Think of the consequences: A decade of discourse about same-sex marriage

Babak Hemmatian¹ · Sabina J. Sloman² · Uriel Cohen Priva¹ · Steven A. Sloman¹

Published online: 11 March 2019
© The Psychonomic Society, Inc. 2019

Abstract

Approaching issues through the lens of nonnegotiable values increases the perceived intractability of debate (Baron & Spranca in *Organizational Behavior and Human Decision Processes*, 70, 1–16, 1997), while focusing on the concrete consequences of policies instead results in the moderation of extreme opinions (Fernbach, Rogers, Fox, & Sloman in *Psychological Science*, 24, 939–946, 2013) and a greater likelihood of conflict resolution (Baron & Leshner in *Journal of Experimental Psychology: Applied*, 6, 183–194, 2000). Using comments on the popular social media platform Reddit from January 2006 until September 2017, we showed how changes in the framing of same-sex marriage in public discourse relate to changes in public opinion. We used a topic model to show that the contributions of certain protected-values-based topics to the debate (religious arguments and freedom of opinion) increased prior to the emergence of a public consensus in support of same-sex marriage (Gallup, 2017), and declined afterward. In contrast, the discussion of certain consequentialist topics (the impact of politicians' stance and same-sex marriage as a matter of policy) showed the opposite pattern. Our results reinforce the meaningfulness of protected values and consequentialism as relevant dimensions for describing public discourse and highlight the usefulness of unsupervised machine-learning methods in tackling questions about social attitude change.

Keywords Protected values · Consequentialism · Same-sex marriage · Latent Dirichlet allocation · Reddit

Introduction

A good negotiator will acknowledge an opponent's most basic values and not try to change them (Atran, Axelrod, & Davis, 2007). All but the most transactional of us have protected or sacred values that we regard as nonnegotiable (Baron & Spranca, 1997). *Protected values* are *deontological*, in Kant's (1797) sense: They are prescriptions or prohibitions regarding actions and are not as much affected by the outcomes of those actions. Decisions that do focus on outcomes (e.g., maximizing expected utility) are sometimes called

consequentialist. In this article, we review evidence suggesting that consequentialist discourse leads to more intellectual openness and potential for attitude change than does discourse about protected values. We tested this claim by examining the amount of discourse of each type during discussions of an issue as public opinion about that issue changed over time. The issue was whether same-sex marriage should be legal.

Protected values are common. Many refuse to consider abortion because they believe that the act is murder, and they possess a protected value that prohibits murder, regardless of the consequences of aborting or not aborting. Sometimes protected values conflict with one another. One might believe both that abortion is murder and that a woman has a fundamental right to choose what happens to her body. People use various strategies to justify violating deeply held moral principles (Baron & Leshner, 2000; Tetlock, 2003).

When two sides with incommensurable protected values are debating policy, focusing on those values is not likely to help achieve conciliation. Indeed, evidence suggests that highlighting protected values hardens attitudes. Dehghani et al. (2009) argued that the Iranian government managed to obtain support from their citizenry to pursue a nuclear

Electronic supplementary material The online version of this article (<https://doi.org/10.3758/s13428-019-01215-3>) contains supplementary material, which is available to authorized users.

✉ Babak Hemmatian
Babak_Hemmatian@Brown.edu

¹ Department of Cognitive, Linguistic and Psychological Sciences, Brown University, Providence, RI, USA

² Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA, USA

program by launching a campaign to frame possession of a nuclear industry as an inalienable right, a protected value.

A more promising approach to overcoming policy divides may be to focus on the consequences of the policy (Baron & Spranca, 1997). Asking people to consider the likely outcomes of a policy makes them more open to compromise. This type of consequentialist reasoning is a kind of causal reasoning: It is concerned with the effects of policies, the probabilities of those effects, and how good or bad they are (i.e., a different type of value).¹ Tanner, Medin, and Iliev (2008) have shown this empirically: They probed people's deontological versus consequentialist orientations and their protected values and found that protected values tended to be more aligned with a deontological than with a consequentialist orientation.

One reason that reasoning about protected values makes disputes intractable is that it simplifies—and often oversimplifies—the issues. Fernbach, Min, and Sloman (2018) found that the degree to which individuals assume protected-values-based over consequentialist orientations toward a given issue predicts how well they think they understand that issue, as well as how intractable they consider it and how extreme their opinion is. Inducing a consequentialist perspective can reveal to people how limited is their understanding. Following Rozenblit and Keil (2002), Fernbach, Rogers, Fox, and Sloman (2013) found that asking people to explain the mechanism by which a policy would lead to some consequence modulated their own sense of understanding. It also made the respondents slightly less extreme in their attitude toward the policy. This adds to the empirical evidence suggesting that when it comes to facilitating open-mindedness and compromise, consequentialist perspectives trump protected-values perspectives.

This brief review suggests that protected-values and consequentialist perspectives are, to at least some extent, matters of framing. Many people can assume both perspectives on many issues. In a period of increasing political polarization across a range of issues (Newport & Dugan, 2017), these findings suggest an avenue for improving public discourse.

We focused on the issue of same-sex marriage because attitudes toward same-sex marriage have changed considerably over the past decade, providing a successful example of consensus-building. According to the Pew Research Center, 35% of Americans favored same-sex marriage in 2001, whereas 62% did in 2017 (Pew Research Center, 2017). The shift in attitudes has licensed major politicians to come out in support of same-sex marriage in recent years, as Democrats Hillary Clinton and Barack Obama did in 2012 and 2013, respectively. Republicans like Rob Portman and Jon

Huntsman also changed their public positions in 2013. The issue came to a head in the United States on June 26, 2015, when the Supreme Court ruled on *Obergefell v. Hodges* that the constitution guarantees a right to same-sex marriage.

Shift in discourse

What kind of changes in discourse should we expect in the face of such opinion change? One possibility is that people on both sides of the issue will have well-formed protected values buttressing their position. In such cases, reframing the issue in consequentialist terms should make people more open-minded, as is suggested by evidence that, relative to protected-values-based discourse, consequentialist conversation leads people to see the complexity of an issue, to reduce their hubris, to focus on the outcomes of a policy rather than on their personal opinion, to see the possibility of compromise, and to engage in less polarized reasoning. When people start with protected values, changes in opinion should be correlated with shifts toward consequentialist discourse. Specifically, the shift in public attitudes surrounding same-sex marriage should be accompanied by a shift in public discourse away from talk about issues related to protected values (like the biblical definition of marriage or the fundamental right to marry who one wants) toward talk about concrete consequences and the causal processes that lead to them (e.g., marital benefits or the welfare of children raised by same-sex parents).

An alternative hypothesis is suggested by a quote from Justice Anthony Kennedy, writing for the majority in the *Obergefell v. Hodges* case: “No longer may this liberty be denied,” a clear framing of a pro-same-sex marriage position in terms of protected values. This illustrates that the shift in attitudes, rather than reflecting a shift from protected-values-based to consequentialist conversation, could reflect a shift from one kind of protected-values frame to another. Presumably this would occur only if people's initial positions were based on protected values that were amenable to change. Sometimes people do not have protected values that they strongly adhere to. For instance, they might rely on protected-values frames only because they are unable to generate convincing consequentialist ones (Sloman & Fernbach, 2017). Lakoff (2004) has argued that changes in opinion are governed by changes between what are essentially different protected-values frames. To illustrate, the US civil rights movement, which led to the Voting Rights Act of 1965, used arguments in favor of the protected right of all citizens to be able to vote, regardless of race, suggesting that protected-values frames likely increased in frequency in the discussion of this issue up to 1965.² It is plausible, in that case, that most people did not hold intransigent protected values about voting;

¹ Of course, any judgment about the downstream effects of a policy is ultimately rooted in some kind of *value* held about the consequences that are more or less desirable. We hold the position prevalent in the literature that the distinction between rhetoric appealing to *the moral principle expressed by a policy itself* and rhetoric appealing to *the causal chain of events a policy induces* is meaningful in thought and discourse.

² Suggested by Robert Goldstone (personal communication, September 30, 2018).

they probably had not thought much about the issue until it became politically charged. It is a distinct possibility that the shift in collective attitude toward same-sex marriage was accompanied by a shift from an anti-same-sex marriage protected-values frame (e.g., a religious one) to a pro-same-sex marriage protected-values frame (like Kennedy's).

Yet another possibility is that the greater prominence of consequentialist discourse over protected values is associated with greater consensus, but that increased consensus is what triggers such a shift in discourse, rather than the other way around. The empirical work we report was designed to reveal how the discourse about same-sex marriage has changed over the past decade. Although correlational, the method's focus on temporal trends can provide us with hints about the causal relations between protected-values-based and consequentialist discourse, on one hand, and changes in public attitude observed in national polls, on the other.

Naturally, we do not intend to suggest that the relative amount of consequentialist versus protected-values discourse is the only factor governing the public conversation. For instance, the Supreme Court decision on same-sex marriage may have had a measurable influence on what people talked about. Our method fundamentally tracks changes in the themes of discourse and can reveal such effects.

Method

Most of the research that has explored the effects of protected-values versus consequentialist framing has been performed either in an experimental setting (e.g., Tanner et al., 2008) or using field methods with small samples (e.g., Atran et al., 2007). However, the increasing popularity over the past decade of social media as a forum for discussion of socially significant topics coincided with rapid changes in public opinion surrounding same-sex marriage. This provides an opportunity to examine the prevalence of these framings in the broader discourse, using a much larger and more naturalistic sample. We used a large corpus of conversations on the popular social-media platform Reddit (<https://www.reddit.com>) to track the evolution of the discourse surrounding same-sex marriage over the past decade.

Self-touted as “the front page of the internet,” Reddit offers users the opportunity to post their opinions publicly in specialized forums (called *subreddits*), where others can upvote or downvote these opinions, making it more or less likely that other users will see the material. We chose Reddit because it is the fifth most visited website in the United States (www.redditinc.com) and because comments shared on this platform have been used in the past to uncover the

determinants of changes in opinions (Tan, Nicolae, Danescu-Niculescu-Mizil, & Lee, 2016). With billions of posts spanning a period of rapid change in public attitude toward same-sex marriage, presumably changes in the broader discourse would be reflected in the discourse on Reddit. Nevertheless, there is a potential for bias in our dataset, due to factors such as the disproportionate representation of young urban males on Reddit (Duggan & Smith, 2013) and Reddit's appropriate-use policy, which may result in the deletion of certain types of messages.

Given the size of our corpus, which included 603,282 posts, it would have been impractical to manually classify the posts as protected-values-based or consequentialist. Supervised machine-learning algorithms provide increasingly refined means of automating such tasks. However, supervised methods often depend on large annotated datasets that can be difficult to obtain. Moreover, the selection of data to annotate for the algorithm's training itself can pose difficulties and can strongly impact the algorithm's outcomes (Bishop, 2006).

We therefore used an unsupervised learning algorithm, which obviated the need for annotated data, as an initial step to characterize the main dimensions of discourse on Reddit. Latent Dirichlet allocation (LDA; Blei, Ng, & Jordan, 2003; Hoffman, Blei, & Bach, 2010) is a method for representing the gist of a body of texts (where each text is called a *document*). It assumes that the gist of every document consists of several themes (*topics*). We used this method to identify the salient themes of more than 10 years of comments related to same-sex marriage and calculated the relative contribution of each topic to Reddit discourse.

We chose LDA because past research had shown that it can provide insights into the semantic content of natural language that go beyond the level of words, partly due to its hierarchical nature, which allows for the disambiguation of different senses of a term, given its immediate context (Griffiths, Steyvers, & Tenenbaum, 2007). There is also evidence that this method is capable of uncovering changes over time in the contributions of generally interpretable topics occurring in natural text (e.g., Cohen Priva & Austerweil, 2015), including on Reddit (e.g., Thompson, Wojtowicz, & DeDeo, 2018). More important for the purposes of this article, the same literature suggests that topic models such as LDA can be used to distinguish between not only specific content, but also different framings of the same content: Cohen Priva and Austerweil showed that articles published in the peer-reviewed journal *Cognition* over the past few decades became more experimental and less theoretical, even when discussing the same areas of research, such as developmental or moral psychology. It is plausible that if protected values and consequentialism are important frames for discourse about same-sex marriage, this fact would be reflected in the topics that LDA provides: Either topics exclusively related to these framings

would be uncovered, or certain topics would reflect consistently more or less use of protected-values-based or consequentialist framing.

We used ratings from participants blind to LDA's representation of the corpus to determine the degree of association between each topic and a *consequentialist* or *protected-values-based* framing. We then tracked the contributions of these two framings to the corpus over time.³

The aim of unsupervised machine-learning methods such as LDA is to provide a generalized summary of data rather than optimized answers to specific questions. However, to demonstrate the usefulness of our method for the latter task, we compared the predictive and classification performance, as well as the interpretability of a linear model based on LDA's topics, with the performance of a similar, keyword-based model with many more parameters that did not benefit from a topic-based representation of the text.

Data

Using all comments on Reddit from January 2006 to September 2017,⁴ we created a corpus of comments matching a regular expression containing negatively or positively valenced words and phrases related to same-sex marriage (see Appendix 1). Quotes and hierarchical dependencies between the comments were ignored.

To improve the quality of our topic model, we applied common preprocessing techniques to the dataset: We changed all comments in our corpus to lowercase, to avoid different cases of the same word being treated as different words, and changed different grammatical forms of the same words to a uniform lemma (a process called *lemmatization*). HTML escape codes, stop words (i.e., common function words that do not distinguish between different contents), extremely rare words (words that appeared in fewer than five documents), and ubiquitous terms (words that appeared in 99% of the documents), as well as nonalphanumeric characters, were removed from the dataset. We used the lemmatizer and set of stop words from the Natural Language Toolkit (NLTK; Bird, Klein, & Loper, 2009), but retained around 20 words with possible relevance to protected-values-based or consequentialist framing (see Appendix 1). These steps were meant to reduce noise in the representation of the gist of each post. Our corpus contained 603,282 comments (15,711,221 words in total, comprising 32,925 unique words) with a mean length of 26 words (median = 20, $SD = 24.99$).

Figure 1 shows the number of comments in our dataset from each year. Most of the comments came from later years,

even though the percentage of comments on the website that were associated with same-sex marriage decreased over time, reflecting the growing popularity of Reddit in recent years (see Appendix 1).

Latent Dirichlet allocation (LDA)

Latent Dirichlet allocation is a generative hierarchical Bayesian model for discrete data (Blei et al., 2003). The model is applied to a set of collections of discrete data (a set of *documents* composed of a set of *words*) to recover latent dimensions (*topics*) that reflect statistical regularities among the words in each document. The structure of an LDA model lends itself naturally to recovering generally interpretable recurrent themes in data that are organized into separable chunks, such as in natural language (Griffiths et al., 2007). For our purposes, each Reddit comment was considered a document.

An LDA model is uniquely characterized by a set of words w , a set of n topics z , and the hyperparameters α and η , which determine the granularity of topics; n , α , and η are free parameters. LDA models the generative process of each document as follows: First, a multinomial distribution $\theta \sim \text{Dirichlet}(\alpha)$ over the topic indices $\{k : k \in [1, n]\}$ is defined. To determine each word $w_i \in w$ in the document, a topic index j is drawn from θ . Finally, w_i is drawn from the probability distribution defined by z_j (Blei et al., 2003). In short, each document is characterized as a probability distribution over topics, and each topic is characterized as a probability distribution over words. The estimated model is a hierarchical probability distribution fit to the training corpus (word order is ignored).

Training

We use a freely available and open-source implementation of an online variational Bayes algorithm (Hoffman et al., 2010) to train an LDA model on our corpus and to query the resultant model (Řehůřek & Sojka, 2010). The estimation procedure for topics is beyond the scope of this article; for more information, see Blei et al. (2003) and Hoffman et al. Ninety-nine percent of the documents (randomly chosen) were used to train the model over 1,000 iterations. The rest of the dataset was used as an evaluation set to ensure that the latter set had no more uncertainty than the training set, as a means of preventing overfitting (see Appendix 3).

Choice of hyperparameters

Following Cohen Priva and Austerweil (2015), α and η were set to .1, which encourages the model to represent each document as composed of only a few topics and to assign high probability to only a few words for a certain topic. The lower bound on per-term topic probability for inclusion in analyses was set to .01.

³ The Python and R code used in this study, as well as the learned models and the results of all analyses, are openly available at https://github.com/BabakHemmatian/Gay_Marriage_Corpus_Study.

⁴ Access was provided by <http://files.pushshift.io/reddit/comments>.

Fig. 1 Numbers of posts on Reddit per year that were relevant to same-sex marriage, as determined by our regular expression (see Appendix 1)

To determine the number of topics that offered the most intuitively interpretable representation of the corpus, we trained models with up to 100 topics in increments of 25. For each model, we looked at a combination of quantitative indicators of the model’s predictive capacity and the qualitative, semantic content of the topics.

Quantitative indicators We looked at the per-word perplexity of each model and the rate at which it changed with different values of n for our evaluation set (Zhao et al., 2015). *Per-word perplexity* reflects how uncertain the model is on average when predicting each word in a document, given the other words in the document. We also calculated UMass coherence values for all models (Mimno, Wallach, Talley, Leenders, & McCallum, 2011). *UMass coherence* measures how much, within the words used to describe a topic, a common word is a good predictor, on average, for a less common word. All measures showed a preference for fewer topics (see Appendix 3 for the values).

Semantic content Although per-word perplexity and UMass coherence provide an approximation to the interpretability of topics, they often do not align well with humans’ internal semantic spaces (the gold standard of coherence; Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009; Stevens, Kegelmeier, Andrzejewski, & Buttler, 2012). We therefore also manually inspected the 80 words most strongly associated with each topic in every model. We made note of whether the vast majority of the words for each topic intuitively belonged to the same conceptual category (e.g., “finance” or “religion”) and whether most of them could presumably be used in making the same kind of appeal for or against same-

sex marriage. This revealed the topics with $n = 25$ to mainly consist of mixtures of disparate arguments and framings. We therefore will report the results with 50 topics, because this value led to the optimal trade-off between, on the one hand, the interpretability and distinguishability of topics, and on the other, the predictive power of the model. Similar inspections for $n = 45$ and $n = 55$ suggested that our results are stable for similar numbers of topics and not an artifact of the exact value used (for all the models with various numbers of topics, the set of most representative words and their probabilities under the relevant topics can be found in the study’s online repository). The choice for n topics was made before the rest of the analysis presented in this article was conducted.

Categorizing topics as consequentialist versus protected-values-based

To determine which topics in the learned model were representative of consequentialist or protected-values-based discourse, 2,000 posts were selected from the corpus that had the greatest impact on the Reddit discourse, as proxied by the absolute difference between the numbers of up- and downvotes.⁵ The differences in the numbers of upvotes and downvotes in this set ranged from – 609 to 10,343, with an average difference of 543. Eight hundred of the 2,000 posts were chosen from our corpus such that, to the extent possible, the comments most representative of each topic were evenly sampled (henceforth called *impactful posts*). A comment’s

⁵ Because of “fuzzing” algorithms employed by Reddit to combat the effect of bots on upvotes and downvotes, it is impossible to recover the absolute number of votes from the freely available data. Only the difference between the number of upvotes and downvotes can be accurately determined.

representativeness of a topic z_j was operationalized as the percentage of words in the comment for which the topic z_j was the most probable topic according to the LDA model. The sampled comments were on average 25% representative of their most likely topic. One topic out of the 50 was not the most likely topic for any comment in the eventual sample (see Appendix 2 for the distribution of topic assignments among the 2,000 comments, as well as among the rated subset).

Ratings for the resulting set of 800 posts were gathered from six trained participants (three men and three women) blind to the predictions of the model. The average age in the sample was 34 years ($SD = 17$, ranging from 21 to 65). All of the participants were supportive of same-sex marriage. The participants read instructions for distinguishing between protected-values-based and consequentialist discourse based on Baron and Spranca (1997). The full text of the instructions can be found in Appendix 2. The full set of impactful comments, their topic estimates, and the participant ratings for them can be found in the study's online repository.

After comprehension check questions, participants were shown comments in randomized order and first asked to rate whether the attitude expressed in the post was in favor of same-sex marriage or against it, or whether it was impossible to tell from the content. They then rated the degree to which the post showed a consequentialist or protected-values-based framing of same-sex marriage, on a scale from 1 (*completely protected-values-based*) to 7 (*completely consequentialist*), where 4 was labeled *neither*. Five participants each rated 120 posts, and one participant rated 360 posts. To determine the reliability of the ratings, 20% of the posts were rated by two different raters.

The contributions of different topics to each rated impactful comment were queried from the LDA model and entered into a linear regression for training sets composed of 90% of the 800 comments. Parameter estimates from the regression were then used to predict ratings on the remaining 10%. Predictions were derived after dropping topics from the regression formula that were not significant predictors of ratings in the training set using a stepwise search. To ensure robust results despite the small size of our sample, we performed 10 ten-fold cross-validations, testing the model's performance on 100 different test sets in total: For each ten-fold cross-validation, we randomly divided the 800 comments into 10 mutually exclusive subsets, deriving the predictions for each subset based on the other nine, and aggregating the results across all ten test subsets. This entire process was then repeated 10 times, and the results were aggregated across all 10 ten-fold cross-validations.

We chose topics for tracking over time that were significant predictors ($p < .05$) in the linear model for at least half of the 100 iterations. These topics were classified as either protected-values-based or consequentialist, on the basis of the sign of

their associated parameter estimates averaged across all iterations.

To characterize the theme represented by each topic, we randomly chose 10 posts from the set of 25 posts in the corpus that were most representative of each top topic. Examples of the sampled comments and their associated representativeness values can be found in Appendix 4. All sampled comments can be found in the study's online repository. We examined the comments for use of similar categories of concepts (e.g., religious dogma), as well as for arguments with appeals similar in form and content (e.g., causal analysis involving changes to the structure of mainstream families). We also queried the 40 words that had the highest probabilities under that topic. We then identified words that were only present in the high-probability set of that top topic (*unique top words*) or that were shared with the high-probability set of only one other top topic (*almost unique top words*). See Appendix 4 for the resulting sets. We then examined the sets for properties shared among the vast majority of unique or almost unique top words belonging to each topic, noting whether they belonged to similar categories of concepts or could be used in arguments similar in style or content for or against same-sex marriage.

Calculating topic contributions

To determine the relative popularity of different topics over time, we calculated the monthly contribution of topic z_j (for $j \in [1, n]$) to the learned model. Following Cohen Priva and Austerweil (2015), this measure was defined as

$$p_m(z_j) = \frac{1}{|D_m|} \sum_{d \in D_m} \frac{|\{w_i \in d : \text{topic}(w_i) = z_j\}|}{|d|}, \quad (1)$$

where w_i is a word in document d ; d is a document, represented as an unordered bag of words in D_m (the set of documents from month m); and $\text{topic}(w_i)$ stands for the most likely topic for w_i given the prior distribution over topics and the other words present in d . This measure reflects the percentage of words in a month that are most strongly associated with a certain topic. In calculating the norm of documents, only words were counted for which the conditional probability of at least one topic was more than .01.⁶

Other than tracking the contributions of protected-values-based and consequentialist topics over time, this measure was

⁶ It is possible that for a significant portion of the dataset, similar probabilities could be associated with several topics, and by choosing only the most likely topic we would lose information with significant impact on the trends. The choice of hyperparameters discourages this. However, to ensure the robustness of the trends, we used an alternative calculation in which the contribution of a given topic was determined by the normalized sum of the probabilities assigned to the words of a document, conditioned on that topic and the document's context. Using this method, any probability associated with a topic counts toward its overall contribution. The results did not deviate from what is reported in the text.

used to identify topics that were major contributors to the discourse even though they were not consistently associated with either category. The combination of such topics and topics chosen on the basis of predictive value with respect to the ratings of impactful comments comprises a set that we will henceforth call *top topics*. We considered only these topics in the results reported below.

Comparison with a word-frequency model

Topic models such as LDA posit latent variables that make it possible to learn dependencies in text that may not be uncovered using a nonhierarchical representation of word meanings (Griffiths et al., 2007). To determine whether the LDA model captures more of the variance in the human ratings of consequentialist versus protected-values-based reasoning, we compared the test set prediction performance of the LDA-based regression model with that of a regression model that used the most discriminative words as predictors, defined according to the word-frequency model described below.

Hierarchical models of language can also provide more interpretable dimensions of variation (Griffiths et al., 2007). We therefore will also discuss the predictors included in the two regressions in terms of interpretability.

The word-frequency model used the conditional probability of each word w_{ij} (the j^{th} word of the i^{th} comment) in a training set of impactful comments rated as consequentialist [$P(w_{ij} | \text{consequentialist})$] or protected-values-based [$P(w_{ij} | \text{protected-values-based})$]. The informativeness of each word with respect to the distinction between the two frames is defined as

$$I(w_{ij}) = \log \frac{P(w_{ij} | \text{consequentialist})}{P(w_{ij} | \text{protected-values-based})}, \quad (2)$$

where greater values mean a stronger association with consequentialist framing. One hundred training and test sets were produced with 10 ten-fold cross-validations, using a procedure similar to that of the LDA predictive model described above. The informativeness values for each word were summed across all iterations to calculate each word's overall association with the two frames of discourse. One hundred words with the highest aggregate values and 100 words with the lowest aggregate values were chosen as the best word-level predictors (see Appendix 3 for a list of these predictors).

A linear regression model was learned for each iteration following the same procedure as we described for LDA, but with these 200 words as potential predictors instead of topic contributions. Parameter estimates based on each training set were then used to predict ratings of the comments in the associated test set. Performance was averaged across all 100 iterations and compared with that of LDA.

The presence of the following words was a significant predictor of consequentialist rating of an impactful comment in more than half of the 100 test sets (in decreasing order of number of times; the numbers range from 100, for “used,” to 51, for “barometric”): “used,” “ignored,” “claim,” “fiscal,” “destroy,” “effect,” “zero,” “morning,” “kill,” “forcing,” “work,” “aggressive,” “tried,” “none,” “gain,” and “barometric.” The presence of the following words was a significant predictor of protected-values-based rating of an impactful comment in more than half of the 100 test sets (in decreasing order of number of times; the numbers range from 97, for “devil” to 65, for “directly”): “devil,” “clear,” “stupid,” “believe,” “clapped,” “within,” “holy,” “favor,” “weekend,” “wanting,” “becomes,” “Christian,” “currently,” “away,” “different,” “terrible,” and “directly.” Although some of these words carry connotations associated with the relevant discourse category (such as “fiscal” for consequentialist or “holy” for protected values), others are less intuitive indicators of their assigned discourse category (such as “kill” for consequentialist or “currently” for protected-values-based).

Results

Human ratings of the impactful posts

The average rating for the association of comments with protected-values-based and consequentialist discourse was 4.27 (slightly more consequentialist than not; $SD = 2.09$). On the basis of these ratings, out of 800 impactful posts, 302 (37.8%) were categorized as consequentialist, 237 (29.6%) were categorized as protected-values-based, and the remainder were rated as neither (261 comments, comprising 32.6% of the dataset).⁷

A total of 160 comments were rated more than once. To measure the reliability of ratings, we randomly assigned one of the two ratings for each comment to one of two sets and calculated the Pearson correlation between the resulting sets. This random assignment was repeated 100 times. The resulting correlations between different ratings of the same comments ranged from .18 to .19. These correlations suggest less agreement among the raters than we anticipated, though they are consistently positive. The difficulty that participants had with classifying posts as consequentialist or protected-values-based is in line with what has been found in other studies (e.g., Fernbach et al., 2018).

Of the rated comments, 554 (69.3%) were identified as pro-same-sex marriage, 33 as clearly against same-sex marriage

⁷ To determine classification of comments for which more than one rating was available, we classified them as “neither” if two raters provided ratings on the opposite halves of the scale, and otherwise used the mean.

(4.1%), and the remaining 213 (26.6%) had unclear ramifications for the issue. The strong bias toward pro-same-sex marriage arguments might reflect the population of Reddit users, moderation and acceptable use rules, or the coincidence of increased public support for same-sex marriage with the increased popularity of Reddit and is a limitation of our study. It is also noteworthy that the vast majority of impactful comments had more upvotes than downvotes, suggesting that Reddit users are more likely to react to content they approve of. This may have introduced further bias toward opinions in favor of same-sex marriage in our sample of most impactful comments.

Prediction of human ratings using LDA and the word-frequency model The average Pearson correlation between the predictions of a linear regression based on the LDA topics and the true ratings of held-out test sets was .25 across 10 ten-fold cross-validations. Including random intercepts to account for the rating styles of different raters increased this correlation to .4. The average adjusted *R*-squared for the former model was .13, whereas the same value for the latter model was .34 (see the study's online repository for the results of all regression analyses).

Comments for which the predicted rating was greater than 4 were considered to be classified as consequentialist, and predictions of less than 4 were considered protected-values-based. The accuracy of these classifications across the 100 tests sets was .64. A model that classified every comment as consequentialist would have had an accuracy of .56. The word-frequency model had comparable predictive and classification capabilities, with an average correlation of .37 and a classification accuracy of .64. The adjusted *R*-squared averaged across iterations was .27 for the word-frequency model.

The amount of information contained in the human ratings, and the amount thus captured by the models, is modest. Our results suggest that the LDA model captures as much information about the human intuitions surrounding consequentialist versus protected-values-based reasoning as a word-frequency model with four times the number of parameters.

Breakdown of discourse

See Table 1 for the top five high-probability words associated with each of the *top topics*: topics that were either (a) significant predictors of classification of impactful comments in more than half of all cross-validations or (b) major contributors to the discourse based on the LDA model. The probabilities of the words in this table, given the associated top topic, range from .01 to .25 ($M = .035$, $SD = .036$). The variability in probabilities partly reflects differences in breadth among topics. The probability of any word in this table being associated with any other topic in the model (including nonmajor topics) is less than .056. Only unique high-probability words

are shown, because certain words, such as “gay” or “marriage,” have high probability under all topics and thus do not capture the topic's focus. See Appendix 4 for the full sets of unique and almost unique high-probability words for all mentioned topics. For the 80 words most associated with each of the 50 topics in the model, and their probabilities, see the [supplemental materials](#).⁸

Protected-values-based topics Significant contributions from three topics were significant predictors of a post being rated as more protected-values-based in all 100 cross-validations. Inspection of the unique and almost unique high-probability words and the comments most representative of these topics revealed the following prominent themes: *religious arguments*, *freedom of belief*, and *LGBT rights*. The presence of no other topic was a significant predictor of protected-values-based reasoning in more than half of the cross-validations.

Consequentialist topics Significant contributions from five topics were significant predictors of a post being rated as more consequentialist in more than half of all cross-validations. Inspection of the most representative words and comments revealed that these topics covered a greater range of themes than the protected-values-based topics. The topics were most representative of the following themes: *politicians' stance*, *children of same-sex parents*, *same-sex marriage as a policy issue*, *employer attitude and regulations*, and *cultural and historical status*.

The last two topics had a noticeably broader focus. Their contribution was more distributed across different posts, and their most representative comments were more diverse in content (see Appendix 4 and the representative comments in the online repository). These two topics were also highly associated with several words that did not intuitively cohere with the majority of elements in the same sets. Words such as “evil” and “value” are among the top words for these two topics. Although these words appear to be more associated with protected values, inspection of the impactful and representative comments shows the main focus to be causal analysis of cultural and financial trends related to same-sex marriage. This ability to separate the overall association of such terms with a discourse category from their association in a specific context is one of the benefits of using hierarchical models such as LDA and is absent in simple keyword-based methods (Griffiths et al., 2007).

Major topics rated as neither consequentialist nor protected-values-based For each topic, we calculated the topic's average

⁸ The supplemental materials can also be found at https://github.com/BabakHemmatian/Gay_Marriage_Corpus_Study/blob/master/Supplemental_Materials.pdf.

Table 1 Top five words that are uniquely associated with each top topic, in decreasing order of probability given that topic

Topic	Top five most associated unique words
Protected-values-based topics	
Religious arguments	God, Bible, Islam, Jesus, Christianity
Freedom of belief	Belief, opinion, others, atheist, respect
LGBT rights	LGBT community [*] , trans, movement, right, fight
Consequentialist topics	
Employer attitude and regulations	Job, Sanders, evil, owner, trade
Cultural and historical status	Traditional, modern, Tory, value, destroy
Politicians' stance	Anti, supported, changed, pro-gay, politically
Children of same-sex parents	Child, parent, divorce, adoption, adopt
Same-sex marriage as a policy issue	Political, Left, important, climate, immigration
Neither	
Forcing versus allowing behaviors	Allowed, legally, fine, anybody
Personal anecdotes	Woman, man, lesbian, wife, husband

* The probabilities conditional on the associated topic for “LGBT” and “community” were almost identical, suggesting that these words appeared in conjunction

monthly contribution to the model—that is, the fraction of words in the corpus from a certain month generated by that topic. We then compared that value to the contribution that would have been expected if the topic contributions were uniformly distributed at each time point. Two topics were not consistently associated with consequentialist or protected-values-based discourse but were nevertheless major contributors to the discourse based on this criterion, contributing on average more than twice the uniform baseline. The themes represented by these two major contributors were *forcing versus allowing behaviors* and *personal anecdotes*. The former topic has fewer unique associated words than other topics in Table 1, because words such as “force” and “let” that are most strongly associated with it are common terms also associated (albeit less strongly) with other top topics. Similarly, the word “right” had by far the highest probability under *LGBT rights* but was shared as a high-probability word with other topics, due to its many meanings.

Intertemporal trends

To visualize how the contributions of protected-values-based and consequentialist top topics to the discourse changed over time, we first combined monthly contribution estimates of the topics associated with the two discourse categories. We then visualized the trends of these pooled estimates as a function of time using degree-two locally estimated scatterplot smoothing (LOESS) regressions. We used LOESS to highlight the local perturbations in contributions that characterized the impact of specific events. Although the LDA was trained on the entire dataset, the monthly contributions from 2006 and 2007 were excluded from our regression analyses because each estimate was based on fewer than 100 documents. Figure 2 shows the results for the two discourse categories. The shaded areas

represent 95% confidence intervals. Note that the combination of the two sets of categories accounts for at most about 20% of our corpus at each point in time.

Protected-values-based themes rise in prominence until mid-2012, after which they show a steady decline that predates the Supreme Court ruling in June 2015 (rightmost vertical line). The decline begins not long after the point at which a majority of Americans expressed support for same-sex marriage (leftmost vertical line), which was followed by a number of prominent politicians expressing support for the issue (Gallup, 2017).⁹ In contrast, the contribution of consequentialist topics decreases overall until 2010, despite significant monthly variability during this period. One spike in consequentialist discourse happens around the time of majority support for same-sex marriage, whereas a larger spike can be seen in 2016, presumably associated with the United States' presidential election.

These results are based on the pooled contributions of several topics. To see whether our results generalized to individual topics, we ran a separate regression to predict the logarithmic odds of each topic z_j 's contribution, $\log \frac{P(z_j|D)}{P(\neg z_j|D)}$, where D is the total observed text in the corpus at a given time step, and $\neg z_j$ stands for all other topics, as a function of time. The model accounted for 49% of the variance in contributions. Consequentialist topics and higher time indicators were both associated with a significantly higher contribution ($p < .001$).

⁹ Gallup first reported that more than 50% of Americans supported same-sex marriage in May 2011 (Gallup, 2017). However, the percentage dropped from 53% to 48% in November of the same year (with a 95% confidence interval of 4%). Starting in May 2012, all polls showed a steadily increasing mean approval rate of 50% or more. Note that one of the cases that culminated in the Supreme Court ruling to legalize same-sex marriage was filed in the same year (*DeBoer v. Snyder*, in January 2012).

Fig. 2 Points reflect the proportion monthly contributions of consequentialist and protected-values-based topics to the trained LDA model. Each data point was calculated by summing the monthly contributions of all topics subsumed by the relevant discourse category. The colored lines show the best-fitting locally estimated scatterplot smoothing (LOESS) function (see the text for details). The shaded areas represent

95% confidence intervals. The vertical line on the left marks the beginning of a steady upward trend of support for same-sex marriage among a majority of Americans (Gallup, 2017). The vertical line on the right marks the Supreme Court ruling that legalized same-sex marriage in the United States

The linear and quadratic interaction between these two predictors was positive and highly significant ($p < .001$). In other words, individual consequentialist topics are associated, on average, with increasing contributions over time.

We then visualized the trend of each top topic's contribution to the LDA model as a function of time using LOESS regression. Figure 3 shows the results for protected-values-based topics. *Religious arguments* and *freedom of belief* follow each other closely: They rise in prominence prior to 2013 and see a decrease in importance afterward. The contribution of *LGBT rights* is relatively constant until the Supreme Court ruling and shows a spike in 2016.

To test whether the trends associated with each topic's contribution over time were statistically significant, we ran separate cubic regressions with each topic's contribution as the predicted variable and time indicators as the predictors. This analysis showed the linear and quadratic trends for *religious arguments* and *freedom of belief* to be significant ($p < .001$). The upward linear trend in *LGBT rights* was also significant ($p < .001$).

Figure 4 shows the results of LOESS regression for consequentialist topics. The contribution of *politicians' stance* to discourse about same-sex marriage stays relatively constant prior to 2011 and increases sharply for a period of two years afterward. We conjecture that this increase reflects the time when

many politicians started to voice their support publicly, including Barack Obama in May 2012. The second increase in discussions of *politicians' stance* starts in 2014 and might initially have related to a string of highly publicized court cases regarding LGBT rights that culminated in the Supreme Court decision. The prominence of LGBT issues in the presidential campaigns of 2016 might explain the higher probability of discussion about *politicians' stance* in 2016. The contribution of this topic decreases in 2017 but is still higher on average than in 2015. Discussion of *same-sex marriage as a policy issue* follows a similar pattern, but with smaller spikes. *Same-sex marriage as a policy issue* was also the only topic that did not show any significant cubic polynomial trend, which might be ill-equipped to represent its seemingly periodic pattern of contribution. The contribution of discussions about *children of same-sex parents* decreases slightly but significantly until 2016, whereas those of *employer attitude and regulations* and *cultural and historical status* to discourse are stable.

Figure 5 shows the results of a LOESS regression for the major topics categorized as neither protected-values-based nor consequentialist. *Forcing versus allowing behaviors* has the highest average monthly contribution to the model of any topic. It also shows the clearest downward trend, seemingly unperturbed by major events in this period concerning same-sex

Fig. 3 Points reflect the proportion monthly contributions of individual protected-values-based topics to the trained LDA model. The colored lines show the best-fitting LOESS function, and the shaded areas represent 95% confidence intervals. The black vertical line on the left marks

the beginning of a steady upward trend of support for same-sex marriage among a majority of Americans (Gallup, 2017). The vertical line on the right marks the Supreme Court ruling that legalized same-sex marriage

marriage. The significance of this trend was confirmed via cubic regression ($p < .001$). In contrast, the contribution of *personal anecdotes* to the model increases rapidly in earlier years and more slowly afterward. This trend was also highly significant on the basis of cubic regression ($p < .001$). We conjecture that the latter pattern reflects the increased tendency of individuals to share their experiences or those of their acquaintances with higher societal acceptance of same-sex relationships.

The individual temporal trends and the most strongly associated keywords for the 40 topics not included in Figs. 3–5 can be found in the [supplemental materials](#). The combined average contributions of all those topics to the model did not vary significantly over the timeframe of interest.

Intertemporal trends of the word-frequency model

Figure 6 shows the fractions of words that the word-frequency model classified as consequentialist and protected-values-based in each month. This measure shows a relatively constant advantage for consequentialist discourse throughout the years, which reflects the higher base rate of consequentialist comments in the set of impactful comments. Unlike LDA, the word frequency model is a flat semantic representation that does not posit latent themes, but effectively averages across more nuanced and temporally variant threads of discourse.

This averaging may reconcile the flatness of the trends in word-based classification with the temporal variation discussed in the previous sections.

Considerations in interpretation of the results

Prediction and classification accuracy Although reliably above chance, the accuracy of classification and rating predictions based on both our word-frequency model and the LDA analysis was slightly below two-thirds. The word-based model might have performed better if it had a way to represent the hierarchical nature of language. LDA might have performed better if it were not an unsupervised method that is not optimized for providing such a classification, but rather for offering a summary representation of the corpus. Additionally, much of the information about the discourse content was not present in the individual comments, but was carried by other elements of the discourse context. More recent topic-modeling methods that allow for the inclusion of hierarchical relations between different posts could help extract aspects of the relevant context (e.g., matrix inter-joint factorization; Nugroho, Zhong, Yang, Paris, & Nepal, 2015). However, implied inferences that are not referred to directly in the text would escape most common natural language processing models (Manning, Manning, & Schütze, 1999).

Fig. 4 Points reflect the proportion monthly contributions of individual consequentialist topics to the trained LDA model, separated by color. SSM stands for same-sex marriage. The colored lines show the best-fitting LOESS function, and the shaded areas represent 95% confidence intervals. The black vertical line on the left marks the beginning of a

steady upward trend of support for same-sex marriage among Americans since the first time a Gallup poll showed a majority of Americans in favor of it (Gallup, 2017). The vertical line on the right marks the Supreme Court ruling that legalized same-sex marriage

Coverage of discourse We recognize that the distinction between consequentialist and protected-values-based discourse accounted for only about 20% of our corpus. Although more robust ratings might increase the fraction considerably, we acknowledge that same-sex marriage is a complex topic that can be approached from a variety of standpoints, not all of them characterized by our focal distinction.

Data limitation The quality of an algorithm's outputs is largely a function of its inputs. Most of the posts in our dataset belonged to more recent years. Therefore, the trained model was biased in favor of the statistical properties of discourse in those years. In addition, Reddit posts appear to be strongly biased toward pro-same-sex marriage opinions. According to ratings by the authors, the rate of positive attitudes among the most representative posts we sampled for the 10 top topics was higher than the reported national average (Gallup, 2017). A similar pattern was observed for impactful posts rated by participants blind to our hypothesis.

Properties of LDA The distributions representative of top topics partly reflect parameter choices, such as the number

and granularity of topics. Although we based our choice of parameters on careful analysis of the corpus, better values might exist. We also note that, due to either the strictness of the thresholds that determined inclusion or the nature of discourse on Reddit, our analysis may have excluded some topics relevant to the debate over same-sex marriage.

More generally, the inherent shortcomings of LDA may have affected our results. LDA's ability to pick up interpretable themes in documents emerges from the correspondence of interpretability with the statistical properties of the "bag-of-words" representations of these documents. Such representations lose important aspects of language, including sequential dependencies between words. We urge readers to follow the advice of Blei et al. (2003) and to treat the epistemological claims we make about the top topics with the usual degree of caution. Similarly, the simplicity of our word-frequency model may have concealed interesting patterns in the data. For instance, our word-frequency model did not make use of the base rates of words.

Fig. 5 Points reflect the proportion monthly contributions of major topics that were categorized as neither protected-values-based nor consequentialist to the trained LDA model, separated by color. The colored lines show the best-fitting LOESS function, and the shaded areas represent

95% confidence intervals. The black vertical line on the left marks the beginning of a steady upward trend of support for same-sex marriage among a majority of Americans (Gallup, 2017). The vertical line on the right marks the Supreme Court ruling that legalized same-sex marriage

Fig. 6 Fractions of words in each month classified as consequentialist or protected-values-based by the word-frequency model (see the text for details). The shaded regions indicate the 95% confidence regions, calculated on the basis of 100 instances of the word-frequency model, each trained on a different set of comments consisting of 90% of the labeled

impactful comments. The black vertical line on the left marks the beginning of a steady upward trend of support for same-sex marriage among a majority of Americans (Gallup, 2017). The vertical line on the right marks the Supreme Court ruling that legalized same-sex marriage

Discussion

We hypothesized that shifts in public attitudes surrounding same-sex marriage are accompanied by shifts in public discourse away from protected values and toward consequentialist rhetoric, and that this shift would be reflected in discourse on Reddit over the last decade. To address the question, we used LDA to infer the topics underlying posts appearing on Reddit from January 2006 to September 2017. We categorized major topics into the categories protected-values-based, consequentialist, and neither, based on human ratings and average contributions to the model, and examined the contributions of each category as well as the individual topics to the corpus over time. Many of the trends we observed coincided with turning points associated with the same-sex marriage debate, suggesting that the discourse on Reddit can track shifts on issues of major social interest.

Discussion of the political framing of same-sex marriage (e.g., the impact of *politicians' stance*) ebbed and flowed, following election cycles closely. However, the spikes grew larger over time, showing an overall increase in discussing political ramifications of this issue even after the Supreme Court ruling in favor of same-sex marriage in 2015. Discussion of two protected-values-based topics (*freedom of belief* and *religious arguments*) increased sharply prior to majority support. These trends reversed in the latter half of 2012, when several high-profile court cases related to same-sex marriage were filed, but long before the Supreme Court ruled in favor of equal marriage. Although these topics dominated the overall pattern of consequentialist and protected-values-based discourse, not every topic in one of these two categories changed in contribution considerably during the same period: Among protected values, *LGBT rights* gained some limited prominence in 2016 but was otherwise unaffected by the passage of time. The contribution of children's welfare to the discussion decreased prior to consistent majority support in the US for same-sex marriage (Gallup, 2017) and flattened afterward. Discussion of other major consequentialist topics was flat during the period we studied.

Other significant contributors to the discourse not reliably associated with either category were LGBT-themed anecdotes and discussion of forcing or allowing certain behaviors. The former category increased in contribution monotonically, presumably due to increasing acceptance of LGBT experiences in society at large, while forcing or allowing behaviors consistently decreased in influence.

Our results are correlational and do not speak to the causal relation between changes in discourse and changes in attitudes. They only show that for this one social issue, a relative shift from a certain type of protected-values-based discourse to a certain consequentialist framing on a popular online platform coincided with a shift in public attitude in more recent years. A natural follow-up to our work would be to track to what extent shifts in protected-values-based and consequentialist rhetoric trace the public attitudes associated with other major social

topics (e.g., marijuana legalization). Regardless, the timing of the shift in discussion of same-sex marriage hints at a causal relation that is in the opposite direction from what we initially hypothesized. The data we report suggest that the trajectory of discourse on this important social issue began with a debate about protected values: Should the issue be framed in relation to protected notions of marriage or in terms of freedom of opinion and beliefs? Once attitudes began to change, the discourse changed, too. Even though protected values remained major contributors to the discourse, the debate shifted to be less protected-values-oriented, and the relative concentration of consequentialist discourse—in particular, the political and policy ramifications of the issue—increased.

That the shift coincided with the development of majority support for same-sex marriage could have occurred for several reasons. Individuals might have voiced their support for values they believed were popular. As opinions shift, the tendency to provide arguments for the less popular viewpoints could decrease. With increased public support, hopes for achieving concrete outcomes related to same-sex marriage may have been rekindled, as well, spurring talk of political processes, which itself may have impacted attitudes.

Alternatively, people may be likely to rely on the consensus within their communities to determine their protected values and consequentialist beliefs (Sloman & Fernbach, 2017). There is strong evidence that people do not reason their way to many of their beliefs, but instead inherit those beliefs from the groups they affiliate with (including their families, religious and political communities, etc.; for an early modern discussion, see Hardwig, 1985). Majority support for same-sex marriage may have signaled a shift in these inherited beliefs, itself inducing further changes in discourse and attitudes.

So far, we have discussed protected-values-based and consequentialist framing of the discourse as if they are mutually exclusive. It is possible that protected values and consequentialist thinking fall on separate discursive dimensions rather than on the same continuum (cf. Tanner et al., 2008). Relatedly, some of the themes we discovered in our dataset, such as forcing or allowing behaviors and personal anecdotes, can and were used to appeal to both or neither of these dimensions. Without assuming a strict continuum between the two, we maintain that protected values and consequentialism characterize two important frames of public discourse.

Public discourse is a rugged and complex landscape. We have presented a survey of specific dimensions in that landscape that can be helpful for characterizing societal discourse, with the hope of contributing to an ongoing project of mapping its features more entirely. However, the fact that not all protected-values-based and consequentialist topics showed clear temporal trends highlights the fact that discourse about major social topics is impacted by many factors and is not reducible to a binary dimension. Indeed, only about 20% of the text in our corpus was classified along this dimension. A natural follow-up to our

work would be to attempt to identify the other dimensions that characterize online discourse on major social topics.

We tracked intertemporal trends in high-level themes, but our method does not have the resolution to allow us to delve into the fine-grained ways in which these themes are invoked. For example, although we can track discussion of religious arguments over time, we are unable to say how often these arguments are invoked sincerely and how often as a straw man. The different ways in which discussants invoke protected values and consequences could explain or predict different patterns of attitude shifts. For example, Atran et al. (2007) argued that acknowledging and making small concessions to the other side's protected values is an important and often necessary step in conflict resolution and the achievement of consensus. Brewer (2003) provided evidence of changes in protected values accompanying more positive attitudes toward the LGBT community. Although the increase in discussion of LGBT rights in 2016 and 2017 may reflect the results of such a process, our method may have failed to uncover other examples of similar shifts in the corpus.

From a methodological perspective, our results highlight how limited annotated data combined with unsupervised machine learning can help psychologists extract the properties of large corpora that speak to important facets of and hypotheses about individual and social cognition. The topic modeling of the corpus was independent of our hypothesis, and therefore free of biases introduced by training samples, and our annotated dataset was small as compared to those commonly used in supervised natural language processing. Given the ubiquity of social media in modern life, open access to many of these data, and easy access to packages for performing the associated analyses, such methods will allow researchers to use much larger and more naturalistic samples of human discourse to address psychological hypotheses with modest resources.

Even though LDA is not optimized to classify consequentialist and protected-values-based reasoning, its performance was comparable to a keyword-based approach optimized for that purpose through the use of many more parameters. The results of LDA were more interpretable with respect to a couple of major dimensions of human discourse. Many of the keywords associated with either category in the word-frequency model could not be readily associated with specific arguments. In contrast, LDA topics represented more interpretable clusters of discourse surrounding same-sex marriage. This is partly because the hierarchical structure of LDA allowed us to separate the association of terms with the two types of reasoning from their association to specific classes of arguments. For instance, the term “value” was a strong predictor in both models. However, it was among the 80 words most associated with both the discussion of causal historical processes, categorized by raters as consequentialist, and the *freedom of belief* topic, rated as predominantly protected-values-based. The word-frequency model simply associated this term with protected-values-based reasoning.

LDA also afforded clearer and more fine-grained, argument-specific temporal trends: While the presence of words associated with consequentialist or protected-values-based human ratings suggested a constant advantage for consequentialist discourse over time, a more detailed look at the topics underlying the use of certain words showed the ebb and flow of arguments over time. These trends corresponded to major social events, suggesting that LDA can be used to uncover how discourse reacts to influences that unfold over time. Future work could take advantage of the burgeoning collection of related methods (e.g., Esmaeili, Huang, Wallace, & van de Meent, 2019) to characterize the rugged landscape of social discourse.

Author note This article greatly benefited from discussion with Robert Thorstad and members of the Sloman Lab at Brown University. We thank Elinor Amit, Linda Covington, David Sherman, Leila Sloman, Semir Tatlidil, An Vo, and Luana Pessanha de Mattos for their help with the data gathering. Earlier versions of the results included in this article were presented at the Heterodox Psychology Workshop held at Chapman University in Orange, California, USA, in August 2018, and at the 39th Annual Conference of the Society for Judgment and Decision-Making in New Orleans, Louisiana, USA, in November 2018. This publication was made possible through a grant from the Intellectual Humility in Public Discourse Project at the University of Connecticut and the John Templeton Foundation.

Appendix 1: Corpus preprocessing

Regular expression used to select relevant comments

In creating the regular expression, WordNet synsets (Fellbaum, 1998) were used to determine words close in meaning to “gay” and “marriage.” Some of the terms include homophobic slurs, whereas some are not taboo. We then removed terms with obvious alternative meanings (such as “queen”) to reduce the amount of noise, resulting in the following regular expression:

```
(^(?=.*gay|.*\\bfag|.faggot|.fagot|.queer|.\\bhom-
o|.LGBT|.GLBT|.same.sex|.lesbian|.\\bdike|.\\b-
dyke|.sodom)(?=.*marry|.civil
union|.marri).*$)|(^(?=.*marriage equality|.equal
marriage).*$)
```

Set of words retained from the Natural Language Toolkit's list of stop words

Potentially protected-values-based should, shouldn't, should've, shan't, need, needn't, ought, must, mustn't

Potentially consequentialist how, can, could, couldn't, few, more, most, all, any, against, because, why

Fractions of posts on Reddit related to same-sex marriage

Fig. 7 Fractions of all Reddit posts that were relevant to same-sex marriage (vertical axis) as a function of time (horizontal axis)

Appendix 2: Ratings of impactful comments

Distribution of most likely topics among impactful posts

Among 2,000 most impactful posts Differences in the number of sampled posts may reflect differences in popularity, or the base rate of different topics. The topics identified as protected-values-based by raters (shown in Fig. 8) have IDs 12 (*religious arguments*), 48 (*freedom of belief*) and 49 (*LGBT rights*); topics identified as consequentialist are marked with IDs 4 (*employer attitude and regulations*), 14 (*cultural and historical status*), 22 (*politicians' stance*), 27 (*children of same-sex parents*), and 28 (*same-sex marriage as a policy issue*). The two major topics that did not belong to either category can be identified by IDs

16 (*forcing versus allowing behaviors*) and 33 (*personal anecdotes*). Topic 38, with the greatest number of sampled comments in the set, had a focus similar to Topic 33.

Among 800 rated impactful posts The IDs of topics (shown in Fig. 9) are similar to those in Fig. 8.

Instructions for ratings of impactful posts (emphasis in the original)

Thank you for your participation in this task. We are interested in better understanding the reasons behind people's attitudes for and against the legalization of same-sex marriage. On the next several pages, you will see a sample of comments from the popular social media platform Reddit. Your task is to read each comment

Fig. 8 Relative prominence of the 50 topics in the 2,000 most impactful comments. The x-axis shows the 50 topic indicators, and the y-axis shows the number of comments for which the topic had the highest contribution

Fig. 9 Relative prominence of the 50 topics in the 800 comments for which human ratings were elicited. The x-axis shows the 50 topic indicators, and the y-axis shows the number of comments for which the topic had the highest contribution

carefully, and then answer two questions: whether the comment is *pro* or *against* same-sex marriage, and *why* you think the commenter holds the attitude they do. In particular, we ask you to rate the comment on a scale from “completely protected-values-based” to “completely consequentialist.” Here’s what we mean:

Protected-values-based reasoning invokes values about actions that people think are definitely permissible or should be avoided, and are generally not open to compromise. For example, some people believe that marrying someone of the same sex violates a sacred value because marriage must be between a man and a woman. Others have a protected value that people have the right to marry the person they love. Note that while protected-values-based arguments may imply belief in certain concrete consequences (for example, that pursuing true love leads to happiness), a person engaging in such reasoning is unlikely to change their mind based on the consequences that result from certain actions. It is the action that is right or wrong regardless of its consequences.

Consequentialist reasoners hold positions because of the consequences they believe the policy will lead to. They aim for positive outcomes and want to avoid negative ones. For instance, an opponent may believe that legalizing same-sex marriage will hurt the development of children that may be put under the same-sex couple’s care. Proponents of same-sex marriage may believe that allowing same-sex marriage will allow same-sex couples to have access to better benefits, which will in turn improve their quality of life. Note that while these reasons may be rooted in certain values (e.g., the importance of children’s wellbeing), a person holding a consequentialist attitude would change their mind if they changed their mind about the consequences. For example, if the opponent was convinced that being raised by same-sex couples does *not* negatively impact children’s development, she might change her position.

Note that most arguments are neither completely protected-values-based nor completely consequentialist.

Example of an impactful post

“Why would you think that anyone wants to force churches to marry gay people? I’ve never heard anyone suggest that—ever.” (From January 2015; 1,877 more upvotes than downvotes, 54% representative of *forcing versus allowing behaviors*)

Appendix 3: Modeling procedures and analysis results

LDA model

Measures of predictiveness

Per-word perplexity Smaller values indicate that the model is better at predicting unseen words. Note that the per-word perplexity is lower for evaluation sets with 25 and 50 topics compared with the relevant training sets, providing no evidence of overfitting. The performance on the training set, which comprised most of the corpus, is comparable across different numbers of topics.

25 topics	50 topics	75 topics	100 topics
<i>Lower bound on per-word perplexity for training set:</i>			
164.95	178.35	188.54	197.47
<i>Lower bound on per-word perplexity for evaluation set:</i>			
15.13	64.32	211.51	545.59

Perplexity change rate (PCR) This measure was introduced by Zhao et al. (2015) and has been shown to outperform simple per-word perplexity in determining the optimal number of topics. The PCR is equal to the absolute difference

between the per-word perplexities of two models, divided by the absolute difference between the numbers of topics in the models. This measure also shows a preference for smaller numbers of topics (lower values are preferable):

25 topics	50 topics	75 topics	100 topics
0.61	5.28	5.89	13.36

UMass coherence This measure was calculated over the training set. Larger numbers indicate that the words associated with each topic in that model are more likely to co-occur.

25 topics	50 topics	75 topics	100 topics
– 2.95	– 3.74	– 4.05	– 4.31

Robustness of topic contribution calculation

To test for the robustness of the distribution of topic contributions to random initializations of the estimated parameters,¹⁰ we ran the topic contribution calculation 10 times and calculated the Jensen–Shannon divergence (JSD)¹¹ between the resulting distributions. The pairwise JSD between the distributions from all runs was effectively 0.

Word-frequency model

Most predictive words

These are listed in order of decreasing informativeness regarding the classification task, as defined in the text. Words in *italic* were significant predictors for the linear model in more than half of the iterations.

Most consequentialist words Democratic, voting, third, *tried*, pay, caused, fear, *claim*, *effect*, *aggressive*, ring, crime, decision, campus, character, *work*, blame, *favor*, registered, despite, *kill*, ruling, among, *destroy*, shift, air, shouldn't, somebody, built, venue, lower, *zero*, girlfriend, *fiscal*, likely, November, *used*, little, candidate, rate, adoption, men, benefit, mental, existing, body, site, Iran, career, constitution, hospital, tomorrow, logic, multiple, arguing, *ignored*, unmarried, cock, dollar, overturn, *currently*, administration, *gain*, gave, abolish, net, seek, quickly, using, neutrality, medical, entitled, *barometric*, ahead, hurricane, agenda, widespread, legislation, perfectly, homophobic, blue, sub, trade, banned, whose, awkward, potentially,

¹⁰ The gensim package's (Řehůřek & Sojka, 2010) implementation of the algorithm to estimate topic–word distributions, which we used to estimate each term of the topic contributions (see the Calculating Topic Contributions section in the main text), randomly initializes the model parameters.

¹¹ We used the naïve calculation of JSD implemented by the THOTH Python package, available at thoth-python.org.

Fox, certainly, feminist, *forcing*, *none*, reform, nominee, appointed, DADT, virginia, subject, 000, establishment

Most protected-values-based words *clear*, evidence, Jesus, *terrible*, 2012, *different*, 100, pray, ethnic, lil, culture, moral, month, *directly*, looking, opposition, fundamentalist, wearing, figure, gonna, met, God, felony, repeatable, Zeus, Estonian, trip, music, religion, thinking, unnatural, *within*, *stupid*, board, weird, ask, impose, rally, federally, write, shitty, standing, *weekend*, clerk, *wanting*, debated, licence, technically, faggot, map, ridiculous, knew, experience, required, Netherlands, pro, sin, xd, hispanic, unit, art, home, fuck, pot, *believe*, *Christian*, personal, *away*, treated, Catholic, *holy*, respect, focus, belief, money, brother, *clapped*, committing, silly, *morning*, *becomes*, discriminate, accepted, opposing, recent, thus, quiet, visit, defending, burn, Southern, sitting, defense, provide, mostly, Francis, Chinese, interviewed, attacking, *devil*

Log odds of discourse assuming a consequentialist frame based on the word-frequency model

Figure 10 shows the estimated log odds that the subset of the corpus from each month was generated by a consequentialist frame. In particular, each point is the sum $J(m)$ over every observed word in month m of the log odds that that word was generated under a consequentialist frame:

$$J(m) = \sum_{i=1}^N \sum_{j=1}^{|C_i|} I(w_{ij}), \quad (3)$$

where N is the number of comments from month m , C_i is the i^{th} comment from month m , the cardinality of C_i is defined as

Fig. 10 Log odds of the observed discourse in each month under the assumption of a persistent consequentialist frame (see the text for details). The blue line tracks the mean, and the shaded region indicates the 95% confidence region, calculated on the basis of 100 instances of the word-based model, each trained on a different set of comments consisting of 90% of the labeled impactful comments. The black vertical line on the left marks the beginning of a steady upward trend of support for same-sex marriage among a majority of Americans (Gallup, 2017). The vertical line on the right marks the Supreme Court ruling that legalized same-sex marriage

the number of words in C_i , and $I(w_{ij})$ is defined as in Eq. 2. The log odds are mostly greater than zero, showing a consistent advantage of consequentialist discourse over time (see Fig. 6 in the main text).

Appendix 4: Unique or almost unique high-probability words and example comments for the top topics

Unique high-probability words are words among a topic's list of 40 words with highest prior probability that were not found in similar lists for other top topics. Almost unique high-probability words were shared with only a single other top topic. The words are ordered on the basis of probability given the relevant topic. The two sets of high-probability words are followed by a representative post randomly chosen from the set of 25 posts in the corpus most representative of that topic. Representativeness is operationalized as the percentage of words in the comment for which the associated top topic is the most probable topic. The full sets of high-probability words for all topics can be found in the [supplemental materials](#). The full set of representative comments for top topics can be found in the study's online repository.

Protected-values-based topics

Religious arguments

Unique God, Bible, Islam, Jesus, Christianity, else, tell

Almost unique Christian, sin, believe, Muslim, wrong, everyone, anyone

Representative post (representativeness = .73) I understand that many people interpret the bible to say that God doesn't approve of gay sex, but when I interpret the bible or any other book that makes claims about what God does or doesn't like, I always ask myself what is the reason? If God tells us not to do something, he should have a very good reason for it, like don't murder, that makes sense because it harms people, so I can believe this command comes from God, no problem. But what could God possibly see wrong with two people of the same sex, who love each other, getting married? Who does this harm, you don't believe God gives us arbitrary rules do you?

Freedom of belief

Unique belief, opinion, others, atheist, based, respect

Almost unique believe, wrong, agree, choice, freedom, personal, moral, Christian, disagree, person

Representative post (representativeness = .69) I think religion is based on personal opinions. If you personally don't like gay marriage, abortion and evolution, that is your choice, you don't have to get gay married, have an abortion or believe in evolution. When you try to enforce others to believe what

you believe, that is based on your traditional understanding that people should follow your tradition which they will attribute to religion.

LGBT rights

Unique LGBT community, trans, movement, fight, fighting, plebiscite, conversion, wing, push, anti-LGBT, denying, activist, ya, deserve, overturning, proud, fire, liberalism, rainbow

Almost unique equal, support, away

Representative post (representativeness = .48)

Fundamentally, the right to form coequal legal relations ought to be available to homosexuals as a matter of Due Process. That is, we're getting confused about what a right is. A right is not susceptible to plebiscite. If it were, it would be superfluous qua right. The "people" of California, Arkansas, Florida, etc., can no more vote on the rights of homosexuals to marry than they can vote on the rights of Jews or Puerto Ricans, for example, to speak or to associate. The entire point of a right is to protect minorities from majority tyranny. These ballot measures, as a vehicle to attack an individual right, are extremely dangerous not only to the rights of homosexuals, but also to all other individual rights.

Consequentialist topics

Politicians' stance

Unique anti, supported, changed, pro-gay, politically, opposes, mind, considering, 90's, pro-choice

Almost unique support, Hillary, Clinton, Libertarian, position, said, politician, anti-gay, pretty, actually, campaign, came, Bill

Representative post (representativeness = .68) Either it was politically expedient for him to claim not to support gay marriage, or it was politically expedient for him to not support DOMA. I think his overall posture on gay rights while in office (DOTA, etc.) lends support to the former view.

Children of same-sex parents

Unique child, parent, divorce, adoption, adopt, sign, adult, baby, joke, rate, raise, home, son, raising, worried, parenthood, adopted

Almost unique better

Representative post (representativeness = .76) Because it affects society in general. We know an alcoholic abuse parent has an affect on his/her children. We also know that children who are raised in single parents homes are affected by that. Why wouldn't Gay marriage and the right for gays to adopt affect a child? And children grow up and they affect society.

Same-sex marriage as a matter of policy

Unique political, Left, important, climate, immigration, economic, major, far

Almost unique issue, social, policy, etc., party, politics, side

Representative post (representativeness = .68) People apathetic to politics will stay apathetic to politics. Especially when there's a good chance the issue doesn't make a difference to them. How will gay marriage affect the average heterosexual either way? There's no good reason to *not* be apathetic. Personally, I'm straight and I think the current setup of marriage is inequitable, but most don't.

Employer attitude and regulations

Unique job, Sanders, OP, evil, owner, trade, sell, employee, SSM, 2009, blow, EU, depending, argued, class, reach, store, donor, realise, career, category, guilty, lesser, 2003, fly, tolerate, buying, threw, 2011, assault, Rubio, oil

Almost unique protected

Representative post (representativeness = .27) And it is freedom from consequences if the government is your employer. Especially if it's political speech. If back in 1950 I said I support gay marriage and was fired from a government job for supporting gay rights vocally, I constitutionally could not be fired.

Cultural and historical status

Unique traditional, modern, Tory, value, destroy, Islamic, China, duty, legalisation, Hilary, attack, recognised, dunno, Baptist, cult, nationally, Adam, unit, century, civilization, wasn't, badly, Arab, monogamy, designed, household, aggressive, vague, reached, warrior

Representative post (representativeness = .32) I don't think gay "marriage" devalues anyone else's marriage. This isn't a values issue for me. I'm concerned for society and want to preserve and protect its traditional structure which is based on the nuclear family only. I don't feel that my marriage is under attack, but I do feel the society I grew up in is.

Major topics categorized as neither protected-values-based nor consequentialist

Personal anecdotes

Unique woman, man, men, lesbian, wife, husband, male, girl, female, bi, bisexual, another, always

Almost unique friend, dude, relationship, partner

Representative post (representativeness = .65) Yeah, I've noticed something similar to this. I haven't even been in a long-term relationship with a woman, and my situation is much different (I'm married to a man who allows me to enjoy being with women because he understands). I'm incredibly up front about

this situation when I'm talking to women that I'm interested in 'getting to know further' and they always ask "why don't you just tell your husband you're a lesbian then?" I've gotten that from a couple women and am always slightly irritated by it. I have to keep saying "it's not that I don't enjoy sex with my husband, it's that he's not a woman, and I like those too."

Forcing versus allowing behaviors

Unique allowed, legally, fine, nobody

Almost unique getting, able, shouldn't, forced, let, understand, anyone, happy

Representative post (representativeness = .71) Because it does no impact my life at all. If two straight people get married it does not affect my life, just as if two gay people go get married. So who am I to tell two consenting adults they can't do something that will not actually hurt anyone, but make them very happy.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Atran, S., Axelrod, R., & Davis, R. (2007). Sacred barriers to conflict resolution. *Science*, 317, 1039–1040.
- Baron, J., & Leshner, S. (2000). How serious are expressions of protected values? *Journal of Experimental Psychology: Applied*, 6, 183–194.
- Baron, J., & Spranca, M. (1997). Protected values. *Organizational Behavior and Human Decision Processes*, 70, 1–16.
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. Sebastopol, CA: O'Reilly Media.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Secaucus, NJ, USA: Springer.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Brewer, P. R. (2003). The shifting foundations of public opinion about gay rights. *Journal of Politics*, 65, 1208–1220.
- Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (Vol. 22, pp. 288–296). New York, NY: Curran Associates.
- Cohen Priva, U., & Austerweil, J. L. (2015). Analyzing the history of cognition using topic models. *Cognition*, 135, 4–9. <https://doi.org/10.1016/j.cognition.2014.11.006>
- Dehghani, M., Iliev, R., Sachdeva, S., Atran, S., Ginges, J., & Medin, D. (2009). Emerging sacred values: Iran's nuclear program. *Judgment and Decision making*, 4, 930–933.
- Duggan, M., & Smith, A. (2013). 6% of online adults are Reddit users. *Pew Internet & American Life Project*, 3, 1–10.
- Esmaili, B., Huang, H., Wallace, B. C., & van de Meent, J. W. (2019). *Structured representations for reviews: Aspect-based variational hidden factor models*. arXiv preprint. arXiv:1812.05035
- Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.
- Fembach, P. M., Min, L., & Sloman, S. A. (2018). Values-based and consequence-based policy attitudes (Working article).

- Fernbach, P. M., Rogers, T., Fox, C., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science*, 24, 939–946.
- Gallup. (2017). *US support for gay marriage edges to new high*. Retrieved from <http://news.gallup.com/poll/210566/support-gay-marriage-edges-new-high.aspx>
- Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, 114, 211–244. <https://doi.org/10.1037/0033-295X.114.2.211>
- Hardwig, J. (1985). Epistemic dependence. *Journal of Philosophy*, 82, 335–349.
- Hoffman, M. D., Blei, D. M., & Bach, F. (2010). Online learning for Latent Dirichlet Allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 23, pp. 856–864). New York, NY: Curran Associates.
- Kant, I. (1797). *The metaphysics of morals*. Cambridge, UK: Cambridge University Press.
- Lakoff, G. (2004). *Don't think of an elephant! Know your values and frame the debate*. White River Junction, VT: Chelsea Green.
- Manning, C. D., Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA, USA: MIT Press.
- Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 262–272). Stroudsburg, PA: Association for Computational Linguistics.
- Newport, F., & Dugan, A. (2017). Partisan differences growing on a number of issues (Editorial). Retrieved from <https://news.gallup.com/opinion/polling-matters/215210/partisan-differences-growing-number-issues.aspx>
- Nugroho, R., Zhong, Y., Yang, J., Paris, C., & Nepal, S. (2015). Matrix inter-joint factorization—A new approach for topic derivation in twitter. In *2015 IEEE International Congress on Big Data* (pp. 79–86). Piscataway, NJ: IEEE Press.
- Pew Research Center. (2017, June). Changing attitudes on gay marriage. Retrieved from <http://www.pewforum.org/fact-sheet/changing-attitudes-on-gay-marriage/>
- Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Retrieved from <http://is.muni.cz/publication/884893/en>
- Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, 26, 521–562.
- Sloman, S., & Fernbach, P. (2017). *The knowledge illusion: Why we never think alone*. New York, NY: Riverhead Books.
- Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttlar, D. (2012). Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 952–961). Stroudsburg, PA: Association for Computational Linguistics.
- Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., & 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* (pp. 613–624). Geneva, Switzerland: International World Wide Web Conferences Steering Committee.
- Tanner, C., Medin, D. L., & Iliev, R. (2008). Influence of deontological versus consequentialist orientations on act choices and framing effects: When principles are more important than consequences. *European Journal of Social Psychology*, 38, 757–769.
- Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in Cognitive Sciences*, 7, 320–324. [https://doi.org/10.1016/S1364-6613\(03\)00135-9](https://doi.org/10.1016/S1364-6613(03)00135-9)
- Thompson, W. H. W., Wojtowicz, Z., & DeDeo, S. (2018, December). Levy flights of the collective imagination. Retrieved December 28, 2018, from <https://arxiv.org/abs/1812.04013v1>
- Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics*, 16, S8.