFD, MFD 2.0 and eMFD to score 6,800 posts of the topic *family* on `r/AmItheAsshole` (Nguyen et al., 2022, cf. Chapter 2). The y-axis in the top panel represents the number of words within each post that are contained in each lexicon, whereas the y-axis in the bottom panel is the same count normalized by the number of words in that post. Two-sided Pearson correlation coefficients (r, reported on top of each plot) are all statistically significant with $p < 10^{-10}$. Error bars represent 95% CIs on the predictions of a linear regression model.

the word count methods described above and all tokens are lowercased. For every token that does not exist in the embedding, we use the zero vector to represent it. For the sets of words describing a foundation, we use the same prototypical words in (Trager et al., 2022, Appendix B, Table 20), which is also presented in Table B.2. The word lists describing the sub-categories *equality* and *proportionality* are merged into one that describes *fairness*.

B.3 Limitations of Word Count Methods for Scoring Moral Foundations

In this section, we give some examples in the `r/AmItheAsshole` dataset where the prediction of moral foundation using word counts is problematic. Note that the scoring is only done on the body text, not including the title.

In the following example, we score the post using MFD 2.0. The lexicon detects one word (in bold) that signifies the foundation *authority*, which leads to the prediction that this post contains the mentioning of *authority*.

Title: AITA for giving up on my addict sister?

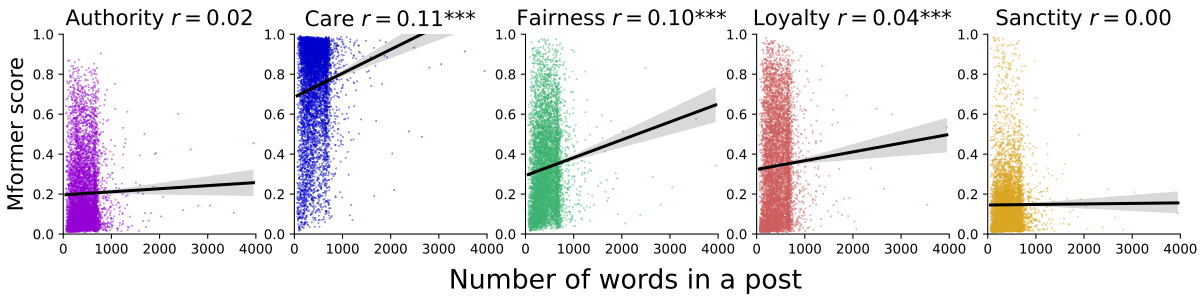


Figure B.2: Correlation between a post's length and its moral foundation scores predicted by Mformer. Two-sided Pearson correlation coefficients are reported on top of each plot, where ***$p < 0.001$. Error bars represent 95% CIs on the predictions of a linear regression model.

> **Body text**: My sister is 2 years older than me and we used to be close but she moved out a few years ago to live with her abusive boyfriend and she got addicted to meth soon after. She kicked out the boyfriend a few months ago and went to rehab and then came to live with us (me and my parents) and about 3 days after she was released from rehab, she went back to her (ex)boyfriend and soon got addicted again. I'm at the point where if she calls me for help or money, I don't answer and I tell her it's her own problem. I know she's going to let it kill her because she **refuses** to let anyone help her despite my parents' multiple attempts. I honestly don't care if she lets it kill her at this point.

B.3.1 Bias on familial roles by word count methods

Here we provide more evidence of social bias which can be revealed from the scoring results by word count methods. Figure 3.1 and Section 3.3 in the main text show that when MFD 2.0 is used to score content in the topic *family*, there is a clear association of familial roles to moral foundations, such as "father" to *authority* and "mother" to *care*.

We find that this distinction directly results in a difference in the distributions of foundation scores for these posts. In Figure B.3, we score the same 6,800 posts using MFD, MFD 2.0, eMFD and Mformer and compare the mean score for three types of posts: those containing the word "father," those containing the word "mother" and those containing neither. As seen in bar plots on the second row for MFD 2.0, posts containing "father" on average score dominantly higher for *authority* than those containing "mother." This pattern is the opposite for *care*, where posts containing "mother" score much higher than the those containing "father."

We do not find any significant difference in the mean foundation scores between these groups when using Mformer, an evidence that our model is unlikely to suffer from the same problem. This does not conclusively show that Mformer successfully avoids all biases, but at least in this context it does not seem to associate particular words or roles with moral foundations directly.

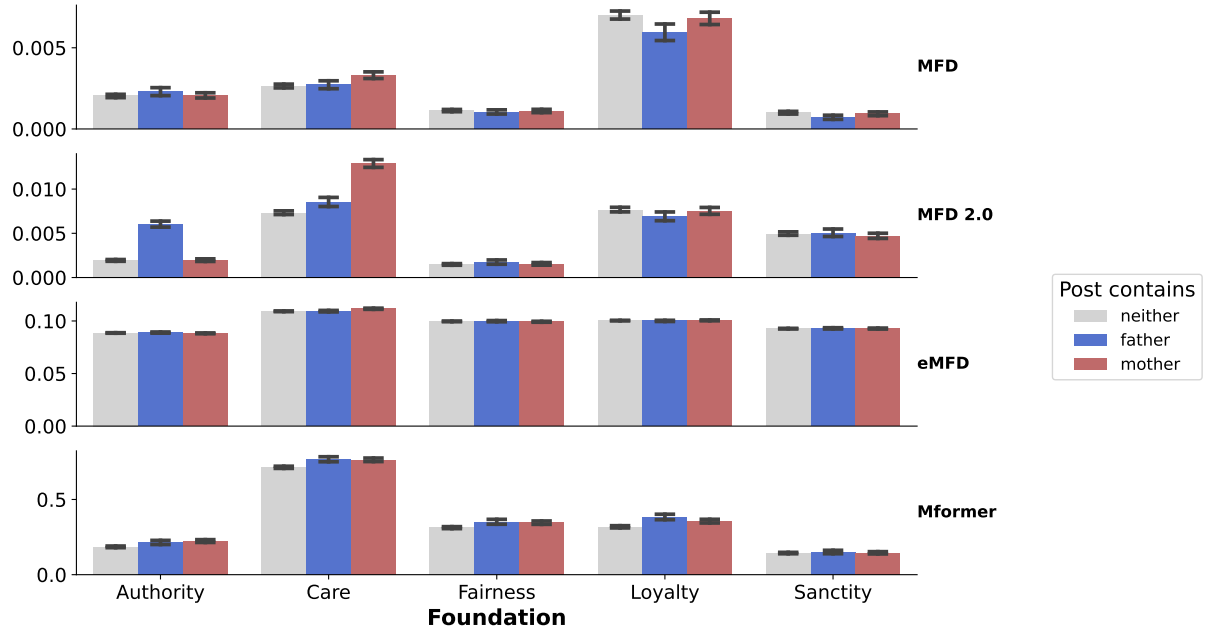


Figure B.3: Mean foundation scores (and 95% CI) for 6,800 posts of the topic *family* on `r/AmItheAsshole` (Nguyen et al., 2022, cf. Chapter 2). The bar colors represent posts that contain the word "father" (blue, $N = 1,472$), those that contain the word "mother" (red, $N = 2,180$) and those that contain neither of these words (grey, $N = 3,960$). Each post is scored by word count (including with MFD, MFD 2.0 and eMFD) and Mformer.

B.4 The Moral Foundations Dataset

In this section, we describe in more detail the collection and processing of the moral foundations dataset in Section 3.4.1. Since this dataset is used to train moral foundation classifiers, it must contain examples as text and five binary labels, one for each foundation. The goal is to build binary classifiers to predict whether each foundation is present in an input.

B.4.1 Moral foundations Twiter corpus

This dataset was introduced by Hoover et al. (2020) and released publicly.[3] It contains a total of 34,987 tweets[4] encompassing seven "socially relevant discourse topics": All Lives Matter, Black Lives Matter, 2016 U.S. Presidential election, hate speech, Hurricane Sandy, and #MeToo Hoover et al. (2020). Thirteen annotators were carefully trained to label the tweets with moral foundations and their sentiments (virtue and vice), with at least three annotations per tweet. Annotators were allowed to choose more than one foundation that match a tweet, or no foundation at all (in this case, the authors defined the label as "non-moral").

The format of this dataset is a tweet plus all labels given by each annotator. For example,

[3] https://osf.io/k5n7y

[4] Hoover et al. (2020) reported 35,108 tweets in their dataset, but the released corpus contains only 34,987 unique tweets.

the tweet

> "There is race war being engineered and financed right in front your eyes in America. Wake up! #alllivesmatter"

was labeled by three annotators, who gave it the following sets of foundations: {*harm*}, {*care*, *fairness*} and {*harm*}. We first merge the moral sentiments (virtue and vice) for all labels, so that *care* and *harm* both become *care*. Then for each tweet, we determine that it contains a foundation f if at least one annotator labeled it with f. So, the labels for the above examples are *care* and *fairness* as at least one annotator gave it one of these foundation labels.

We keep all tweets in this dataset for our use. Finally, we perform train-test splitting for each foundation. Specifically, for foundation f we randomly choose 10% of the dataset as the test set, ensuring that the training and test sets are stratified.

B.4.2 Moral foundations news corpus

This dataset was collected and used to create the eMFD lexicon (Hopp et al., 2021).[5] As described in Appendix B.1.2 above, the dataset contains 1,010 unique news articles and 73,001 raw highlights produced by 854 annotators. Each highlight represents the annotation that the corresponding part of the article contains a specific foundation.

We follow the preprocessing by Hopp et al. (2021) for this raw data. We only keep annotators who spent at least 45 minutes on their task and exclude one annotator who spent too long. Only 63,958 annotations from 557 annotators remain. Then, only documents that had been labeled by at least two annotators who differed in their assigned foundation are kept, yielding 47,650 highlights for 992 articles.

From the remaining articles and highlights, we create a moral foundation-labeled dataset by segmenting the articles into sentences. Specifically, we use `spaCy` to extract all sentences within each article. Then we label a sentence s with a foundation f if s was highlighted (in part or in whole) with f by at least one annotator. For example, if a highlight associated with foundation *sanctity* contains a sentence—either entirely or partially—then that sentence will have the binary label 1 for *sanctity*. This yields 32,262 sentences in total from 992 articles.

Similarly, we perform stratified train-test splitting for each moral foundation. Due to the annotation setting, not every article was labeled with all foundations. Therefore, sentences for which no annotator was tasked with labeling using a particular foundation are excluded from the training and test sets for that foundation.

B.4.3 Moral foundations Reddit corpus

Introduced by Trager et al. (2022), this dataset contains a total of 17,886 comments[6] extracted from 12 different subreddits roughly organized into three topics: U.S. politics, French politics, and everyday moral life (Trager et al., 2022).[7] A total of 27 trained annotators were tasked with labeling these comments with moral foundations. The authors followed a recent work by Atari

[5] https://osf.io/vw85e

[6] The number of comments reported in (Trager et al., 2022) is 16,123.

[7] https://huggingface.co/datasets/USC-MOLA-Lab/MFRC

et al. (In Press) which proposes to separate the foundation *fairness* into two classes: *equality* (concerns about equal outcome for all individuals and groups) and *proportionality* (concerns about getting rewarded in proportion to one's merit). In addition, another label, *thin morality*, was defined for cases in which moral concern is involved but no clear moral foundation is in place. Therefore, the total number of binary labels is seven.

For the labels, to be consistent with the other two sources, we merge both *equality* and *proportionality* into their common class *fairness* and consider *thin morality* as the binary class 0 for all foundations. This results in the same five moral foundation labels. Then, a comment receives a binary label 1 for a foundation if at least one annotator labeled this comment with this foundation.

We similarly perform train-test splitting for each moral foundation.

B.4.4 Further discussions on moral foundations datasets

The moral foundations dataset described in Section 3.4.1 and this appendix section is aggregated from three different sources. Here, we provide more details comparing these datasets in terms of the definition of each moral foundation and agreement rate.

Table B.3 presents the definition of each moral foundation used in the annotation process for each dataset, Twitter, News and Reddit. We find that the way each foundation is described is consistent among the three sources. In all settings annotators were informed about moral foundations both in actions that uphold them (virtues) and those that violate them (vices). We also find consistency in their content. For example, all definitions of the foundation *loyalty* are related to one's priority towards one's ingroup where patriotism is a common example. The major difference among these datasets is with regard to how detailed the examples for each foundation are. News, for instance, includes several examples of each foundation, such as "seeking societal harmony while risking personal well-being" or "verbally, physically, or symbolically attacking outgroup members because they are outgroup members" for *loyalty* (Hopp et al., 2021, see Section 4, Subsection 7 in the Supplemental Materials). In addition, as described in Appendix B.4.3 above, in the Reddit dataset the foundation *fairness* is split into two concepts, *equality* and *proportionality*, both of which are presented in Table B.3. We find that these definitions only represent more specific cases of *fairness* and do not contradict this concept in any particular way. Overall, based on the consistency with which a moral foundation is portrayed throughout the three domains, we think that aggregating them does not pose any challenge from a definitional perspective.

With respect to inter-annotator agreement rate, we note that Hoover et al. (2020) and Trager et al. (2022) report these metrics for this in the Twitter and Reddit datasets, respectively. (We do not have this rate for the News dataset because example sentences are extracted from original news articles.) In particular, both works report Fleiss's Kappa (Fleiss, 1971) and its prevalence- and bias-adjusted variant PABAK (Sim and Wright, 2005) for each foundation. Both works also report relatively medium PABAK values, suggesting the subjective nature of this annotation task. In using these datasets for training supervised classifiers, to account for label noise several methods have been proposed, such as by taking the majority vote. For these datasets, we decide to assign a foundation f to an example if at least one annotator gave it this label. The reason for this choice is because of the inherent subjectivity of this task and

Table B.3: Summary of the definitions of moral foundations used to train annotators of three datasets, Twitter (Hoover et al., 2020), News (Hopp et al., 2021) and Reddit (Trager et al., 2022). These datasets are described in more detail in Appendices B.4.1 to B.4.3. For News, the full definitions and examples can be found in (Hopp et al., 2021, Supplemental Materials). For Reddit, the foundation *fairness* was split into two classes, *equality* and *proportionality*; we report the definitions for both here.

Foundation	Twitter (Hoover et al., 2020, see "Annotation")	News (Hopp et al., 2021, see Section 4 of "Supplemental Materials")	Reddit (Trager et al., 2022, see Section 3.1)
Authority	Prescriptive concerns related to submitting to authority and tradition and prohibitive concerns related to not subverting authority or tradition.	(...) Authority recognizes that leaders and followers represent a mutual relationship that fosters group success. Subversion is a consequence of leaders or followers disrupting the healthy relationship of social hierarchies. (...)	Intuitions about deference toward legitimate authorities and high-status individuals. It underlies virtues of leadership and respect for tradition, and vices of disorderliness and resenting hierarchy.
Care	Prescriptive concerns related to caring for others and prohibitive concerns related to not harming others.	Care is often exemplified through nurturing, assisting, and protecting others. (...) Harm can be considered any time of distress or pain inflicted on another. (...)	Intuitions about avoiding emotional and physical damage or harm to another individual. It underlies virtues of kindness, and nurturing, and vices of meanness, violence and prejudice.
Fairness	Prescriptive concerns related to fairness and equality and prohibitive concerns related to not cheating or exploiting others.	(...) Fairness is when the give and take between two parties is equivalent. Reciprocity, unity, collaboration, cooperation, solidarity, and proportionality are key concepts of fairness. (...) Cheating is when one party violates the norms of justice by reaping the benefits without contribution or consideration of the other party. Selfishness, freeloader, cheater, slacker, liar, and manipulator describe the actions of one party against another. (...)	(Proportionality) Intuitions about individuals getting rewarded in proportion to their merit (i.e., effort, talent, or input). It underlies virtues of meritocracy, productiveness, and deservingness, and vices of corruption and nepotism. (Equality) Intuitions about egalitarian treatment and equal outcome for all individuals and groups. It underlies virtues of social justice and equality, and vices of discrimination and prejudice.
Loyalty	Prescriptive concerns related to prioritizing one's ingroup and prohibitive concerns related to not betraying or abandoning one's ingroup.	(...) Loyalty includes showing group pride, patriotism, and a willingness to sacrifice for the group. (...) Betrayal is demonstrated when group members put themselves prior to the group or damage the group image and identity. (...)	Intuitions about cooperating with in-groups and competing with out-groups. It underlies virtues of patriotism and self-sacrifice for the group, and vices of abandonment, cheating, and treason.
Sanctity	Prescriptive concerns related to maintaining the purity of sacred entities, such as the body or a relic, and prohibitive concerns focused on the contamination of such entities.	(...) Sanctity relates to things that are noble, pure and elevated, such as the human body which should not be harmed. (...) Degradation deals with ideas of physical disgust, particularly when it has to do with the body. (...)	Intuitions about avoiding bodily and spiritual contamination and degradation. It underlies virtues of grossness, impurity, and sinfulness.

the reliance on "moral intuitions rather than [...] deliberations" (Hopp et al., 2021), which we believe make it more likely for annotators to "miss" a true label than mislabeling a negative example (higher false negative rate than false positive rate). Further, taking the majority vote will decrease the size of the training set significantly; a favor for data quantity at the expense of higher label noise can be justified

B.5 Training Moral Foundation Classifiers

In this section, we provide more detail on training/fine-tuning moral foundation classifiers described in Section 3.4.2.2. Specifically, we use two models, logistic regression and RoBERTa for sequence classification (Liu et al., 2019), and describe the hyperparameter tuning setting for each model.

B.5.1 Logistic regression

For each foundation f, we train an ℓ_2-regularized logistic regression model to estimate $p(f \mid d)$, the probability that a document d is labeled with f. We choose to tune C, the hyperparameter equal to the inverse of the ℓ_2 regularization strength, in $\{10^{-7}, 10^{-6}, \ldots, 10^{6}, 10^{7}\}$. The best value of C is one that achieves the highest average validation AUC over 10 folds in the training set. We use four document embeddings for this model:

- Tf-idf (Manning et al., 2008): We tokenize every training example using `spaCy`. Then we lemmatize and lowercase every token. We filter out all tokens that are in the set of standard English stop words, then remove all tokens with non-alphabetic characters and those with fewer than 3 characters. The vocabulary discovered from the training set contains 12,586 unique tokens, corresponding to the dimensionality of the embedding.

- SpaCy (Honnibal et al., 2020): We also use the built-in static word embedings by `spaCy`. Specifically, after tokenizing a document, each token within the document is associated with a word vector with 200 dimensions. The vector representation for the document is the average of these word embeddings.

- GloVe (Pennington et al., 2014): Similar to the `spaCy` embedding, the GloVe embedding of a document is the average of all of its token embeddings. We follow the same setting described in Appendix B.2 above, where we use the "Twitter" 200-dimensional version of GloVe and use the zero vector to represent all out-of-vocabulary tokens.

- Sentence-RoBERTa (Reimers and Gurevych, 2019): This is RoBERTa (Liu et al., 2019) fine-tuned for a sentence similarity task. To encode a document, we perform a forward pass through this network. The output provided by the final layer (before fitting it through a classification head) contains embeddings for all of the document's tokens. The vector representation for the document is taken as the average of all token vectors. We specifically choose the RoBERTa-large architecture, which outputs a 1,024-dimensional embedding.

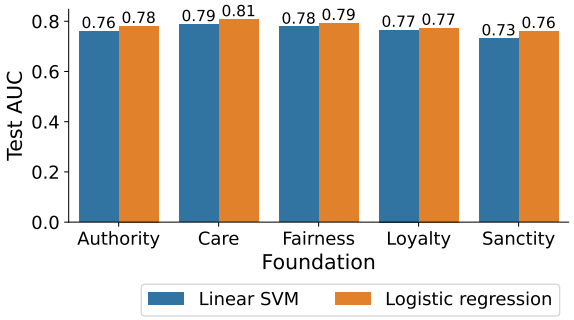


Figure B.4: Performance comparison between linear SVM (trained using the tf-idf embedding) and logistic regression (trained using the Sentence-RoBERTa embedding).

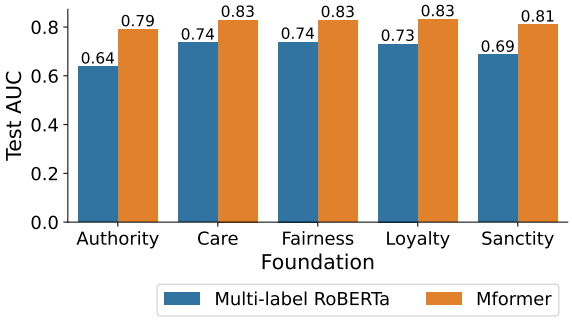


Figure B.5: Performance comparison between two RoBERTa variants: multi-label and Mformer. For the multi-label variant, RoBERTa's final classification layer contains 5 neurons, each followed by a *sigmoid* activation to represent the binary probability for each class. For Mformer, each foundation is associated with one version of RoBERTa binary classifier.

Alternatives to logistic regression. In addition to logistic regression, we also consider support vector machines (SVM), another type of linear classifier that can achieve robustness by finding a "maximum-margin" decision function. This method was used by Hoover et al. (2020) in setting up a prediction baseline for the Twitter dataset.

In particular, we embed each training example using tf-idf similar to above. Different from Hoover et al. (2020, see "Methodology"), we do not consider only words belonging to MFD or MFD 2.0 but consider the entire vocabulary (12,586-dimensional), which already contains words in these lexicons, for our embedding, leading to a more versatile representation. Then, for each moral foundation, we train a linear SVM with ℓ_2 regularization using the binary presence of that foundation as the supervised signal. Following Hoover et al. (2020), the regularization hyperparameter is set to $C = 1.0$.

Using the same test set, we compare linear SVM and the highest-performing logistic regression models (using the Sentence-RoBERTa embedding) in Figure B.4. The results show that logistic regression achieves a higher test AUC than SVM for all foundations, although the difference can be marginal, especially for *loyalty*. We therefore only report the results by logistic regression in the main text for further evaluation (cf. Figures 3.2 and 3.3).

B.5.2 RoBERTa

The moral foundation classifiers described and used in this work are RoBERTa (Liu et al., 2019). This language model is a direct extension of BERT (Devlin et al., 2019) with a more advanced and robust training procedure. RoBERTa has two versions: RoBERTa-base, which has a hidden size of 768, and RoBERTa-large with a hidden size of 1,024. We choose the former architecture for this task.

For each moral foundation, we fine-tune RoBERTa on a binary text classification task:

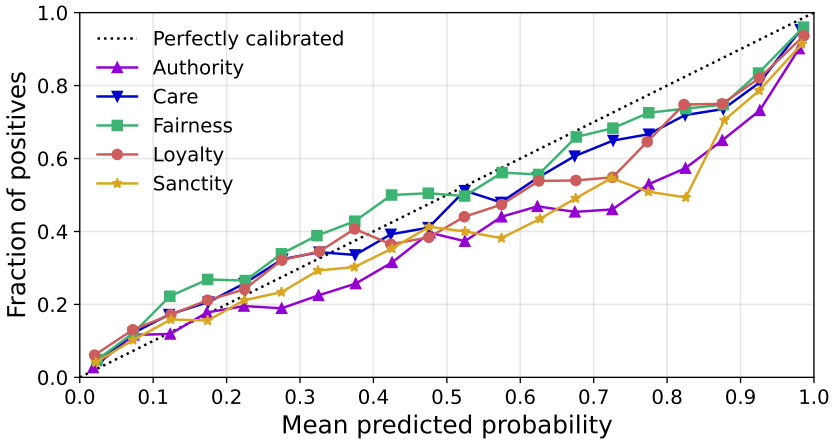


Figure B.6: Calibration curves for the five Mformer classifiers.

whether the foundation is present in the input or not. We first tokenize a document using HuggingFace's built-in tokenizer. Then, the sequence of tokens is truncated to a maximum of 510 tokens if it is longer than that. Two special tokens, `<s>` and `</s>` are added to the beginning of the end of the sequence, respectively, ensuring the maximum sequence length is 512. These are often called the `CLS` ("classification") and `SEP` ("separation") tokens.

In fine-tuning RoBERTa, we start with the weight checkpoints released on HuggingFace.[8] After the final self-attention layer, the 768-dimensional embedding of the token `<s>` is fed through a linear layer with 768 neurons, followed by the tanh activation. Finally, this goes through another linear layer with two outputs, followed by the softmax transformation to estimate the probability of the two classes. We use the typical categorical cross-entropy loss to compare the probabilities with the ground-truth binary label of the input. All RoBERTa weights are optimized using the AdamW optimizer (Loshchilov and Hutter, 2019), where we set its parameters to $\epsilon = 10^{-8}$, $\beta_1 = 0.9$, $\beta_2 = 0.99$. We also add an ℓ_2 regularization term for the weights with the regularization strength $\lambda = 0.01$. We use a batch size of 16 during optimization.

With respect to hyperparameter tuning, we control two quantities: the learning rate $\alpha \in \{10^{-6}, 10^{-5}, 3 \times 10^{-5}\}$ and the number of training epochs in $\{2, 3, 4, 5, 6\}$. To do so, we randomly split the training set into a training and a validation portion of relative size 9:1. The splitting is done in a stratified manner, ensuring the same proportion of the positive class in the two subsets. We then perform a grid search over all combinations of α and the number of epochs and choose the combination with the highest validation AUC as the final hyperparameters. Finally, we combine the training and validation subsets into the original training set and fine-tune RoBERTa using the best hyperparameters to give the final models.

We call the final five fine-tuned classifiers Mformer.

[8] https://huggingface.co/roberta-base

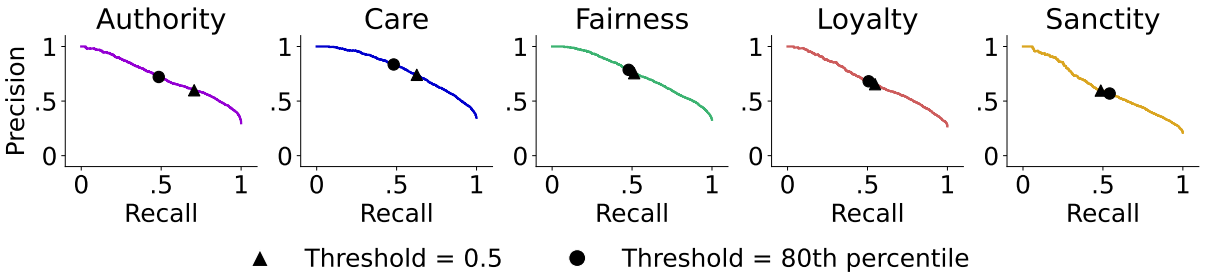


Figure B.7: Precision-recall curves for the five Mformer moral foundation classifiers. Two thresholding values for the prediction scores are displayed: 0.5 (black triangles) and the 80th percentile of all predicted scores on the test set (black circles).

B.5.3 Multi-label RoBERTa

We also consider a version of RoBERTa which performs multi-label classification. In particular, different from Mformer which has five distinct models, this version only has one RoBERTa model but the final classification layer contains 5 neurons instead of 2. Each of the five neurons is followed by a sigmoid activation, which outputs a probability that the input contains each moral foundation. Note that the five outputs by multi-label RoBERTa do not necessarily add up to 1, because each output represents the probability for each foundation. Compared to Mformer, multi-label RoBERTa only requires one copy of RoBERTa, thereby reducing the space requirement by 5 times.

Note that our training dataset is multi-label, and that there exist examples for which some labels for moral foundations are missing. For these reasons, train-test splitting must be done differently. We employ the iterative stratification method by Sechidis et al. (2011); Szymański and Kajdanowicz (2017), setting 10% of the dataset for testing and splitting the training set into a training and validation portion of ratio 9:1. In fine-tuning multi-label RoBERTa, we explore the same hyperparameters as above and choose the model with the highest validation AUC.

We find that the multi-label variant to Mformer leads to suboptimal performance. As shown in Figure B.5, which compares these two alternatives using the AUC on the same test set, Mformer performs significantly better than multi-label RoBERTa, obtaining an increase of 12.0–24.0% in AUC. We therefore decide to use Mformer instead of multi-label RoBERTa in later analyses.

B.6 Evaluation of the Fine-Tuned RoBERTa Classifiers

In this section we provide more detail on evaluation of Mformer models in Section 3.4.3.

B.6.1 Calibration

To assess whether the scores predicted by Mformer closely approximate actual probabilities, we present a calibration plot in Figure B.6. For each moral foundation, we collect Mformer

Table B.4: Precision at different thresholding levels.

Foundation	95th	90th	80th	70th	60th	50th
Authority	0.94	0.86	0.72	0.64	0.57	0.51
Care	0.98	0.95	0.84	0.74	0.65	0.58
Fairness	0.98	0.92	0.79	0.69	0.61	0.54
Loyalty	0.92	0.84	0.68	0.59	0.52	0.45
Sanctity	0.88	0.72	0.57	0.48	0.41	0.36

Table B.5: Recall at different thresholding levels.

Foundation	95th	90th	80th	70th	60th	50th
Authority	0.16	0.29	0.48	0.64	0.77	0.85
Care	0.14	0.27	0.48	0.64	0.74	0.83
Fairness	0.15	0.28	0.48	0.63	0.74	0.82
Loyalty	0.17	0.31	0.51	0.66	0.77	0.85
Sanctity	0.21	0.34	0.54	0.68	0.79	0.86

Table B.6: F-1 score at different thresholding levels.

Foundation	95th	90th	80th	70th	60th	50th
Authority	0.27	0.43	0.58	0.64	0.66	0.64
Care	0.25	0.42	0.61	0.68	0.69	0.68
Fairness	0.26	0.43	0.60	0.66	0.67	0.65
Loyalty	0.29	0.46	0.58	0.62	0.62	0.59
Sanctity	0.34	0.47	0.56	0.56	0.54	0.51

Table B.7: Accuracy at different thresholding levels.

Foundation	95th	90th	80th	70th	60th	50th
Authority	0.75	0.77	0.79	0.78	0.76	0.71
Care	0.70	0.74	0.79	0.79	0.77	0.73
Fairness	0.72	0.76	0.79	0.78	0.76	0.71
Loyalty	0.77	0.80	0.80	0.78	0.74	0.69
Sanctity	0.83	0.83	0.82	0.78	0.72	0.65

scores for all test examples. Then, the scores are discretized into 20 equal-width bins. The x-axis represents the average of the scores within each bin and the y-axis represents the fraction of examples in each bin that are in fact positive (i.e., contain the foundation). A perfectly calibrated classifier should have a diagonal calibration curve, as, for instance, a score of 0.7 implies that 70% of examples predicted at that level are positive.

It is known that classifiers with outputs produced by sigmoid or softmax functions are well calibrated, which is the case in Figure B.6 since most curves are close to the diagonal line. We find that the calibration curves for *care* and *fairness* are the best, while those for *authority* and *sanctity* tend to deviate from the perfectly calibrated curve as predicted scores get higher.

B.6.2 Choosing classification thresholds

In this work far we have shown that Mformer (fine-tuned RoBERTa) outperforms existing methods—namely, word count, embedding similarity, and logistic regression—on predicting moral foundations in text based on the AUC metric. In most downstream analyses, binary labels are required from the raw prediction scores. Determining a threshold above which a prediction becomes positive reflects the tradeoff between precision and recall, which we show in Figure B.7 for each moral foundation.

The good calibration curves (Figure B.6) described above suggest that a threshold of 0.5 is reasonable. In Figure B.7, the precision and recall for each foundation is represented as black triangles. Another widely-chosen method, especially when predicting in a novel domain, is to set the top xth percentile of all scores as the classification threshold, where $x \in [0, 1]$. A higher value of x is typically justified by the researchers' preference for higher precision at the cost of low recall. To illustrate the precision-recall tradeoff, we present four metrics—precision, recall, F-1 and accuracy—for different levels of x in Tables B.4 to B.7. We find that when $x = 80$ or $x = 70$, both F-1 and accuracy remain relative high. Therefore, we choose either of these values when binary labels are needed in Section 3.6.

B.7 External Moral Foundations Datasets and Evaluation

This section provides more details on the four external datasets used to examine the generalizability of our model, Mformer, described in Section 3.5.

B.7.1 Datasets

B.7.1.1 Moral foundations vignettes dataset (VIG)

This dataset contains 132 vignettes describing behaviors that violate a moral foundation (Clifford et al., 2015). Each vignette is short (14–17 words) and in the "You see [behavior]" format. An example that violates the foundation *authority* is "You see a woman refusing to stand when the judge walks into the courtroom." In addition to the five moral foundations, two classes are considered: *liberty*, relating to freedom of choice, and *social norms*, describing actions that are unusual but are not considered morally wrong. In total, each class contains 16 to 17 examples except for the foundation *care*, which contains 32 examples where half are about actions causing physical harm to humans and the other half to nonhuman animals.

We consider this dataset to contain "gold labels" for three reasons. First, the vignettes were carefully generated by the authors to clearly violate only one foundation. Second, each vignette was independently rated by a large pool of annotators (approximately 30 respondents per example) and in every case, at least 60% of annotators agreed on the intended foundation. Third, all vignettes which were believed by 20% or more of annotators to violate an unintended foundation were excluded. In other words, the dataset only contains examples with high-confidence moral foundation labels.

We preprocess this dataset as follows. First, we merge the two sub-cases for *care*, namely physical harm to humans and to nonhuman animals, into simply *care*. Then, we remove the 17 instances violating *liberty*. Finally, for the 16 instances of *social norms*, we consider them to contain no foundation at all, i.e., the five binary labels for moral foundations are all 0. We decide against considering *liberty* examples to contain no foundation, as it could be shown to correlate with *fairness* and *authority* (Haidt, 2013) which might unnecessarily complicate our evaluation later on. On the other hand, *social norms* cases were clearly designed not to violate any explicit moral foundation at all, hence they are good candidates for negative instances. The final dataset contains 115 vignettes with 16–17 examples for each foundation except for *care*, which has 32 examples.

B.7.1.2 Moral arguments dataset (ARG)

Kobbe et al. (2020) aimed to study the effect of moral sentiment in the analysis of arguments, especially in the context of argument mining. To do so, the authors pulled 320 arguments taken from the online debate platforms `createdebate.com` and `convinceme.net`; the dataset is called Dagstuhl ArgQuality Corpus (Wachsmuth et al., 2017; Habernal and Gurevych, 2016). Different from the vignettes above, the arguments are much more diverse in length (min = 11, max = 148, median = 64 tokens per example) and in topic (e.g., evolution vs. creation, plastic water bottle ban, etc.). Then the two authors manually labeled these arguments with every moral foundation.

Since this dataset already contains the necessary binary labels for moral foundations, we thus perform no preprocessing and use all 320 instances. Of all arguments, nearly a third (96/320) were labeled with no foundation, over half (179/320) contain one foundation, and the maximum number of foundations per example is 3 (with 3/320 examples). An example argument containing *fairness* is "Religion in the past has caused many wars. It encourages racism, sexism, and homophobia. It is something that gives us prejudice. It makes us hate one another. [T]he time has come to put a stop to it."

B.7.1.3 Social chemistry 101 dataset (SC)

This dataset contains 292K rules-of-thumb (RoTs), defined as "descriptive cultural norms structured as the judgment of an action" (Forbes et al., 2020). Each RoT was created by an online annotation worker based on a given *situation*. For example, the situation "Not wanting to be around my girlfriend when she's sick" prompted an annotator to give the RoT "It's kind to sacrifice your well-being to take care of a sick person." Then, the RoTs were labeled with categorical attributes such as social judgments, cultural pressure, legality, and moral foundations. Similar to the previous two datasets, in this case, every RoT received five binary labels for the foundations.

We choose the test set, which contains 29,239 RoTs (10% of the entire dataset) to evaluate the moral foundation classifiers. In this set, the RoTs are relatively short (median = 11 tokens per example). Over 60% of the RoTs contain only one foundation, while 21% contain no foundation at all. An example RoT labeled with *authority* is "It is controlling to tell a partner what they can and cannot do with their body."

B.7.1.4 Moral integrity corpus (MIC)

Similar to SC, this dataset features 99K RoTs, which were created based on *prompts* such as "Do you ever smoke marijuana illegally?" (Ziems et al., 2022). However, unlike SC, here the RoTs were generated by humans based on *prompt-reply* pairs partially created by chatbot systems. In particular, with the prompt-reply pair "Do you ever smoke marijuana illegally?"–"I smoke it to relax. It helps my anxiety," an annotator generated the RoT "It is bad to use harmful substances." Subsequently, each RoT was labeled on several dimensions including the severity of violating that RoT, the degree to which other people agree with the RoT, and the moral foundations involved in the RoT.

We also choose the test set, containing 11,375 examples. The examples are also short (median = 9 tokens per instance) and 67% contain one foundation. An example containing *loyalty* is "It is wrong for a person to betray their friends."

B.7.2 Scoring moral foundations

In all four datasets described above (VIG, ARG, SC, and MIC), we score every instance using five methods described in Section 3.2.2. The summary of these methods is as follows:

- Word count: We use the three dictionaries—MFD, MFD 2.0, and eMFD—for scoring every example. Exact detail about how examples are scored can be found in Appendix B.1.

Even though MFD 2.0 tends to perform better than MFD in most cases, we also want to examine MFD, especially regarding its limited vocabulary.

- Embedding similarity: Appendix B.2 describes how scoring is done. We use the same word embedding and seed words.
- Supervised classifiers: We use the five Mformer models described in Appendix B.5. No further fine-tuning is done on any external datasets. We decide not to use any logistic regression models because Mformer is shown to outperform them in all previous benchmarks on the moral foundations test set.

B.7.3 Evaluation

Once we obtain the moral foundation scores, we also use the AUC to compare them against the ground-truth binary labels. The result is presented in Figure 3.3 in the main text. Some analysis of the AUC for Mformer has been discussed in Section 3.5 in the main text.

We find that Mformer classifiers give the best performance in all cases, often by a large margin compared to the second-highest scorer (e.g., 0.95 vs. 0.69 obtained by MFD 2.0). The only tie in the highest AUC is between Mformer and embedding similarity on predicting the foundation *loyalty* on the VIG dataset (both AUC = 0.75).

We also find that, despite being based on dense word embeddings and not restricted to a small lexicon, embedding similarity is not clearly better than simple word count methods based on the MFDs. Sometimes, its performance is even worse: for example, when predicting *fairness* on the ARG dataset, the AUC for embedding similarity is 0.53—somewhat equal to random guessing—compared to 0.60 by MFD.

When comparing the three lexicons for word count, we find that MFD 2.0 often performs better than MFD. This is expected since MFD 2.0 is an extension of MFD with more than three times as many words. The eMFD also tends to improve from MFD 2.0, with some exceptions such as when predicting *loyalty* in ARG. However, this improvement is much smaller than what Mformer makes.

In the following, we provide further evaluation of Mformer models for each dataset.

B.7.3.1 VIG

We find that Mformer performs very well on this dataset. Given that the vignettes were carefully designed to elicit one particular moral foundation, and the fact the majority of a large number of annotators agreed with the ground-truth labels, the results are in strong favor of our models. We find that Mformer performs well when predicting the foundations *authority*, *fairness*, and *sanctity* in particular, with the recorded AUC between 0.88 and 0.95. However, the performance on *care* is not as impressive (AUC = 0.81). Upon inspection, we find that some examples for *care* tend to be misclassified as *fairness*; an example is "You see a man quickly canceling a blind date as soon as he sees the woman." (As Clifford et al. (2015) reported, for this this example 16% of respondents also thought it is about *fairness*.) The foundation hardest to score is *loyalty*, with an AUC of 0.75. Similarly, we find that some

examples tend to be misclassified as *authority* as in this vignette: "You see a head cheerleader booing her high school's team during a homecoming game."

We consider this good evidence in support of adopting Mformer as a new classifier of moral foundations. However, limitations exist, including the fact that the dataset is small; the vignettes are relatively short and simple; the scenarios are only about actions that exhibit a *violation* moral foundation and not those that constitute a *virtue*.

B.7.3.2 ARG

Similar to VIG, we think that this dataset contains high-quality labels since Kobbe et al. (2020) directly studied moral foundations in arguments and the annotation process was executed carefully, although with fewer annotators. The results for Mformer on this dataset is also very good with all AUC between 0.81 and 0.86. The foundations *care* (AUC = 0.86), *fairness* (AUC = 0.86) and *loyalty* (AUC = 0.86) are the easiest to classify. For *sanctity* (AUC = 0.83), we find that low-scoring examples often score high on *care*. The following example was only labeled with *sanctity* but is scored low on this foundation while very high on *care*:

> Americans spend billions on bottled water every year. Banning their sale would greatly hurt an already struggling economy. In addition to the actual sale of water bottles, the plastics that they are made out of, and the advertising on both the bottles and packaging are also big business. In addition to this, compostable waters bottle are also coming onto the market, these can be used instead of plastics to eliminate that detriment. Moreover, bottled water not only has a cleaner safety record than municipal water, but it easier to trace when a potential health risk does occur. [URL]

Similarly, we also find some examples for *authority* score low on this foundation but high on *care*, as in the argument: "Some kids don't learn by spanking them..So why waste your time on that, when you can always take something valuable away from them."

B.7.3.3 SC

This is the largest dataset for evaluation with 29K examples (RoTs). While Mformer still performs better than all other methods, we find the AUC on SC, between 0.70 and 0.80, to be relatively low compared to that on VIG and ARG. Of all five foundations, *loyalty* (AUC = 0.80) is the easiest to score, followed by *authority* (AUC = 0.74) and *fairness* (AUC = 0.73).

We suspect that the relatively lower performance of Mformer is due to the fact that there exists some considerable label noise in this dataset. First, we note that moral foundations are one of many categorical attributes studied by Forbes et al. (2020), beside cultural pressure and social judgment for example. Since annotators were not exclusively tasked with identifying foundations in an RoT, it is possible that the foundation labels were noisily chosen. Second, while the examples in VIG and ARG were annotated by more than one annotator and subsequently went through manual inspection, this is not the case for SC's RoTs. Each RoT was only labeled by one worker in SC, and since this annotation task is known to be subjective and has low inter-annotator agreement (Hoover et al., 2020; Trager et al., 2022; Kobbe et al., 2020), we believe the noise level is significant and deserves a formal inspection. Third, the annotators for VIG and ARG went through careful training while the construction of SC tended

to simplify this process: annotators were only asked "Which moral foundations best fit this RoT?" and given short descriptions and examples of each foundation. This could have had a significant effect on the labels.

Here we give some examples of label noise in SC. An RoT which was only labeled with the foundation *care* but scores very high on *fairness* reads: "You shouldn't call people homophobic slurs." While this RoT also scores high on *care*, it lacks a true label for *fairness*, and the false positive prediction for *fairness* should be true positive if the label was correct. An example of false negative prediction in the presence of an incorrect label is this RoT: "It's fun to go swimming on a hot day." This example was labeled with foundations *authority* and *care* but it scores very low on all foundations according to Mformer.

B.7.3.4 MIC

The results for MIC are similar to SC. In particular, we find a relatively low AUC for Mformer, even though it performs better than all other methods. The AUC range is 0.65–0.75, with the highest performance achieved when scoring the foundations *authority* and *loyalty*. We also believe that the MIC dataset contains significant label noise due to the same reasons as SC. A false positive example under label inaccuracy is "It's interesting to learn about the shows you watch," which was labeled by a worker to contain the foundation *sanctity* but scores very low using Mformer. A false negative example is "People should unite and fight against terrorism," which scores high on *loyalty* and should contain *loyalty* but does not according to the annotators.

B.8 Analyzing Moral Discussions on Reddit using Moral Foundations Theory

In this section, we describe in more detail our study of moral stories and judgments on `r/AmItheAsshole` (AITA), a popular subreddit. Users on AITA post about a non-violent moral conflict they recently experience and invite the community to judge if they were in the right. This online community has garnered the attention of social NLP and computational social science researchers who study moral story framing and judgment from different perspectives. For example, Zhou et al. (2021) analyzed the relationship between linguistic features pertaining to moral role (e.g., victim or aggressor) and moral valence. Botzer et al. (2022) studied language models' ability to classify moral judgments based on a described story. Recently, Nguyen et al. (2022, cf. Chapter 2)]nguyenMappingTopics1002022 mapped the AITA moral domain, consisting of over 100K stories, into a set of 47 topics encompassing several aspects of daily life such as *marriage*, *work*, *appearance* and *religion*. The authors found that topics are an important covariate in studying everyday moral dilemmas, and demonstrated that the framing (how a story is told) and judgment (how people perceive a story's author) vary across topics and topic pairs in non-trivial ways.

In this section, we perform two analyses. First, we replicate a study of AITA content using moral foundations in Chapter 2. We show that the detection of foundations using the MFD 2.0, which was used in that study, significantly differs from our method, Mformer. This

leads to non-trivial differences in the downstream results and, hence, interpretation. Second, we demonstrate the utility of the MFT in studying conflicting judgments within a thread, in which different moral valence appeal to moral foundations in distinct ways. These findings echo the prior work's argument that MFT is a useful framework to study moral judgments; however, researchers aiming to adopt the theory for studying moral content should be aware of word count's limitations.

B.8.1 Moral foundation prevalence in topics and topic pairs

In the first analysis, we replicate the study performed in Chapter 2. Specifically, we aim to see the difference in their downstream results of moral foundation prevalence when the posts and judgments on AITA are scored by our more performant Mformer. In the below, we describe this study in more detail.

B.8.1.1 The AITA subreddit and dataset

AITA is organized into discussion *threads*. Each thread contains a *post* made by an *original poster* (OP) and *comments* made by Reddit users. A post, which describes a moral conflict its OP experiences, contains a short *title* of the form "AITA (am I the a**h****)..." or "WIBTA (would I be the a**h****)...", followed by a *body text* which gives more detail to the story. Other users make comments below a post detailing their judgment and reasoning.

Five moral judgments are used on AITA: `YTA` ("you're the a**h***"), `NTA` ("not the a**h***"), `ESH` ("everyone sucks") and `NAH` ("no a**h****s here") and `INFO` ("more information needed"). Each time a user makes a judgment, they must give one of these acronyms. Reddit users can upvote or downvote the post and comments, and the *score* of a post/comment is the difference between its upvotes and downvotes. After 18 hours since a post was made, a Reddit bot assigns the judgment in the highest-scoring comment as the post's *verdict*.

We use the dataset released by Nguyen et al. (2022),[9] which contains 102,998 threads, including about 13M comments, from 2014 to 2019. For each post, we also use its highest-scoring comment as the final verdict. We remove all threads for which the final verdict is `INFO`, since we are only interested in those with a concrete verdict. Finally, we place the rest four judgments into two groups: `YA` (containing `YTA` and `ESH`, representing judgments with *negative valence* against the OP) and `NA` (containing `NTA` and `NAH`, representing judgments with *positive valence* toward the OP). The final dataset contains 94,970 post-verdict pairs.

B.8.1.2 Labeling posts and verdicts with moral foundations

To predict the presence of moral foundations within a post or verdict, we use two methods:

- Word count with MFD 2.0: this was used by Nguyen et al. (2022, cf. Chapter 2). Specifically, if a post/verdict contains a word that maps to foundation f according to MFD 2.0, then the post/verdict is considered to contain f.

[9] https://doi.org/10.5281/zenodo.6791835

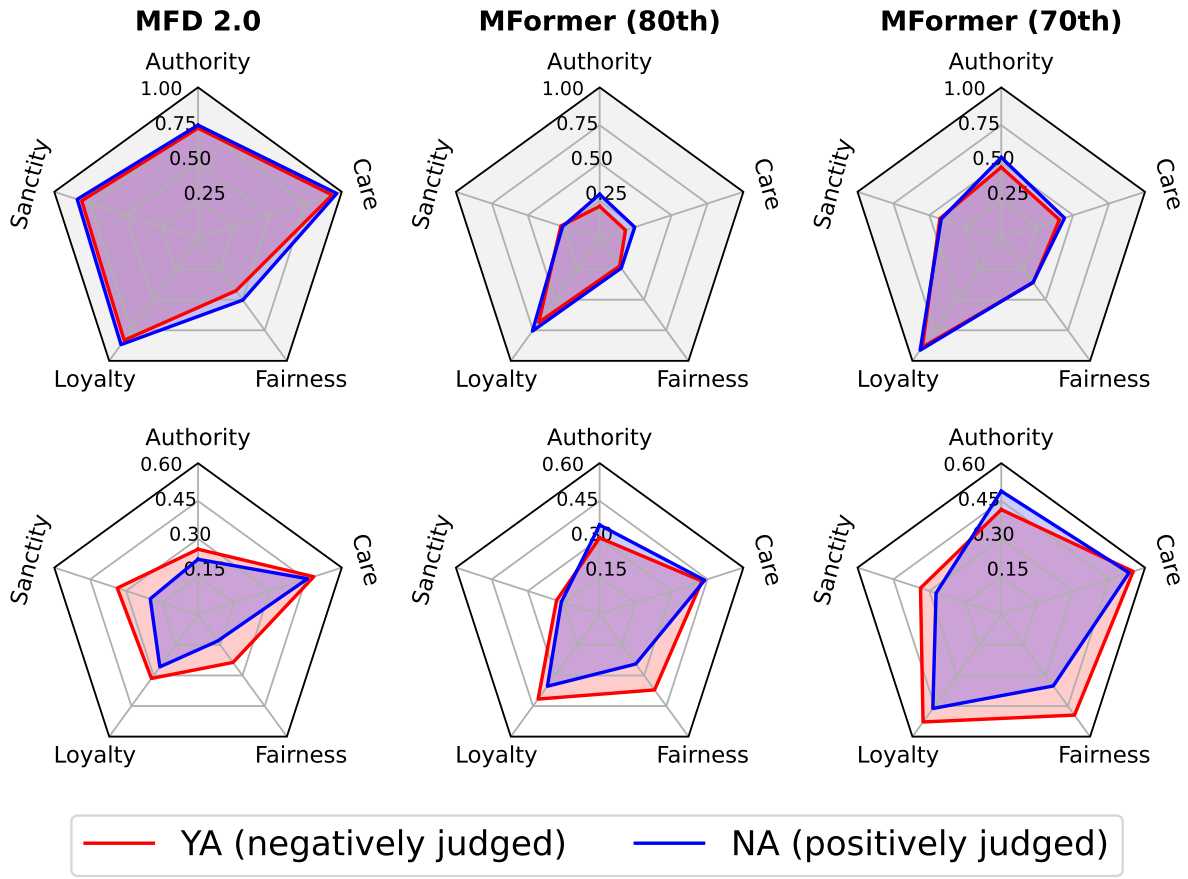


Figure B.8: Posts (top) and verdicts (bottom) in the (*family*, *marriage*) topic pair on AITA. Each number in a radar plot indicates the proportion of posts (or verdicts) that contain each moral foundation. The moral foundations are detected by two methods: MFD 2.0 and Mformer. For Mformer, two thresholding values are displayed (80th and 70th percentiles). Red (resp. blue) indicates YA (resp. NA) valence.

- Mformer: we score every post and verdict on the five foundations using Mformer in Section 3.4.2.1. Then, for each foundation f, we assign a positive label to the instances in the top 20% of the scores. That is, to contain foundation f, a post/verdict must score higher than at least 80% of the dataset.

B.8.1.3 Measuring moral foundation prevalence in topics and topic pairs

In Chapter 2, we used topic modeling to map the domain AITA to 47 fine-grained topics such as *family*, *religion* and *appearance*. They demonstrated three important points. First, a post is typically best described by two nominal topics such as *family* and *money*, hence the focus on *topic pairs* as the thematic unit of the data. Second, topic pairs are an important covariate in AITA analysis, in which the authors found that every topic pair appeals to moral foundations in different ways. For example, when talking about *family* and *marriage* (Figure 3.4, first

column), an OP tends to focus on the foundations *authority*, *care*, *loyalty* and *sanctity* but relatively downplays the aspect of *fairness*. On the other hand, verdicts in this topic pair only primarily focus on *care*. And third, moral foundations can help characterize the difference between positive (NA) and negative (YA) verdicts within each topic pair. Specifically, the study found that YA judgments typically appeal to all moral foundations more often than do NA judgments.

We perform the same measurements, with the only difference in the moral foundation labels: instead of using MFD 2.0, we replace them with Mformer-predicted labels. Then, for each topic pair, we count the number of posts that contain a foundation f and divide that by the total number of posts in that topic pair. This ratio gives the *prevalence* for f, which is interpreted as the frequency with which a post in this topic pair is about f. We do the same for the verdicts. Finally, we also separated the posts and verdicts into YA and NA classes to investigate any difference between them. The prevalence of moral foundations within each topic is presented in Figure B.9 (for posts) and Figure B.10 (for verdicts). Additionally, Figure 3.4 in the main text displays the foundation prevalence for posts and verdicts in the topic pair (*family*, *marriage*).

B.8.1.4 Results

We first compare the differences in foundation prevalence reported in Chapter 2—where we used the MFD 2.0 to infer the foundations within each post and verdict—and that produced using Mformer, which is in Figures B.9 and B.10 in this work. In most topics, the prevalence of moral foundations changes dramatically. For example, in Chapter 2 we showed that among posts in the topic *family*, most of them are concerned with *care*, *loyalty*, and *sanctity*. However, in our result, we find that the only dominating foundation is *loyalty*, whereas the rest four foundations are somewhat equal in their prevalence. In another example, within verdicts in the topic *roommates*, we found that foundation *care* is the most relevant; we, on the other hand, find that it is *fairness* that verdicts tend to highlight the most.

This difference is also observable, even to a larger extent, when topic pairs are considered. In addition to our discussion of Figure 3.4 in the main text, here we also show that the relative importance of moral foundations produced by Mformer is not sensitive to the cutoff level of 80th percentile. In Figure B.8, we present in the last column the same radar plots but at the 70th-percentile cutoff level, i.e., for each post/verdict to contain a foundation, it must score higher than at least 70% of all posts/verdicts. The pentagons' shapes on the second and third columns of Figure B.8 do not change, indicating that they are not sensitive to the binary classification threshold of Mformer's predictions.

B.8.2 Characterizing conflicting judgments in highly controversial discussions

In the second, and novel, analysis, we aim to investigate in more detail the *positive* (NA) and *negative* (YA) judgments on AITA from the perspective of the MFT. Specifically, we hypothesize that in a controversial, highly conflicting thread, there is a systematic difference in the way NA judgments are made compared to YA judgments when it comes to which moral foundations they adhere to. Below we give more detail of this analysis.

B.8.2.1 Dataset

We focus on threads with highly conflicting communities of judgment. To do so, we search over all threads in the dataset in Chapter 2 and only keep threads satisfying the following conditions:

- Only comments within 18 hours from the time the original post was made are considered.[10]
- Only top-level comments are considered. That is, these comments must reply directly to their original post but not to another comment.
- Only comments with a valid judgment (`YTA`, `NTA`, `ESH`, `NAH`) are kept. We remove posts labeled with `INFO` as they do not present a concrete judgment. Then, `YTA` and `ESH` comments are grouped into `YA` judgments of negative valence. Similarly, `NTA` and `NAH` comments are grouped into `NA` judgments of positive valence.
- To be considered controversial, each thread must contain at least 50 such judgments. Further, the proportion of `YA` or `NA` judgments in a thread must not exceed 70% of all judgments.[11]

Based on these, we obtain a dataset of 2,135 threads with a total of 466,485 judgments. Each thread contains an original post and a median of 110 judgments (max = 3,251).

B.8.2.2 Labeling posts and judgments with moral foundations

We use the Mformer models to score every post and judgment on the five moral foundations. Similar to before, we label a post/judgment with a foundation f if it scores in the top 20% of the dataset. In other words, to contain f a post/judgment must score higher than at least 80% of all posts or judgments.

B.8.2.3 Comparing `YA` and `NA` judgments based on moral foundations

To estimate the difference in their appeal to moral foundations between `YA` and `NA` judgments in one thread, we use the odds ratio (OR) defined as

$$\mathrm{OR}_f = \frac{N_{f,\texttt{NA}} \cdot N_{\neg f,\texttt{NA}}}{N_{\neg f,\texttt{NA}} \cdot N_{f,\texttt{YA}}}, \tag{B.4}$$

where

- f is one of the five foundations;
- $N_{f,\texttt{NA}}$ is the number of `NA` judgments labeled *with* foundation f;
- $N_{\neg f,\texttt{YA}}$ is the number of `YA` judgments labeled *without* foundation f;

[10] This condition is consistent with the way the AITA bot determines the final verdict for each post. We do aim to argue that the comments made after this period are less important.

[11] This condition is inspired by the "AITA Filtered" subreddit: https://www.reddit.com/r/AITAFiltered.

- $N_{\neg f,\texttt{NA}}$ is the number of `NA` judgments labeled *without* foundation f; and
- $N_{f,\texttt{YA}}$ is the number of `YA` judgments labeled *with* foundation f.

A value of OR greater than 1 (resp. less than 1) means that the presence of foundation f in a judgment raises the odds that the judgment is `NA` (resp. `YA`). For example, if $\text{OR}_{loyalty} = 2.5$, then the presence of the foundation *loyalty* raises the odds that a judgment is `NA` (i.e., positive toward the author) by 2.5 times. On the other hand, if $\text{OR}_{loyalty} = 0.2$, then the presence of *loyalty* raises the odds that a judgment is `YA` (i.e., negative toward the author) by $1/0.2 = 5$ times. Finally, to establish statistical significance, we can estimate the standard error of the natural logarithm of the OR as

$$\text{SE} = \sqrt{\frac{1}{N_{f,\texttt{NA}}} + \frac{1}{N_{\neg f,\texttt{YA}}} + \frac{1}{N_{\neg f,\texttt{NA}}} + \frac{1}{N_{f,\texttt{YA}}}}. \tag{B.5}$$

The log OR can be approximated by the normal distribution $\mathcal{N}(\log(\text{OR}_f), \text{SE}^2)$ and, hence, the 95% confidence interval for the log OR is $(\log(\text{OR}_f) - 1.96\text{SE}, \log(\text{OR}_f) + 1.96\text{SE})$. In presenting the results, we often use the following two conventions:

- If $\text{OR}_f > 1$, that is, when f is associated with `NA` judgments, we present the OR and its 95% CI as OR_f and $(\text{OR}_f - \exp(1.96\text{SE}), \text{OR}_f + \exp(1.96\text{SE}))$, respectively.
- If $\text{OR}_f < 1$, that is, when f is associated with `YA` judgments, we present the OR its 95% CI as $\frac{1}{\text{OR}_f}$ and $\left(\frac{1}{\text{OR}+\exp(1.96\text{SE})}, \frac{1}{\text{OR}-\exp(1.96\text{SE})}\right)$, respectively. In this case, the OR represents the number of times the odds of a `YA` judgment is raised in the presence of foundation f.

If the 95% CI does not contain 1, then the OR is said to be significant at the 0.05 level.

B.9 Moral Foundations and Stance toward Controversial Topics

Stance, in general, can be understood as a person's opinion—in favor, against or neutral—toward a proposition or target topic (Mohammad et al., 2017). In this work, we analyze the association between people's stances on some controversial topics and moral foundations. We use the dataset by Mohammad et al. (2016), which contains three components in each of its instances: a topic, a tweet, and the stance of the tweet's author toward the topic. We aim to reproduce some of the results by Rezapour et al. (2021) using our new method of detecting moral foundations, i.e., Mformer. We also present new results in light of our classifiers.

B.9.1 Dataset

This data contains 4,870 instances each with three components: a topic, a tweet, and an annotated stance. Six "controversial" topics are considered: *Atheism*, *Climate Change is a Real Concern* (henceforth *Climate Change*), *Donald Trump*, *Feminist Movement*, *Hillary Clinton* and *Legalization of Abortion* (henceforth *Abortion*). Each tweet was annotated with a stance toward

Table B.8: Example tweets and their authors' stance on some controversial topics.

Topic	Stance	# Tweets	Example
Atheism	In favor	124	Now that the SCOC has ruled Canadians have freedom from religion, can someone tell Harper to dummy his 'god bless Canada'. #cdnpoli
	Against	464	dear lord thank u for all of ur blessings forgive my sins lord give me strength and energy for this busy day ahead #blessed #hope
Climate Change	In favor	335	We cant deny it, its really happening.
	Against	26	The Climate Change people are disgusting assholes. Money transfer scheme for elite. May you rot.
Donald Trump	In favor	148	Donald Trump isn't afraid to roast everyone.
	Against	299	@ABC Stupid is as stupid does! Showedhis true colors; seems that he ignores that US was invaded, & plundered,not discovered
Feminist Movement	In favor	268	Always a delight to see chest-drumming alpha males hiss and scuttle backwards up the wall when a feminist enters the room. #manly
	Against	511	If feminists spent 1/2 as much time reading papers as they do tumblr they would be real people, not ignorant sexist bigots.
Hillary Clinton	In favor	163	Hillary is our best choice if we truly want to continue being a progressive nation. #Ohio
	Against	565	@tedcruz And, #HandOverTheServer she wiped clean + 30k deleted emails, explains dereliction of duty/lies re #Benghazi,etc #tcot
Abortion	In favor	167	@tooprettyclub Are you OK with #GOP males telling you what you can and can't do with your own body?
	Against	544	Just laid down the law on abortion in my bioethics class. #Catholic

Table B.9: Comparison moral foundation scores (produced by fine-tuned Mformer models) between tweets in favor and those against each controversial topic. **F** >A indicates that a randomly chosen tweet in favor of a topic scores significantly higher than a randomly chosen tweet against that topic. Similar for F <**A**. The higher-scoring stance is in **bold**. Statistical significance is established by the two-sided Mann-Whitney U test using the asymptotic method with continuity correction. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. All insignificant results at the 0.05 level are replaced by the "–" symbol.

Topic	# Favor	# Against	Moral foundations				
			Authority	Care	Fairness	Loyalty	Sanctity
Atheism	124	464	**F** >A*	F <**A****	**F** >A***	F <**A****	F <**A****
Climate Change	335	26	F <**A***	–	F <**A***	–	–
Donald Trump	148	299	**F** >A***	F <**A****	F <**A****	**F** >A***	–
Feminist Movement	268	511	F <**A***	F <**A***	–	–	–
Hillary Clinton	163	565	F <**A****	–	F <**A****	–	F <**A****
Abortion	167	544	–	F <**A***	**F** >A***	–	F <**A****

the topic it is about: whether the author of the tweet is in favor, against, or neither toward the topic.

For each topic, our goal is to characterize the difference between those in favor of and those against it; therefore, we remove all tweets labeled with "None" as its author's stance, resulting in 3,614 tweets left. Table B.8 presents some example tweets for each stance in every topic as well as the number of tweets for each stance. In all topics except for *climate change*, there are more tweets against it than there are tweets in favor of it.

Finally, we score every tweet using our five Mformer moral foundation classifiers.

B.9.2 Comparing conflicting stances on each topic using moral foundation scores

Here we examine the difference in foundation scores for tweets in favor of a topic and those against it. Formally, for each topic t and foundation f, we compare Mformer scores when predicting whether f exists in two groups: tweets in favor of topic t and those against topic t. We choose the Mann-Whitney U test to examine statistically significant differences in foundation scores between the groups.

Table B.9 presents the results. For some topics, there are clear differences between the two conflicting stances. For example, on the topic *atheism*, tweets in favor of it score significantly higher on *authority* ($p < 0.05$) and *fairness* ($p < 0.001$). These tweets often criticize the political authority of religion, as in the tweet "Religious leaders are like political leaders - they say what they think people want to hear. #freethinker"; some also focus on discrimination, e.g., "if u discriminate based on ur religion, be ready to be discriminated against for having that religion." On the other hand, those against this topic score higher on *care* ($p < 0.01$), *loyalty* ($p < 0.001$), and *sanctity* ($p < 0.001$). They often focus on the virtues of compassion, faithfulness, and holiness such as the tweet "Give us this day our daily bread, and forgive us our sins as we forgive those who sin against us. #rosary #God #teamjesus."

For some other topics, we only find one-way difference between the stances. For instance, on the topic *feminist movement*, tweets against it score significantly higher than those in favor

of it on two foundations: *authority* ($p < 0.05$) and *care* ($p < 0.05$). Those who voice their disagreement with the movement often mention government authority such as in the tweet "I hate government, but if I were in government i'd want to be a District Attorney or a Judge to hold #YesAllWomen accountable.". However, there is no significant evidence suggesting that tweets in favor of the movement appeal to any foundation more than those against it.

B.9.3 Comparison based on binary predictions

In this subsection, we estimate the effect sizes of the difference between tweets in favor and those against a topic with respect to their adherence to moral foundations. First, we convert all raw scores outputted by Mformer into *binary labels* as follows. For each foundation f, we give the binary label 1 to the highest-scoring 20% of the tweets. In other words, to be considered to contain f, a tweet must score higher than at least 80% of all tweets in the dataset. We choose a rather high threshold as we prefer high precision at the cost of a lower recall. Refer to Appendix B.6.2 for a more detailed discussion on setting binary classification thresholds for Mformer; results on the test set suggest that setting the threshold at the 80th percentile is reasonable as it gives both high F-1 and accuracy.

B.9.3.1 Significance of association between stance and moral foundations

We replicate a prior analysis by Rezapour et al. (2021) in which the authors used the chi-square test to examine the dependence between two binary variables: whether a foundation is found in a tweet and whether the tweet is in favor or against some topic. Whereas Rezapour et al. (2021) labeled each tweet with a moral foundation using the MFD,[12] which has been shown in this work to produce misleading results, we use the labels produced by threshold Mformer prediction scores described above. We use the scipy implementation of this test and invoke Yates' correction for continuity.

The results of this significance test are given in Table 3.4. The entries in the "MFD" columns are taken from Rezapour et al. (2021, Table 5) and those based on our method are in the "Mformer" columns. We first observe some similar results: for example, Rezapour et al. (2021) found no significant result for the topic *feminist movement*, which agrees with our analysis. Similarly, the prior finding that there is a significant dependence between the foundation *fairness* and the stance toward the topic *abortion* agrees with our result, although our result yields a higher level of significance ($p < 0.001$ compared to $p < 0.05$).

However, we find that a lot of significant dependencies discovered by Mformer were overlooked by the prior study. For example, in (Rezapour et al., 2021), the authors found no significant results for the topics *Donald Trump* and *climate change*. Our findings, on the other hand, show that very strong associations in these topics exist. For example, on the topic *climate change*, we find a strong relationship between stance and *fairness* ($p < 0.001$). On the topic *Donald Trump*, three foundations correlate significantly with stance: *authority* ($p < 0.001$), *care* ($p < 0.05$) and *loyalty* ($p < 0.001$).

[12]The authors used a slightly modified version of the MFD in which each word is annotated with its part of speech (Rezapour et al., 2019).

B.9.3.2 Odds ratios between moral foundations and stance toward a topic

Finally, to quantify the association between the presence of a moral foundation in a tweet and the tweet's stance toward a topic, we calculate the odds ratios (ORs) as follows. In each topic, for foundation f, the OR between f and stance is

$$\mathrm{OR}_f = \frac{N_{f,\texttt{favor}} \cdot N_{\neg f,\texttt{against}}}{N_{\neg f,\texttt{favor}} \cdot N_{f,\texttt{against}}}, \tag{B.6}$$

where

- $N_{f,\texttt{favor}}$ is the number of tweets labeled *with* foundation f and are *in favor* of the topic;
- $N_{\neg f,\texttt{against}}$ is the number of tweets labeled *without* foundation f and are *against* the topic;
- $N_{\neg f,\texttt{favor}}$ is the number of tweets labeled *without* foundation f and are *in favor* of the topic; and
- $N_{f,\texttt{against}}$ is the number of tweets labeled *with* foundation f and are *against* the topic.

If there is no association between f and stance, the OR equals 1. A value of OR greater than 1 (resp. less than 1) means that the presence of foundation f in a tweet raises the odds that the tweet is in favor of (resp. against) the target topic. The standard error for the natural logarithm of the OR is

$$\mathrm{SE} = \sqrt{\frac{1}{N_{f,\texttt{favor}}} + \frac{1}{N_{\neg f,\texttt{against}}} + \frac{1}{N_{\neg f,\texttt{favor}}} + \frac{1}{N_{f,\texttt{against}}}}. \tag{B.7}$$

We calculate the 95% confidence interval (CI) of the OR as $(\mathrm{OR} - \exp(1.96\mathrm{SE}), \mathrm{OR} + \exp(1.96\mathrm{SE}))$. If this interval does not contain 1, we consider the strength of the relationship statistically significant. Equivalently, we present the log ORs and their CIs: if an interval contains 0 the OR is significant.

Figure 3.6 presents the log ORs with their 95% CIs for all moral foundations in each topic. Significant values are plotted with colors: blue for positive (i.e., a moral foundation is associated with the "in favor" stance) and red for negative (association with the "against" stance). We note that the missing value for the foundation *loyalty* on the topic *climate change* is because no tweet was predicted to contain *loyalty*.

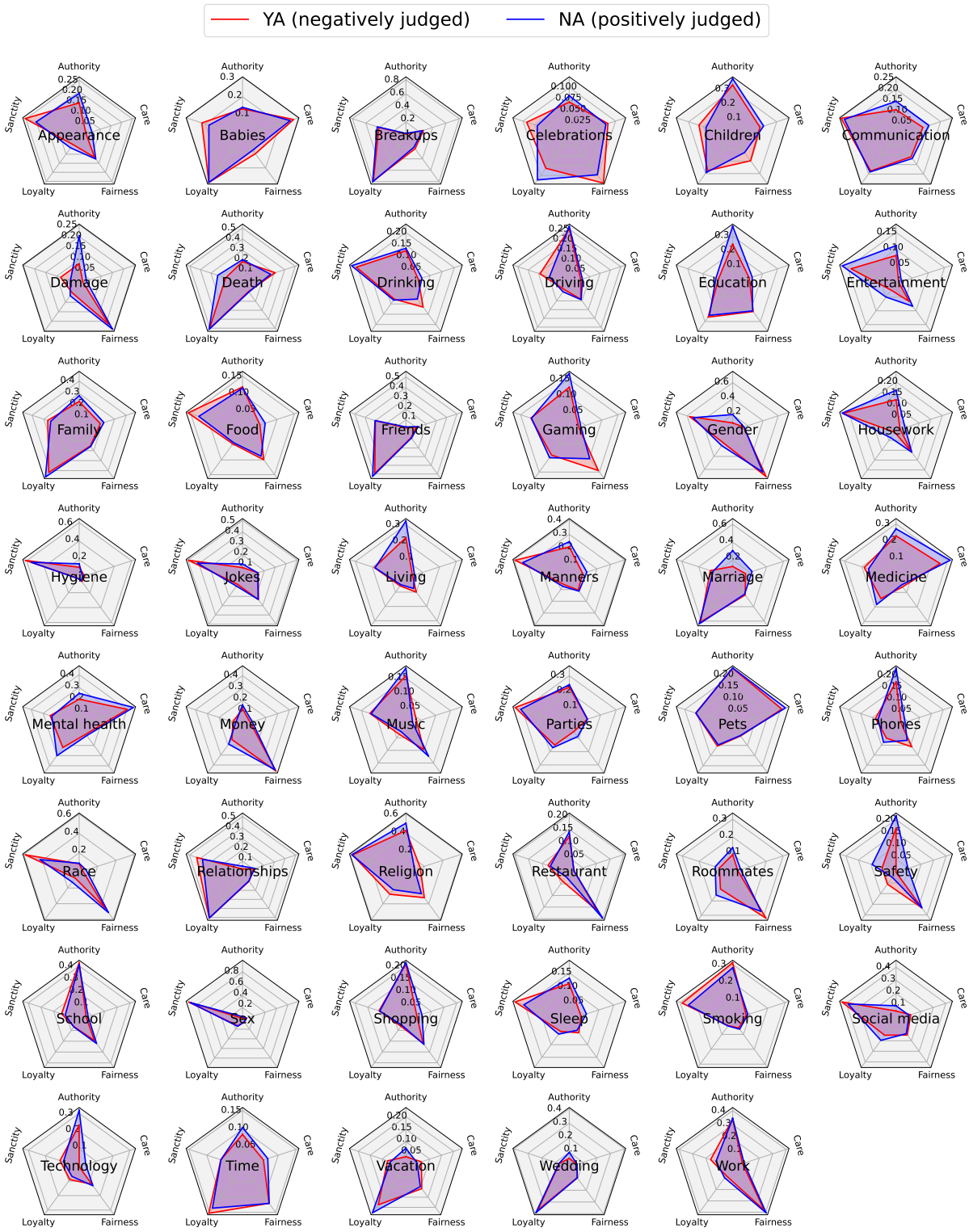


Figure B.9: Prevalence of each moral foundation in each of the 47 topics on AITA. In each radar plot for a topic, each vertex represents the proportion of posts in that topic that contain the corresponding moral foundation. The moral foundations are predicted using our Mformer model. Blue (resp. red) pentagons correspond to NA-judged (resp. YA-judged) posts. These radar plots are reproduced from (Nguyen et al., 2022, Fig. G2), which was made using MFD 2.0.

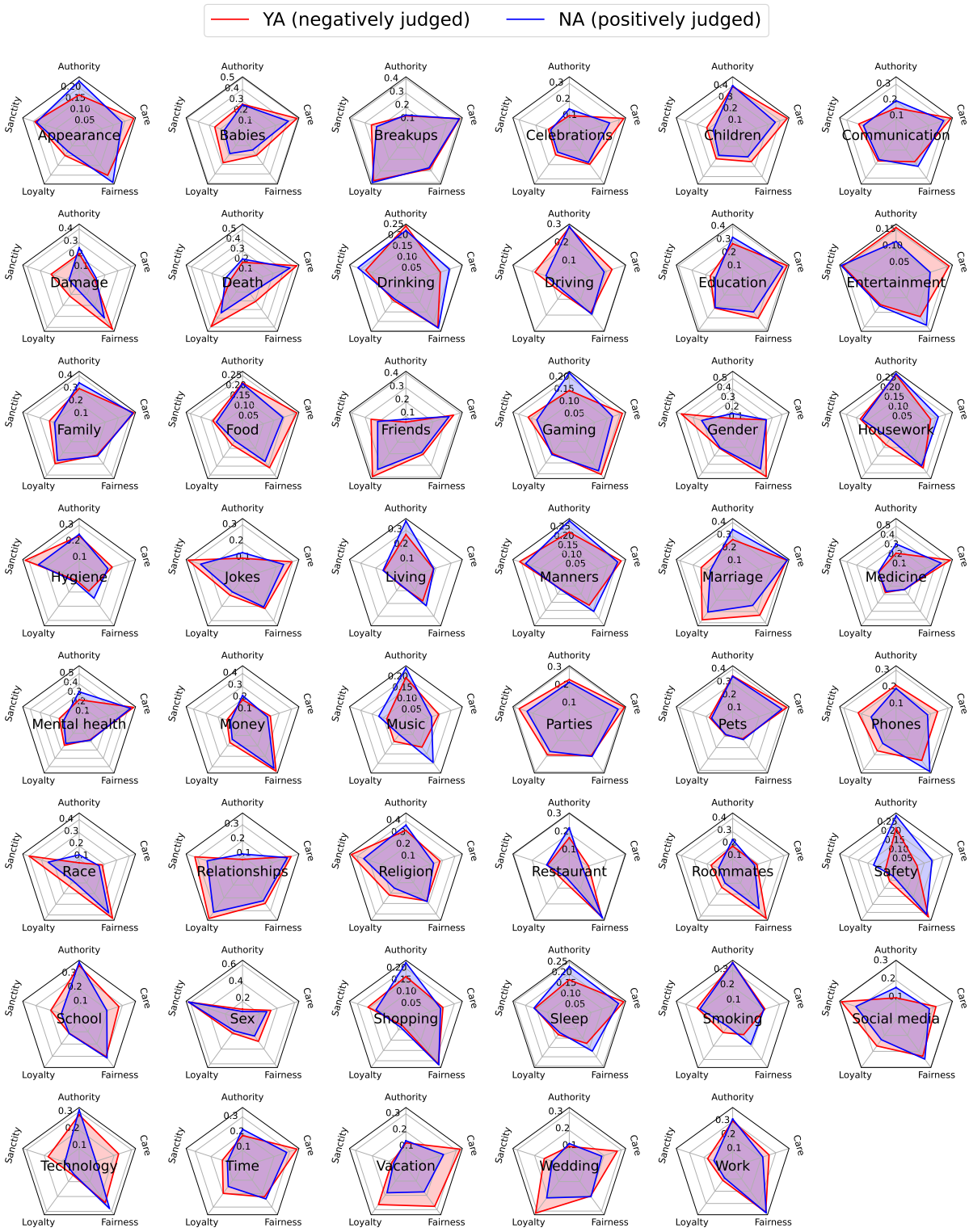


Figure B.10: Prevalence of each moral foundation in each of the 47 topics on AITA. In each radar plot for a topic, each vertex represents the proportion of verdicts in that topic that contain the corresponding moral foundation. The moral foundations are predicted using our Mformer model. Blue (resp. red) pentagons correspond to NA (resp. YA) judgments. These radar plots are reproduced from (Nguyen et al., 2022, Fig. G1), which was made using MFD 2.0.

Bibliography

Amin, A. B.; Bednarczyk, R. A.; Ray, C. E.; Melchiori, K. J.; Graham, J.; Huntsinger, J. R.; and Omer, S. B., 2017. Association of moral values with vaccine hesitancy. *Nature Human Behaviour*, 1, 12 (2017), 873–880. doi:10.1038/s41562-017-0256-5.

Antoniak, M.; Mimno, D.; and Levy, K., 2019. Narrative Paths and Negotiation of Power in Birth Stories. *Proceedings of the ACM on Human-Computer Interaction*, 3, CSCW (2019), 1–27. doi:10.1145/3359190.

Appiah, K. A., 2008. *Experiments in Ethics*. The Mary Flexner Lectures. Harvard University Press, Cambridge, Mass.

Atari, M.; Haidt, J.; Graham, J.; Koleva, S.; Stevens, S. T.; and Dehghani, M., In Press. Morality Beyond the WEIRD: How the Nomological Network of Morality Varies Across Cultures. *Journal of Personality and Social Psychology*, (In Press). doi:10.31234/osf.io/q6c9r.

Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; and Rahwan, I., 2018. The Moral Machine experiment. *Nature*, 563, 7729 (2018), 59–64. doi:10.1038/s41586-018-0637-6.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J., 2020. The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14 (2020), 830–839. doi:10.1609/icwsm.v14i1.7347.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I., 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, Jan (2003), 993–1022.

Botzer, N.; Gu, S.; and Weninger, T., 2022. Analysis of Moral Judgment on Reddit. *IEEE Transactions on Computational Social Systems*, (2022), 1–11. doi:10.1109/TCSS.2022.3160677.

Boyd-Graber, J.; Hu, Y.; and Mimno, D., 2017. Applications of Topic Models. *Foundations and Trends® in Information Retrieval*, 11, 2-3 (2017), 143–296. doi:10.1561/1500000030.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; 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.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D., 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, vol. 33, 1877–1901. Curran Associates, Inc.

Chang, J.; Gerrish, S.; Wang, C.; Boyd-graber, J.; and Blei, D., 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, vol. 22, 288–296. Curran Associates, Inc.

Choi, Y., 2022. The Curious Case of Commonsense Intelligence. *Daedalus*, 151, 2 (2022), 139–155. doi:10.1162/daed_a_01906.

Clifford, S.; Iyengar, V.; Cabeza, R.; and Sinnott-Armstrong, W., 2015. Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior Research Methods*, 47, 4 (2015), 1178–1198. doi:10.3758/s13428-014-0551-2.

Clifford, S. and 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 (2013), 659–671. doi:10.1017/S0022381613000492.

Collins, S., 2015. *The Core of Care Ethics*. Palgrave Macmillan, New York, NY.

Curry, O. S., 2016. Morality as Cooperation: A Problem-Centred Approach. In *The Evolution of Morality* (Eds. T. K. Shackelford and R. D. Hansen), 27–51. Springer International Publishing, Cham.

Curry, O. S.; Mullins, D. A.; and Whitehouse, H., 2019. Is It Good to Cooperate? Testing the Theory of Morality-as-Cooperation in 60 Societies. *Current Anthropology*, 60, 1 (Feb. 2019), 47–69. doi:10.1086/701478.

Dancy, J., 1983. Ethical particularism and morally relevant properties. *Mind*, 92, 368 (1983), 530–547.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota. doi:10.18653/v1/N19-1423.

Driver, J., 1992. The suberogatory. *Australasian Journal of Philosophy*, 70, 3 (1992), 286–295. doi:10.1080/00048409212345181.

Dunn, J. C., 1973. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, 3, 3 (1973), 32–57. doi:10.1080/01969727308546046.

Dupré, J., 1993. *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*. Harvard University Press, Cambridge, Mass.

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

Fast, E.; Chen, B.; and Bernstein, M. S., 2016. Empath: Understanding Topic Signals in Large-Scale Text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4647–4657. ACM, San Jose California USA. doi:10.1145/2858036.2858535.

Fawcett, T., 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, 27, 8 (2006), 861–874. doi:10.1016/j.patrec.2005.10.010.

Feinberg, M. and Willer, R., 2013. The Moral Roots of Environmental Attitudes. *Psychological Science*, 24, 1 (2013), 56–62. doi:10.1177/0956797612449177.

Fleiss, J. L., 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 5 (1971), 378–382. doi:10.1037/h0031619.

Foot, P., 1972. Morality as a System of Hypothetical Imperatives. *The Philosophical Review*, 81, 3 (1972), 305. doi:10.2307/2184328.

Forbes, M.; Hwang, J. D.; Shwartz, V.; Sap, M.; and Choi, Y., 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 653–670. Association for Computational Linguistics, Online. doi:10.18653/v1/2020.emnlp-main.48.

Freedman, R.; Schaich Borg, J.; Sinnott-Armstrong, W.; Dickerson, J.; and Conitzer, V., 2018. Adapting a Kidney Exchange Algorithm to Align With Human Values. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32, 1 (2018), 1636–1643. doi:10.1609/aaai.v32i1.11505.

Frimer, J., 2019. Moral Foundations Dictionary 2.0. doi:10.17605/OSF.IO/EZN37.

Gaffney, D. and Matias, J. N., 2018. Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *PLOS ONE*, 13, 7 (2018), e0200162. doi:10.1371/journal.pone.0200162.

Garten, J.; Hoover, J.; Johnson, K. M.; Boghrati, R.; Iskiwitch, C.; and Dehghani, M., 2018. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis: Distributed dictionary representation. *Behavior Research Methods*, 50, 1 (2018), 344–361. doi:10.3758/s13428-017-0875-9.

Giorgi, S.; Zhao, K.; Feng, A. H.; and Martin, L. J., 2023. Author as Character and Narrator: Deconstructing Personal Narratives from the r/AmITheAsshole Reddit Community. *Proceedings of the International AAAI Conference on Web and Social Media*, 17 (2023), 233–244. doi:10.1609/icwsm.v17i1.22141.

González-Bailón, S. and Paltoglou, G., 2015. Signals of Public Opinion in Online Communication: A Comparison of Methods and Data Sources. *The ANNALS of the American Academy of Political and Social Science*, 659, 1 (2015), 95–107. doi:10.1177/0002716215569192.

Graham, J. and Haidt, J., 2012. The Moral Foundations Dictionary. https://moralfoundations.org/other-materials/.

Graham, J.; Haidt, J.; Koleva, S.; Motyl, M.; Iyer, R.; Wojcik, S. P.; and Ditto, P. H., 2013. Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In *Advances in Experimental Social Psychology* (Eds. P. Devine and A. Plant), vol. 47, 55–130. Academic Press.

Graham, J.; Haidt, J.; and Nosek, B. A., 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96, 5 (2009), 1029–1046. doi:10.1037/a0015141.

Graham, J.; Nosek, B. A.; Haidt, J.; Iyer, R.; Koleva, S.; and Ditto, P. H., 2011. Mapping the moral domain. *Journal of Personality and Social Psychology*, 101, 2 (2011), 366–385. doi:10.1037/a0021847.

Greene, J. D.; Sommerville, R. B.; Nystrom, L. E.; Darley, J. M.; and Cohen, J. D., 2001. An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science*, 293, 5537 (2001), 2105–2108. doi:10.1126/science.1062872.

Griffiths, T. L. and Steyvers, M., 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101, suppl_1 (2004), 5228–5235. doi:10.1073/pnas.0307752101.

Guo, S.; Mokhberian, N.; and Lerman, K., 2023. A Data Fusion Framework for Multi-Domain Morality Learning. *Proceedings of the International AAAI Conference on Web and Social Media*, 17 (2023), 281–291. doi:10.1609/icwsm.v17i1.22145.

Habernal, I. and Gurevych, I., 2016. What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in Web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1214–1223. Association for Computational Linguistics, Austin, Texas. doi:10.18653/v1/D16-1129.

Haidt, J., 2013. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Vintage, New York.

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

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

Haworth, E.; Grover, T.; Langston, J.; Patel, A.; West, J.; and Williams, A. C., 2021. Classifying Reasonability in Retellings of Personal Events Shared on Social Media: A Preliminary Case Study with /r/AmITheAsshole. *Proceedings of the International AAAI Conference on Web and Social Media*, 15 (2021), 1075–1079. doi:10.1609/icwsm.v15i1.18133.

Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J., 2021. Aligning AI with shared human values. In *International Conference on Learning Representations*. Vienna, Austria.

Hofmann, T., 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50–57. ACM, Berkeley, California, USA. doi:10.1145/312624.312649.

Honnibal, M.; Montani, I.; Van Landeghem, S.; and Boyd, A., 2020. spaCy: Industrial-strength natural language processing in python. (2020). doi:10.5281/zenodo.1212303.

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.; and Dehghani, M., 2020. Moral Foundations Twitter Corpus: A Collection of 35k Tweets Annotated for Moral Sentiment. *Social Psychological and Personality Science*, 11, 8 (2020), 1057–1071. doi:10.1177/1948550619876629.

Hopp, F. R.; Fisher, J. T.; Cornell, D.; Huskey, R.; and Weber, R., 2021. The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*, 53, 1 (2021), 232–246. doi:10.3758/s13428-020-01433-0.

Hopp, F. R.; Fisher, J. T.; and Weber, R., 2020. A Graph-Learning Approach for Detecting Moral Conflict in Movie Scripts. *Media and Communication*, 8, 3 (2020), 164–179. doi:10.17645/mac.v8i3.3155.

Jiang, L.; Hwang, J. D.; Bhagavatula, C.; Bras, R. L.; Liang, J.; Dodge, J.; Sakaguchi, K.; Forbes, M.; Borchardt, J.; Gabriel, S.; Tsvetkov, Y.; Etzioni, O.; Sap, M.; Rini, R.; and Choi, Y., 2022. Can Machines Learn Morality? The Delphi Experiment. *arXiv preprint arXiv:2110.07574*, (2022).

Jockers, M. L. and Mimno, D., 2013. Significant themes in 19th-century literature. *Poetics*, 41, 6 (2013), 750–769. doi:10.1016/j.poetic.2013.08.005.

Kagan, S., 1988. The additive fallacy. *Ethics*, 99, 1 (1988), 5–31.

Kobbe, J.; Rehbein, I.; Hulpus, I.; and Stuckenschmidt, H., 2020. Exploring morality in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, 30–40. Association for Computational Linguistics, Online.

Koleva, S. P.; Graham, J.; Iyer, R.; Ditto, P. H.; and Haidt, J., 2012. Tracing the threads: How five moral concerns (especially Purity) help explain culture war attitudes. *Journal of Research in Personality*, 46, 2 (2012), 184–194. doi:10.1016/j.jrp.2012.01.006.

Krügel, S.; Ostermaier, A.; and Uhl, M., 2023. ChatGPT's inconsistent moral advice influences users' judgment. *Scientific Reports*, 13, 1 (2023), 4569. doi:10.1038/s41598-023-31341-0.

Leetaru, K. and Schrodt, P. A., 2013. GDELT: Global Data on Events, Location and Tone, 1979–2012. *ISA Annual Convention*, 2, 4 (2013), 1–49.

Liscio, E.; Dondera, A.; Geadau, A.; Jonker, C.; and Murukannaiah, P., 2022. Cross-domain classification of moral values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 2727–2745. Association for Computational Linguistics, Seattle, United States. doi:10.18653/v1/2022.findings-naacl.209.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V., 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, (2019).

Loshchilov, I. and Hutter, F., 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*. New Orleans, Louisiana, United States.

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

Manning, C. D.; Raghavan, P.; and Schütze, H., 2008. *Introduction to Information Retrieval*. Cambridge university press, Cambridge.

Martin, J. and Stent, G. S., 1990. I think; therefore I thank: A philosophy of etiquette. *The American Scholar*, 59, 2 (1990), 237–254.

McAdams, D. P.; Albaugh, M.; Farber, E.; Daniels, J.; Logan, R. L.; and Olson, B., 2008. Family metaphors and moral intuitions: How conservatives and liberals narrate their lives. *Journal of Personality and Social Psychology*, 95, 4 (2008), 978–990. doi:10.1037/a0012650.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, vol. 26, 3111–3119. Curran Associates, Inc.

Mimno, D.; Wallach, H.; Talley, E.; Leenders, M.; and McCallum, A., 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 262–272. Association for Computational Linguistics, Edinburgh, Scotland, UK.

Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; and Cherry, C., 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 31–41. Association for Computational Linguistics, San Diego, California. doi:10.18653/v1/S16-1003.

Mohammad, S. M.; Sobhani, P.; and Kiritchenko, S., 2017. Stance and Sentiment in Tweets. *ACM Transactions on Internet Technology*, 17, 3 (2017), 1–23. doi:10.1145/3003433.

Mokhberian, N.; Abeliuk, A.; Cummings, P.; and Lerman, K., 2020. Moral Framing and Ideological Bias of News. In *Social Informatics* (Eds. S. Aref; K. Bontcheva; M. Braghieri; F. Dignum; F. Giannotti; F. Grisolia; and D. Pedreschi), vol. 12467, 206–219. Springer International Publishing, Cham.

Newman, D.; Lau, J. H.; Grieser, K.; and Baldwin, T., 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 100–108. Association for Computational Linguistics, Los Angeles, California.

Newman, D. J. and Block, S., 2006. Probabilistic topic decomposition of an eighteenth-century American newspaper. *Journal of the American Society for Information Science and Technology*, 57, 6 (2006), 753–767. doi:10.1002/asi.20342.

Nguyen, T. D.; Lyall, G.; Tran, A.; Shin, M.; Carroll, N. G.; Klein, C.; and Xie, L., 2022. Mapping Topics in 100,000 Real-Life Moral Dilemmas. *Proceedings of the International AAAI Conference on Web and Social Media*, 16 (2022), 699–710. doi:10.1609/icwsm.v16i1.19327.

Olberding, A., 2016. Etiquette: A confucian contribution to moral philosophy. *Ethics*, 126, 2 (2016), 422–446. doi:10.1086/683538.

Paatero, P. and Tapper, U., 1994. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5, 2 (1994), 111–126. doi:10.1002/env.3170050203.

Pang, B. and Lee, L., 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, 2, 1–2 (2008), 1–135. doi:10.1561/1500000011.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, É., 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12 (2011), 2825–2830.

Pennington, J.; Socher, R.; and Manning, C., 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. Association for Computational Linguistics, Doha, Qatar. doi:10.3115/v1/D14-1162.

Puschmann, C. and Powell, A., 2018. Turning Words Into Consumer Preferences: How Sentiment Analysis Is Framed in Research and the News Media. *Social Media + Society*, 4, 3 (2018), 1–12. doi:10.1177/2056305118797724.

Rawls, J., 1971. *A Theory of Justice*. Harvard University Press, Cambridge, Mass.

Reimers, N. and Gurevych, I., 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Association for Computational Linguistics, Hong Kong, China. doi:10.18653/v1/D19-1410.

Rezapour, R.; Dinh, L.; and Diesner, J., 2021. Incorporating the Measurement of Moral Foundations Theory into Analyzing Stances on Controversial Topics. In *Proceedings of the 32st ACM Conference on Hypertext and Social Media*, 177–188. ACM, Virtual Event USA. doi:10.1145/3465336.3475112.

Rezapour, R.; Shah, S. H.; and Diesner, J., 2019. Enhancing the measurement of social effects by capturing morality. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 35–45. Association for Computational Linguistics, Minneapolis, USA. doi:10.18653/v1/W19-1305.

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

Schwartz, H. A. and Ungar, L. H., 2015. Data-Driven Content Analysis of Social Media: A Systematic Overview of Automated Methods. *The ANNALS of the American Academy of Political and Social Science*, 659, 1 (2015), 78–94. doi:10.1177/0002716215569197.

Scott, D. W., 2015. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, Hoboken, New Jersey, 2nd edn.

Sechidis, K.; Tsoumakas, G.; and Vlahavas, I., 2011. On the Stratification of Multi-label Data. In *Machine Learning and Knowledge Discovery in Databases* (Eds. D. Gunopulos; T. Hofmann; D. Malerba; and M. Vazirgiannis), vol. 6913, 145–158. Springer, Berlin, Heidelberg. doi:10.1007/978-3-642-23808-6_10.

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

Simonson, I. and Tversky, A., 1992. Choice in Context: Tradeoff Contrast and Extremeness Aversion. *Journal of Marketing Research*, 29, 3 (1992), 281. doi:10.2307/3172740.

Sinnott-Armstrong, W., 1988. *Moral Dilemmas*. Blackwell.

Sinnott-Armstrong, W., 2008. Framing moral intuitions. In *Moral Psychology: The Cognitive Science of Morality: Intuition and Diversity* (Ed. W. Sinnott-Armstrong), vol. 2, 47–76. The MIT Press, Cambridge, Mass.

Southwood, N., 2011. The Moral/Conventional distinction. *Mind*, 120, 479 (2011), 761–802.

Spencer, D., 2009. *Card Sorting: Designing Usable Categories*. Rosenfeld Media, Brooklyn, New York.

Szymański, P. and Kajdanowicz, T., 2017. A network perspective on stratification of multi-label data. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, vol. 74 of *Proceedings of Machine Learning Research*, 22–35. PMLR, Skopje, Macedonia.

Tan, C.; Niculae, V.; Danescu-Niculescu-Mizil, C.; and Lee, L., 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In *Proceedings of the 25th International Conference on World Wide Web*, 613–624. International World Wide Web Conferences Steering Committee, Montréal, Québec, Canada. doi:10.1145/2872427.2883081.

Tausczik, Y. R. and Pennebaker, J. W., 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29, 1 (2010), 24–54. doi:10.1177/0261927X09351676.

Tetlock, P. E., 1983. Accountability and complexity of thought. *Journal of Personality and Social Psychology*, 45, 1 (1983), 74–83. doi:10.1037/0022-3514.45.1.74.

Thomson, J. J., 1976. Killing, Letting Die, and the Trolley Problem. *Monist*, 59, 2 (1976), 204–217. doi:10.5840/monist197659224.

Trager, J.; Ziabari, A. S.; Davani, A. M.; Golazizian, P.; Karimi-Malekabadi, F.; Omrani, A.; Li, Z.; Kennedy, B.; Reimer, N. K.; Reyes, M.; Cheng, K.; Wei, M.; Merrifield, C.; Khosravi, A.; Alvarez, E.; and Dehghani, M., 2022. The Moral Foundations Reddit Corpus. *arXiv preprint arXiv:2208.05545*, (2022).

Van Leeuwen, F. and Park, J. H., 2009. Perceptions of social dangers, moral foundations, and political orientation. *Personality and Individual Differences*, 47, 3 (2009), 169–173. doi:10.1016/j.paid.2009.02.017.

Vinh, N. X.; Epps, J.; and Bailey, J., 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11, 95 (2010), 2837–2854.

von Luxburg, U.; Williamson, R. C.; and Guyon, I., 2012. Clustering: Science or art? In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, vol. 27 of *Proceedings of Machine Learning Research*, 65–79. PMLR, Bellevue, Washington, USA.

Wachsmuth, H.; Naderi, N.; Hou, Y.; Bilu, Y.; Prabhakaran, V.; Thijm, T. A.; Hirst, G.; and Stein, B., 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 176–187. Association for Computational Linguistics, Valencia, Spain.

Webb, E. J.; Campbell, D. T.; Schwartz, R. D.; and Sechrest, L., 1999. *Unobtrusive Measures*, vol. 2. Sage Publications.

Weinzierl, M. A. and Harabagiu, S. M., 2022. From Hesitancy Framings to Vaccine Hesitancy Profiles: A Journey of Stance, Ontological Commitments and Moral Foundations. *Proceedings of the International AAAI Conference on Web and Social Media*, 16 (2022), 1087–1097. doi:10.1609/icwsm.v16i1.19360.

Xi, R. and Singh, M. P., 2023a. The Blame Game: Understanding Blame Assignment in Social Media. *IEEE Transactions on Computational Social Systems*, (2023), 1–10. doi:10.1109/TCSS.2023.3261242.

Xi, R. and Singh, M. P., 2023b. Morality in the mundane: Categorizing moral reasoning in real-life social situations. *arXiv preprint arXiv:2302.12806*, (2023).

Zhou, K.; Smith, A.; and Lee, L., 2021. Assessing cognitive linguistic influences in the assignment of blame. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, 61–69. Association for Computational Linguistics, Online. doi:10.18653/v1/2021.socialnlp-1.5.

Ziems, C.; Yu, J.; Wang, Y.-C.; Halevy, A.; and Yang, D., 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3755–3773. Association for Computational Linguistics, Dublin, Ireland. doi:10.18653/v1/2022.acl-long.261.