Ideology as topic correlation structure: Inference method and example application to carbon tax public opinion

Journal:	*Environmental Data Science*
Manuscript ID	EDS-2023-0007.R1
Manuscript Type:	Application Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Puelma Touzel, Maximilian; Universite de Montreal, Computer Science and Operations Research Lachapelle, Erick; Universite de Montreal, Political Science
Keywords:	Machine Learning, Data Science, Policy < Climate Change, Artificial Intelligence, Public Opinion
Abstract:	Ideology is a system of ideas that helps people rationalize their beliefs and explain the world. As an object of study, ideology is amenable to a generative modelling approach within the text-as-data paradigm. Here, we characterize ideology as the relational structure among topics. To access this latent structure, we use the Structural Topic Model to infer topics and the weights with which individual opinions mix topics. We fit the model to survey data of open-ended responses of Canadians elaborating on their support of or opposition to a carbon tax. We demonstrate that the latent topic mixture, in contrast to the observed word usage, is an efficient representation for the machine classification of responses, allowing for near-perfect accuracy in discriminating opposition from support. We operationalize relational structure as the correlation structure of the topic mixture statistics over a surveyed group. We evaluate four principled low-order statistics with intuitive interpretations for ideology: specificity, variability, expressivity, and alignment. We find higher specificity and alignment, much less expressivity, and similar variability in the ideology behind oppose responses relative to support responses. We discuss the implications of the results for climate policy debates and of our approach to analyzing ideology for computational social science.

Environmental Data Science (2023), 4: 1–17
doi:10.1017/eds.2020.xx

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Ideology as topic correlation structure: Inference method and example application to carbon tax public opinion

Maximilian Puelma Touzel[1,2]* and Erick Lachapelle[3]

[1]Mila, Québec AI Institute, Montréal, Québec, Canada
[2]Department of Computer Science and Operations Research, Université de Montréal, Montréal, Québec, Canada
[3]Department of Political Science, Université de Montréal, Montréal, Québec, Canada
*Corresponding author. Email: max.puelmatouzel@gmail.com

Received: 31 January 2023; **Revised:** In revision; **Accepted:** Pending revision

Keywords: Machine Learning, Data Science, Climate Change Policy, Artificial Intelligence, Public Opinion

Abstract

Ideology is a system of ideas that helps people rationalize their beliefs and explain the world. As an object of study, ideology is amenable to a generative modelling approach within the text-as-data paradigm. Here, we characterize ideology as the relational structure among topics. To access this latent structure, we use the Structural Topic Model to infer topics and the weights with which individual opinions mix topics. We fit the model to survey data of open-ended responses of Canadians elaborating on their support of or opposition to a carbon tax. We demonstrate that the latent topic mixture, in contrast to the observed word usage, is an efficient representation for the machine classification of responses, allowing for near-perfect accuracy in discriminating opposition from support. We operationalize relational structure as the correlation structure of the topic mixture statistics over a surveyed group. We evaluate four principled low-order statistics with intuitive interpretations for ideology: specificity, variability, expressivity, and alignment. We find higher specificity and alignment, much less expressivity, and similar variability in the ideology behind oppose responses relative to support responses. We discuss the implications of the results for climate policy debates and of our approach to analyzing ideology for computational social science.

Impact Statement

What people think about climate change and climate policy strongly depends on their politics. To better understand the role played by ideology in shaping people's attitudes, we assess people's ideological thinking by analyzing the ways in which they explain their stance in open-ended survey responses. We develop new ways of measuring how topics are talked about together. Independent of what respondents discussed, we find people elaborating on opposition to the carbon tax did so in a more focussed and structured way than did those who support a carbon tax. This knowledge could be used to help design more effective climate policy and communications.

1. Introduction

Personal values and prior beliefs can have an out-sized effect in shaping how individuals interpret the world and how they respond when prompted to opine about specific aspects of it. One helpful construct for understanding this cognitive and collective process is ideology, a system of interconnected beliefs that is much discussed in the context of debates surrounding the role of political partisanship and the causes of political polarization Baldassarri and Gelman (2008). However, how to analyze ideology as a correlation structure on beliefs is a methodological challenge yet unresolved Kalmoe (2020). In this study, we propose an approach to inferring ideological reasoning from open-ended survey data, and ask to what extent topic correlations can reveal differences in the way ideological groups rationalize their positions on carbon tax policy. Given the well-documented anti-tax rhetoric of conservative political discourse (Lakoff, 2010), we expect ideologically-driven differences in the topic structures called to mind among supporters and opponents of carbon tax policy.

There are many key societal challenges to which methods that reveal ideology could contribute. A prime example is mitigating the worst effects of climate change. Climate change mitigation is in many countries a hotly contested political issue that is characterized by extensive ideological and partisan conflict (Birch, 2020; McCright and Dunlap, 2011; McCright et al., 2016). Addressing this challenge requires significant government intervention to shift energy production and consumption away from CO_2-emitting fossil fuels (McGlade and Ekins, 2015; Welsby et al., 2021). In this context, a price on carbon emissions is often proposed by economists as being simple, flexible, and easy to put on a rising schedule, encouraging emissions reductions across economic sectors at a low aggregate cost (Baranzini et al., 2000; Aldy and Stavins, 2012). Policy design has addressed public resistance to carbon tax policy measures via the targeted use of funds in the form of lump sum rebates to taxpayers (Klenert et al., 2018). Despite these mechanisms, however, a lack of public acceptability remains a major obstacle to the adoption of substantive carbon taxes across countries (Haites, 2018; Lachapelle, 2017; Rabe, 2018).

Public support for carbon pricing is typically structured by ideology, with conservatives on the right of the political spectrum generally in opposition, and those on the left generally supporting the policy (Drews and Van den Bergh, 2016). This structuring of carbon tax policy attitudes by ideology is evident even in countries that have implemented tax-and-rebate schemes designed to build public support by providing rational, monetary incentives that offset policy costs. For instance, at current pricing levels in Canada (one of few countries that have implemented a tax-and-rebate scheme), 8 of 10 households are estimated to receive more than they pay as a result of the policy, and yet roughly 8 out 10 surveyed conservatives oppose the policy (Mildenberger et al., 2022). Mildenberger et al. (2022) probe that result using interventions and find conservatives underestimate the size of the rebate they receive. Such behaviour is not necessarily irrational: strong priors can make rational decision-making insensitive to new data (Gershman, 2019). Even after being provided with information on the size of their rebate, individuals may continue to believe that they are net losers. One explanation for these results is that a broader system of values–an ideology–is driving subjects' decision-making. This putative cognitive belief system would then also likely manifest when subjects justify their opposition to the tax. A better understanding of how ideology underlies carbon tax opinion could thus help inform the design of more effective carbon pricing policy in general, perhaps via more effective communication of the policy's benefits. This might involve first identifying, then appealing to (and in some instances directly challenging) issue frames commonly associated with carbon tax policy beliefs.

Issue frames, when written out, can be partially quantified by the semantic content of the writing. Quantitative semantic representations typically focus on frequency of word-use[1]. However, single-word frequency measures do not expose how the same words can be used when talking about different things. Other widely used approaches, such as sentiment analysis, classify responses into only a few affective classes ('like'/'dislike'). By formulating a rich latent topic structure, *topic models* address both these limitations. Topic models are now an established approach to understanding human-generated responses in the social sciences (Grimmer et al., 2022). The Structural Topic Model in particular has been applied to understand open-ended responses on a carbon tax in Spain (Savin et al., 2020), Norway (Tvinnereim et al., 2017), and the US (Povitkina et al., 2021). Here, we make a similar application to data collected in Canada. Unlike these previous works, however, we focus on topic-topic correlations as a means to interrogate ideology. Our contribution is a set of measures and analyses that characterize this latent correlation structure as a means of inferring ideology from open-ended survey responses.

[1] This includes more refined frequency measures such as Tfidf (see fig. 3).

To further motivate our application to carbon tax opinion, we hypothesized that opposition to carbon taxes arises from a well-worn ‘tax is bad’ ideology, involving a handful of correlated ideas (‘distrust in government’, ‘unfairness’, ‘infringement on personal freedom’, etc.) that mutually enforce each other (Lakoff, 2010).

Here, we use the Canadian dataset collected by (Mildenberger et al., 2022), unique in the richness of its meta-data, to fit the parameters of a generative bag-of-words model of word responses having a latent structure of topics. We fit models to different response types: the subset of responses of those who stated separately that they support the tax, of those who stated they opposed the tax, and of those two groups combined. After validating the fitted models, we use it to infer topic structure conditioned on the support and oppose response groups. We focus on the ways in which respondents mix topics when supporting or opposing the carbon tax. We not only find that responses are highly discriminable using these topic mixtures, but that there are clear differences in the correlation structure of the two response types that supports our hypothesis.

2. Methods

2.1. Dataset

We analyzed a dataset of 3313 response documents of survey respondents from across Canada in 2019 on the question of support for the recently implemented carbon tax as published in Mildenberger et al. (2022). In addition to asking respondents for a categorical opinion response in *support/opposition/notsure* (response type), demographic survey data were collected such as age, gender, and province. Lifestyle data (car use, residence environment) and belief data (partisanship (measured using party voting) and political ideology) were also collected. Ideology was measured with a unidimensional scale asking respondents, “In politics, people sometimes talk of left and right. Where would you place yourself on the following scale?”, with a 0 to 10 scale with extreme values labeled “Far to the left” (0) and “Far to the right (10). The mid-point was also labeled as “Centre” (5). In line with research documenting a general trend in partisan dealignment (Dalton and Wattenberg, 2002), partisanship was measured with a vote choice question asking respondents, “If a federal election were held today, for which party would you vote for?” Reflecting the multiparty nature of Canadian politics, response options included items for each of Canada’s main federal political parties: Liberal, Conservative, New Democratic Party, People’s Party, and Bloc Québécois (for respondents in Quebec). An option was also included for “I would not vote.” In the analysis, we recoded Ideology and Partisanship into 3 categories each. See Mildenberger et al. (2022) for more details on this rich meta-data and see the Supplemental Material for the specific reductions of categories that we used. In additional to categorical responses, a central open-ended question in the survey asked respondents to elaborate on why they choose the categorical response *support/oppose/not sure*. For this analysis, we preprocessed these open-ended responses. French responses were first translated into English using the Google Translate API. The response corpus was pre-processed through automated spell-checking, removal of stop words, and reduction to word stems. Words were tokenized: each word in the vocabulary was assigned an index and each word sequence in each response by transformed to its vector representation whose sequence of elements are the sequence of the corresponding word indices (e.g. the response $\{‘\mathrm{I}’, ‘\mathrm{love}’, ‘\mathrm{icecream}’\}$ might map to $\{3403, 5937, 2345\}$).

The *not sure* response types were analyzed in addition to the *support* and *oppose* types. They were only 11% of respondents, the majority of which gave responses to the effect of "I don’t know enough to answer" (see Supplemental Material). Thus, here we focus the presentation of our analysis on contrasting the *support* and *oppose* types.

2.2. Word relevance

We used a standard term-relevance measure that combines term frequency (as a measure of term relevance) and inverse document frequency (as a measure of term specificity):

- *term frequency*, $\mathrm{tf}(v,d)$. The frequency of a vocabulary word v in a document d is $\mathrm{tf}(v,d) := \frac{n(v,d)}{n(d)}$, where $n(v,d)$ is the number of times term v is in document d and $n(d) = \sum_{v\in\mathcal{V}} n(v,d)$ is the total number of words in document d. The term frequency over all the documents is then,
$$\mathrm{tf}(v) := \frac{\sum_{d=1}^{D} n(d)\mathrm{tf}(v,d)}{N_{\mathrm{total}}},$$
where the denominator $N_{\mathrm{total}} = \sum_{d=1}^{D} n(d)$ is just the total word count across all D considered documents.
- *inverse document frequency*, $\mathrm{idf}(v)$. Here,
$$\mathrm{idf}(v) = \frac{\log(D+1)}{\log(n(v)+1)+1},$$
where $n(v)$ is the number of documents in which term v appears, i.e. $n(v,d) > 0$. Idf is like an inverse document frequency.

Term frequency-inverse document frequency (Tfidf) for a word-document pair is then defined simply as the product, $\mathrm{Tfidf}(v,d) := \mathrm{tf}(v,d)\mathrm{idf}(v)$. This is then averaged over documents to create a word-only measure. We used the *sklearn* package implementation of Tfidf that averages the normalized values to remove dependence on the length of a document

$$\mathrm{Tfidf}(v) = \frac{1}{D}\sum_{d=1}^{D}\frac{\mathrm{Tfidf}(v,d)}{||\mathrm{Tfidf}(\cdot,d)||},$$

where $||\boldsymbol{x}|| = \sqrt{\sum_{i=1}^{N} x_i^2}$ is the Euclidean norm (throughout the paper, we denote vectors with bold symbols). As a word relevance metric, Tfidf aims to add more discriminability than frequency alone by downweighting words that appear in many documents since these common words are less discriminative. We computed word clouds and rank histograms using the Tfidf values.

2.3. Response-type Classification

We aimed to predict the response type label (*oppose*/*support*) for each response, first using the word-level representation (*i.e.* which words appear and how many times), then using the inferred topic mixture weights obtained from specific (e.g. K-dependent) fitted STMs on the combined dataset of support and oppose responses (see the following sections for details on the topic model). As preprocessing, we applied maximum absolute value rescaling to handle sparsity in the data for the word-level representation, and standard rescaling in the topic mixture case. In both cases, we then performed logistic regression. We then ranked features (words or topic weight vector components, respectively) by their weighted effect size and then trained a classifier for each subset of ranked features of increasing number, n. For each n, we ran 100 trials of a 2:1 train/test split and report the mean and variance of classification accuracy.

2.4. The Structural Topic Model

Topic models are generative models that generate a set of words in a response document according to some distribution. Topic models are typically *bag of word* models, which eschew grammar and syntax

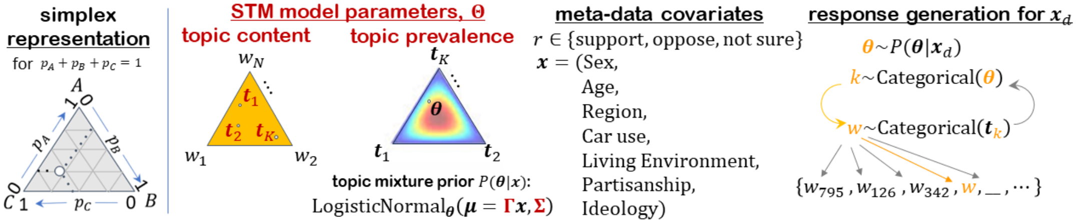
simplex representation
for $p_A + p_B + p_C = 1$
A
B
C
1
0
p_A
p_B
p_C
STM model parameters, Θ
topic content
topic prevalence
w_N
w_1
w_2
t_1
t_2
t_K
θ
topic mixture prior $P(\theta|x)$:
LogisticNormal$_\theta(\mu = \Gamma x, \Sigma)$
meta-data covariates
$r \in$ {support, oppose, not sure}
x = (Sex,
Age,
Region,
Car use,
Living Environment,
Partisanship,
Ideology)
response generation for x_d
$\theta \sim P(\theta|x_d)$
$k \sim$ Categorical(θ)
$w \sim$ Categorical(t_k)
$\{w_{795}, w_{126}, w_{342}, w, \_, \cdots\}$

***Figure 1. The Structural Topic Model for analysis of public opinion survey data**. From left to right: Recall the simplex as the space of frequency variables. STM parameters define topic content through the topics' word frequencies and topic prevalence through parameters associated with the topic mixture prior. We use seven meta-data covariates. We also store the respondent's categorical response (in support, in opposition or not sure) to the issue question (here carbon tax). By sequentially sampling topics and words, the model produces a bag-of-words open-ended response elaborating on this opinion.*

to focus only on the content and prevalence of words (*i.e.* sampling distributions do not condition on previously sampled words in a response). We select a topic model with rich latent structure: the *Structural Topic Model* (STM) (Roberts et al., 2014). Like the Correlated Topic Model (Blei and Lafferty, 2005), the STM uses a logistic normal distribution from which to sample the topic mixture weights on a document and can thereby exhibit arbitrary topic-topic covariance via the covariance parameter matrix of the logistic normal distribution. Unlike the Correlated Topic Model in which the mean and covariance matrix of the logistic normal distribution are learned parameters, the STM allows for parametrizing these using a linear model of the respondents' meta-data. We discuss our choice of meta-data model below. In the standard implementation of STM that we use here (fig. 1), only the mean of the logistic normal is made dependent on the meta-data. This adds flexibility beyond that offered by the CTM and makes the STM appropriate for datasets with non-trivial latent topic structure as we expect here. Specifically, our use of the STM model specifies the parameter tuple $\boldsymbol{\Theta} = (\{\boldsymbol{t}_k\}_{k=1}^K, \Sigma, \Gamma)$ as follows: (fig. 1):

1. *Topic content*: an underlying set of K topics indexed by k, $\{\boldsymbol{t}_k\}_{k=1}^K$, where each topic, $\boldsymbol{t}_k = (\beta_{k,1}, \ldots, \beta_{k,|\mathcal{V}|})$, is a list of word frequencies on a given vocabulary $\mathcal{V}$ indexed by v ($\sum_{v=1}^{|\mathcal{V}|} \beta_{k,v} = 1$ for all k), and
2. *Topic prevalence*: a prior distribution $P(\boldsymbol{\theta}|\boldsymbol{x})$ on the topic mixture vector, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ (weights are normalized, $\sum_{k=1}^K \theta_k = 1$) that is conditioned on the meta-data class, $\boldsymbol{x}$. Here, the distribution is chosen as a logistic normal with covariance matrix parameter Σ assumed to be the same across meta-data classes, and mean $\boldsymbol{\mu} = \Gamma \boldsymbol{x}$ is a linear transformation of $\boldsymbol{x}$ with matrix parameter Γ.

A single response sample for a given response document of length N words that comes from a respondent in meta-data class $\boldsymbol{x}$ is generated by sampling a topic mixture weight vector $\boldsymbol{\theta}$ once from the weight vector distribution $P(\boldsymbol{\theta}|\boldsymbol{x})$, then iteratively sampling a topic index, $z_n \in \{1, \ldots, K\}$, from Categorical$_K(\boldsymbol{\theta})$ and a word, $w_n \in \mathcal{V}$ from the distribution formed from that topic's frequencies, Categorical$_{|\mathcal{V}|}(\boldsymbol{t}_{k=z_n})$, N times to make the response $\{w_1, \ldots, w_N\}$, with $n = 1, \ldots, N$. Topic frequencies are represented in log-space according to the Sparse Additive Generative text model (Eisenstein et al., 2011). This topic content model allows for meta-data dependence. However, again confronted with the challenging model selection problem, and without strong hypotheses about how demographic-biased word use affect topic correlations (our primary interest), we chose to leave out this dependence.

2.5. STM Model Fitting

We built a Python interface to a well-established and computationally efficient STM inference suite written in the R programming language (Roberts et al., 2019). This STM package has a highly optimized parameter learning approach that largely overcomes the inefficiency of not using a conjugate prior[2]. In particular, STM uses variational expectation maximization, with the expectation made tractable through a Laplace approximation. Accuracy and performance of the method is improved by integrating out the word-level topic and through a spectral initialization procedure. For details see (Roberts et al., 2014). The Python code that we developed to access this package is publicly accessible with the rest of the code we used for this paper in the associated GitHub repository.

The parameter fitting performed by this STM package uses a prior on its covariance matrix parameter Σ of the topic mixture weight prior that specifies a prior on topic correlations: $\sigma \in [0, 1]$ giving a uniform prior for $\sigma = 0$ and an independence (i.e. diagonal) prior for $\sigma = 1$ (see the package documentation for more details on σ). Unless otherwise stated, we used the default value of $\sigma = 0.6$. Note that our main analysis of statistical measures does not use Σ. The package also offers priors on Γ. We used the recommended default "Pooled" option. This uses a zero-center normal prior for each element, with a Half-Cauchy(1,1) prior on its variance.

We fit three model types based on what dataset they are trained on. Here, a dataset $\mathcal{D} = \{(r_d, \boldsymbol{w}_d, \boldsymbol{x}_d)\}_{d=1}^{D}$ is a set of (categorical response, open-ended response, respondent metadata) tuples. We fit one model to each of the support and oppose respondent data subsets, $\mathcal{D}_{\text{support}}$ and $\mathcal{D}_{\text{oppose}}$, as well as one to a combined dataset $\mathcal{D}_{\text{combined}} = \mathcal{D}_{\text{support}} \cup \mathcal{D}_{\text{oppose}}$ that joins these two subsets of tuples into one. The model parameters obtained from maximum likelihood fitting to a dataset are denoted $\hat{\boldsymbol{\Theta}}$. As with other topic models, an STM takes the number of topics, K, as a fixed parameter (on which most quantities depend so we omit it to simplify notation). We thus obtained three K-dependent families of model parameter settings $\hat{\boldsymbol{\Theta}}_m$, for $m \in \{\text{support}, \text{oppose}, \text{combined}\}$. With model parameters in hand, we obtained our object of primary interest: the models' topic mixture posterior distributions conditioned on response type $r' \in \{\text{support}, \text{oppose}\}$,

$$P(\boldsymbol{\theta}|\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}_m, r') = \sum_{\boldsymbol{x}} P_m(\boldsymbol{\theta}|\boldsymbol{x}) p(\boldsymbol{x}|r') \tag{1}$$

$$\approx \frac{1}{|\mathcal{D}_{r'}|} \sum_{d \in \mathcal{D}_{r'}} \text{LogisticNormal}(\Gamma_m \boldsymbol{x}_d, \Sigma_m) \ , \tag{2}$$

where we have substituted the Logistic Normal distribution for $P_m(\boldsymbol{\theta}|\boldsymbol{x})$ and use the empirical average over $\boldsymbol{x}$. To compare the ideologies behind support and oppose responses, we make the two comparisons from evaluating eq. (1):

$$P(\boldsymbol{\theta}|\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}_{\text{support}}, \text{support}) \overset{?}{\longleftrightarrow} P(\boldsymbol{\theta}|\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}_{\text{oppose}}, \text{oppose}) \text{ and} \tag{3}$$

$$P(\boldsymbol{\theta}|\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}_{\text{combined}}, \text{support}) \overset{?}{\longleftrightarrow} P(\boldsymbol{\theta}|\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}_{\text{combined}}, \text{oppose}). \tag{4}$$

Since we are interested in topic correlations, we do not need to derive effect sizes for particular meta-data covariates, which in general depend on the particular meta-data model, here given by which elements of Γ in $\mu = \Gamma \boldsymbol{x}$ are non-zero and their sign. Consequently, we can sidestep the challenging meta-data model selection problem (Wysocki et al., 2022) by simply including all the measured covariates that might contribute (*i.e.* all elements of Γ are specified as real-valued free parameters). We thus selected a broad subset of the meta-data from Mildenberger et al. (2022): $\boldsymbol{x} = (\text{age}, \text{sex}, \text{region}, \text{car use}, \text{partisanship}, \text{ideology})$.

[2] The widely used Latent Dirichlet Allocation topic model uses a Dirichlet prior and the variational Bayes algorithm for this reason, but lacks expressiveness as a result.

We analyzed results across a range of K to ensure validity of our conclusions. Models with larger K typically give higher model likelihood but are at higher risk of over-fitting. The model likelihood on a heldout test dataset will however exhibit a maximum likelihood at finite K as overfitting to the remaining 'train' samples becomes significant with increasing K. The value of K where the heldout likelihood achieves its maximum will depend on the relative size of the test data relative to the train data in the train-test split. Nevertheless, a loose upper-bound on K is the value at which the maximum of the split realization-averaged held-out likelihood for a train-biased train-test data split (here 9:1) is obtained. For the support, oppose, and combined response type models, maximums were observed at $K = 15$, 18, and 20, respectively for averages over 100 and 200 train-test split realizations for separate and combined response type models, respectively (see fig. 6e,j). Note that the strong subadditivity (20/(15+18)=0.61) in these data-selected topic numbers suggests a high amount of overlap in the topic content between support and oppose responses.

We can directly compare the response-type conditioned posteriors of the combined model at any given K. In contrast, comparison of posteriors for the response-type specific models should account for the fact that the likelihood will in general suggest distinct K for each. With a prior in hand, we could simply marginalize over K in the posterior for each. We do not have such a posterior in hand unfortunately. A natural alternative is simply to compare models at the K at which each achieves the respective maximum in the likelihood. We highlight these maximum comparisons when plotting results over K.

2.6. Topic Quality

We assessed topic quality across different topic numbers using two related measures (the standard analysis in STM literature). First, exclusivity (high when a topic's frequent words are exclusive to that topic) was measured using the FREX score: the linearly-weighted harmonic mean of normalized rank of a word's frequency relative to other words within a topic ($\beta_{k,v}$) and the same across topics,

$$\mathrm{FREX}_{k,v} = \left(\frac{\omega}{\mathrm{NormRank}(\beta_{\cdot,v})_k} + \frac{1-\omega}{\mathrm{NormRank}(\beta_{k,\cdot})_v}\right)^{-1} \tag{5}$$

$$\mathrm{FREX}_k = \sum_{i=1}^{N_{\mathrm{top}}} \mathrm{FREX}_{k,v_i}(\omega) \ , \tag{6}$$

where $\mathrm{NormRank}(\boldsymbol{x})$ returns the n-dimensional vector of n-normalized ascending ranks of the n-dimensional vector $\boldsymbol{x}$ (high rank implies high relative value) and ω is the linear weight parameter (set by default to 0.7). In this and the following definition, the notation v_i is the index having descending rank i in $\beta_{k,\cdot}$ (low rank implies high relative value, *i.e.* the more weighted topic words) and the sum is over the top N_{top} such words for topic k (the default setting of $N_{\mathrm{top}} = 10$ was used). Note that this implies that the second term simplifies using the identity $\mathrm{NormRank}(\beta_{k,\cdot})_{v_i} = 1 - \frac{i}{|\mathcal{V}|}$. Second, semantic coherence (high when a topic's frequent words co-occur often), was defined as

$$C_k = \sum_{i=2}^{N_{\mathrm{top}}} \sum_{j=1}^{i-1} \log\left(\frac{D(v_i, v_j) + 1}{D(v_j)}\right) \ , \tag{7}$$

where $D(v)$ counts the documents (i.e. responses) in which the word v occurs and $D(v, v')$ counts the documents in which both words v and v' occur together. See the R package's documentation for more details (Roberts et al., 2019). We can study topic quality by analyzing the set of (FREX_k, C_k) pairs over topics in the plane spanned by exclusivity and coherence, where higher quality topics are positioned further to the top right (projection onto the diagonal gives one possible ordering of topics by quality). The center these points over k gives point that we use to evaluate particular STM models in fig. 3.

Table 1. *Four low-order statistics of a data cloud in a $(K-1)$-dimensional space ($\boldsymbol{\mu} \in \mathbb{R}^{K-1}$) with mean position $\bar{\boldsymbol{\mu}}$ and covariance matrix $\Sigma_{\boldsymbol{\mu}}$, having eigenvalues $\{\lambda_i\}_{i=1}^{K-1}$.*

Data Cloud Property	Topic Mixture Interpretation	Normalized Estimator
Position	Specificity	$\frac{1}{\sqrt{K-1}}\mathbb{E}\left[\|\|\boldsymbol{\mu}\|\|\right]$, average distance from the origin (here: equal usage)
Size	Variability	$\frac{1}{K-1}\sqrt{\sum_{i=1}^{K-1}\lambda_i}$, $\sqrt{\text{total variance}}$
Eccentricity	(Lack of) Expressivity	$\frac{1}{K-1}(\sum_{i=1}^{K-1}\lambda_i)^2/(\sum_{i=1}^{K-1}\lambda_i^2)$, intrinsic dimension
(+ve/-ve) Orientedness	Alignment	$\frac{2}{(K-1)(K-2)}\sum_{\rho_{ij}\gtrless 0, i<j}\rho_{ij}$, sum of signed correlation, $(\rho_{ij} = \frac{(\Sigma_{\mu})_{ij}}{\sigma_i\sigma_j})$

2.7. Statistical measures of topic mixture statistics

Our primary interest is in the statistical structure of the posteriors over topic mixture weight vectors, $\boldsymbol{\theta}$ eq. (1). However, the geometry of a domain can have a strong effect on the estimation of statistical properties and estimation using the $\boldsymbol{\theta}$-representation is biased because of the normalization constraint on weight vectors. A now well-established approach to performing statistical analysis on variables defined on the simplex (Pawlowsky-Glahn and Egozcue, 2001; Aitchison, 1982) maps these variables to the space of logarithms of relative weights (relative to the geometric mean, $g(\theta) = \prod_{i=1}^{K}\theta_i^{1/K}$). This bijective transformation maps the simplex to $\mathbb{R}^{K-1}$ such that LogisticNormal distributions map back into their Normal counterparts. Applying this transformation for the computed models, we find that the average distance between all pairs from the set of estimated means $\hat{\boldsymbol{\mu}}_d = \hat{\Gamma}\boldsymbol{x}_d$ over the three datasets is typically much larger than the size of the variation arising from the fitted covariance matrix parameter $\hat{\Sigma}$ (more than 95% of pairwise distances were larger than the total variance/dimension for 3 of the 4 comparisons (eqs. (3) and (4)) with the remainder at 85%; See Supplemental Material for details). Thus, the global geometry of the distribution is well-captured by the distribution of these means. We then focus on the empirical distribution of over $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{K-1})$, *i.e.* the set $\{\hat{\boldsymbol{\mu}}_d\}_{d=1}^{D}$, one for each response in the response type dataset and clustering according to demographic statistics.

We developed and used four intuitive characteristics of a data cloud (see table 1). The first is simply the mean position $\bar{\mu}$, while the remaining three rely on the covariance, $\Sigma_{\boldsymbol{\mu}}$, estimated using the standard estimator $\Sigma_{\hat{\mu}} = \frac{1}{D-1}\sum_{d=1}^{D}(\hat{\boldsymbol{\mu}}_d - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}_d - \hat{\boldsymbol{\mu}})^{\top}$, where $\hat{\boldsymbol{\mu}} = \frac{1}{D}\sum_{d=1}^{D}\hat{\boldsymbol{\mu}}_d$ is the mean estimate. In order to plot results across K without pure dimension effects displaying prominently, we apply appropriate normalization to each measure that gives them constant scaling in K. We label each measure according to its interpretation as a measure of ideology as follows. The position relative to equal usage (the origin in $\boldsymbol{\mu}$-space) measures how specific the usage is across topics. The size measures how much variability there is in how different individuals discuss that ideology. Eccentricity measures how the correlations limit the expressivity in how the ideology mixes topics. We measure this via reductions in a natural global measure of intrinsic dimension (Recanatesi et al., 2022). Finally, orientedness measures how strongly aligned or anti-aligned pairs of topics are.

Each measure is largely independent from the others and together the set is a largely comprehensive representation of unimodal, non-skewed data, since it covers global statistical features up to second-order. We discuss limitations and missing features of this set of measures in the discussion (section 4).

3. Results

3.1. Ideology and carbon tax support/opposition

To motivate our study of open-ended responses, we first show how support for the carbon tax depends on the demographic characteristics of the study's respondents. We find that support for carbon tax transitions from a majority to a minority opinion consistently across ideology-related feature dimensions

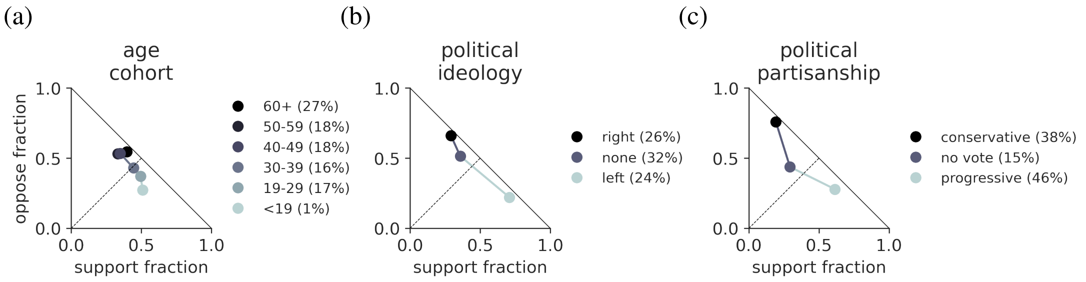

Figure 2. *Carbon tax opposition vs. support varies strongly with political demographics. Fraction of respondents that stated they opposed vs. the fraction who stated they support the carbon tax over (a) age cohort, (b) political ideology, and (c) political partisanship (measured as voting preference; see Methods for details;). Addition of the remaining "not sure" fraction (not shown) sums to 1. Opposition is a majority (minority) opinion for dots above (below) the dashed line.*

of left-to-right political ideology and liberal-to-conservative partisanship (see fig. 2). These transitions are relatively strong (they extend further away from equal support and opposition) in comparison with a transition with age, shown in fig. 2a as a baseline.

3.2. Discriminability of the word-level representation

Moving onto analysis of the open-ended responses, we first present the response type-conditioned word statistics in fig. 3a. One salient feature is how much the word 'tax' dominates in frequency within oppose responses, as compared with the support responses that recruit many frequent words. We then asked to what degree word content (which words appear and how many times) could be used to distinguish support and oppose responses. We performed logistic regression on the responses to distinguish support from oppose response type, finding best test accuracy around 80% (fig. 3a; see section 2 for details about the classification procedure). Word-level features are thus moderately informative in distinguishing support from opposition. This word-level information is spread heterogeneously over the vocabulary. In particular, by rerunning the classification on the n most contributing words (with contribution by frequency-weighted effect size) to each of the two target labels (support/oppose), we show in fig. 3b that test accuracy is maximized for about 100 words for each label (200 words overall) out of the more than 2500 stem words appearing in the data. We also show word clouds of the frequency-weighted effect size for $n = 100$ that display the most discriminating words. In sum, many words seem to contribute to a moderate level of discriminability.

A principal drawback of this word-level approach to assessing the discriminating power of the text responses is that individual words are assigned to one or the other response type. Yet, we know the words used by oppose and support responses overlap. For example, 'carbon' and 'tax' are used in high frequency by both response types. Removing this pair of words allows for higher discriminability (84% accuracy; not shown). That removing these central words increases discriminability suggests first-order word statistics are a poor representation of the semantics of the responses and that we should pursue approaches that employ latent representations that make use of words in different ways. In particular, it is the context in which these terms are used–what distinct ideas are evoked by the term–that ultimately distinguish the two response types and serves as a better semantic representation.

3.3. Topic modelling

To investigate the contextual origin of response type discriminability using the open-ended responses, we used a generative topic model with latent topic structure. We analyzed two model settings: (1)

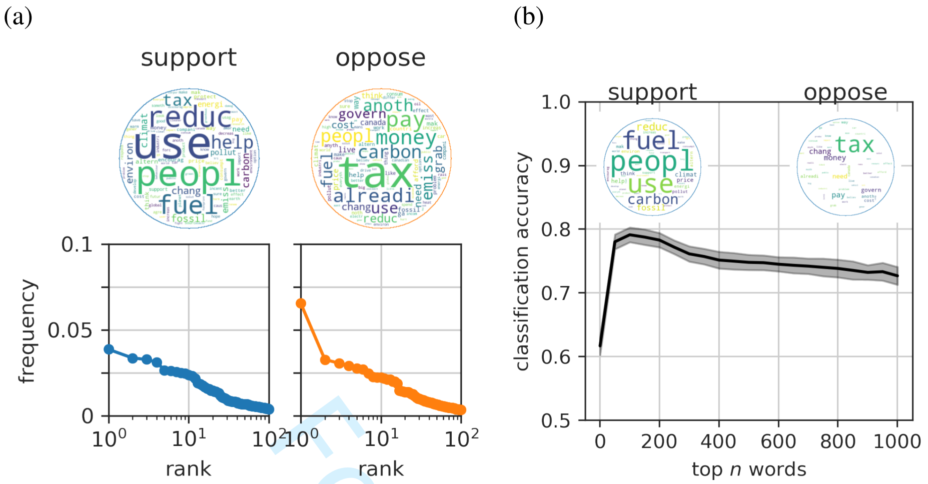

Figure 3. Word-level representations*. (a) Tfidf word cloud (top) and rank plot (bottom) for the word statistics of support and oppose response types. (b) Classification test accuracy of word-level content to predict response type (support/oppose; gray shows ± standard deviation over 100 random train-test splits). Accuracy is plotted for projections of the data onto the n most predictive words (See Methods for details). Inset are word clouds of log-transformed frequency-weighted effect size for the $n = 100$ most predictive words for support (left) and oppose (right) classifications.*

response type-specific models fit to responses of certain type (support/oppose) and (2) a single model fit to both types of responses combined.

To illustrate the topic content, in table 2, we show the top words associated with a given topic in a given model for the support and oppose response-specific models, as well as for the combined model (all for $K = 7$). A topic's top words can evoke an unambiguous semantic label for some topics, but not others. For example, topic a and e in the combined model evoke opposition because of the focus on cost of living and because of the word 'grab' as in 'tax grab', respectively. However, subjective inspection leaves many cases largely ambiguous as to whether a topic in the combined model is weighed more by support or oppose responses. Complementary to top words, top responses can also convey topic quality and even topic labels. For example, we perceived high semantic coherence in topic d, about unfairness of the tax on an international level (as well as, though to a lesser degree, in topic f about economic benefits, in topic c about its effectiveness in bringing about behaviour change, etc.). The latter were mostly not transparent in the list of top words, but only in the responses, *i.e.* in how these words were used. We list some top responses for these topics in the Supplemental Material for reference. It is hard to determine if the ambiguity in determining a topic label from a set of top responses arises from low topic quality reflecting a weakly clustered topic structure in the data or if it reflects limitations in our ability to identify salient topics. Another convoluting factor is that these topics are specific to the chosen value of K insofar as it imposes a specific number of topics, independent of any natural structure in data. For example, clusters of co-occuring words that are well-separated from each other will nevertheless likely be joined when K is too low and single clusters will likely be labeled incorrectly as multiple distinct clusters when K is too high.

To gain a deeper understanding of the relative changes in the topics as K increased without having to resort to *ad hoc* interpretation of the topics' top words, we developed two complementary analyses. First, we analyzed how topics breakup as K is increased (we call this a depth phylogeny). Inspired by methods of representing genetic phylogenies, we drew a tree structure obtained from connecting each topic to the topic in the model learned with 1 fewer topics that was nearest in the space of vocabulary word frequencies. In order to see how support and oppose responses spread their use of these topics

Table 2. *Example topic content (for $K = 7$ and independent topic mixture prior, $\sigma = 1$).*

Topic label	Frequency-ranked words
	Support model
a	chang, environ, climat, pollut, pay, mak, effect, protect, product, earth
b	peopl, help, energi, think, price, better, like, car, decrea, transport
c	emiss, less, increa, caus, govern, anoth, save, solut, drive, option
d	fuel, reduc, fossil, compani, consumpt, must, warm, global, incent, renew
e	tax, money, support, put, good, planet, believ, may, make, work
f	use, need, way, encourag, cost, hope, someth, gener, futur, thing
g	carbon, altern, sourc, find, get, develop, look, forc, creat, corpor
	Oppose model
a	use, fuel, work, car, altern, price, electr, like, stop, better
b	emiss, canada, countri, get, make, world, pollut, less, thing, big
c	carbon, alreadi, live, price, heat, caus, global, high, economi, warm
d	chang, reduc, anoth, grab, climat, noth, effect, believ, hurt, spend
e	tax, much, consum, problem, compani, industri, pocket, year, back, person
f	peopl, money, pay, fuel, need, cost, fossil, way, enough, mak
g	govern, put, think, environ, canadian, drive, emiss, incom, mani, averag
	Combined model
a	pay, alreadi, cost, enough, peopl, know, live, much, work, afford
b	money, mak, way, think, get, use, make, consum, environ, altern
c	fuel, reduc, carbon, chang, use, emiss, fossil, peopl, climat, help
d	canada, carbon, countri, effect, world, problem, pollut, believ, one, heat
e	tax, govern, anoth, grab, noth, put, price, car, canadian, still
f	need, like, support, thing, energi, someth, option, wast, resourc, gener
g	environ, tax, peopl, less, energi, pollut, encourag, altern, use, compani

for topics learned on both responses, we computed a topic depth phylogency for the combined model (see fig. 4a). Branch thickness denotes closeness[3]. We also added the relative support and oppose contribution to the mixture weight component of that topic as a piechart at each topic node in the tree. This allowed us to observe the cleaving of more coarse topics at low values of K into their finer-grained internal components as K increased. Reminiscent of phylogenetic analyses, we define a semantic-agnostic topic nomenclature using the row and column labels, respectively. In most cases, we see that topics are recruited by both response types so that single topics alone are insufficient for discriminating oppose from support responses. It is the weighting across multiple topics that gives sufficient information. That said, one salient feature of this particular tree is that topic 7f is the common ancestor for many of the topics at larger values of K. It predominantly loads onto support responses, recapitulating existing results in the literature of a diverse and diffuse set of topics discussed by those who support a carbon tax. In the following section, we will see that this $K = 7$ model is highly discriminative and this topic in particular conferred more than 99% accuracy.

In a second analysis, we assessed the quality of the topics over different K using the standard pair of metrics of exclusivity and semantic coherence (see Methods section). Running the model on each response type separately produced topics whose quality (fig. 4b) show that topic-averaged values give a linear trade-off between the two, with topic number K setting where along the trade-off the model resides. This suggests the linear combination as a family of scalar metrics of topic quality. Across different response types and values of K, the fixed sum (*i.e.* equal weights) of exclusivity and semantic coherence is highest in *oppose* responses, followed by *support* and *not sure*. We attribute this ordering to the singular focus that *oppose* responses have on the word ‘tax’, as compared to the much more diffuse responses (in word use) of *support* responses. Note, however, that the topic variance in each of the two quality metrics is at least as large as the difference in topic means at fixed topic number so the

[3] We use the inverse Euclidean distance between topics, normalized by the sum of these distances over topics in a given model.

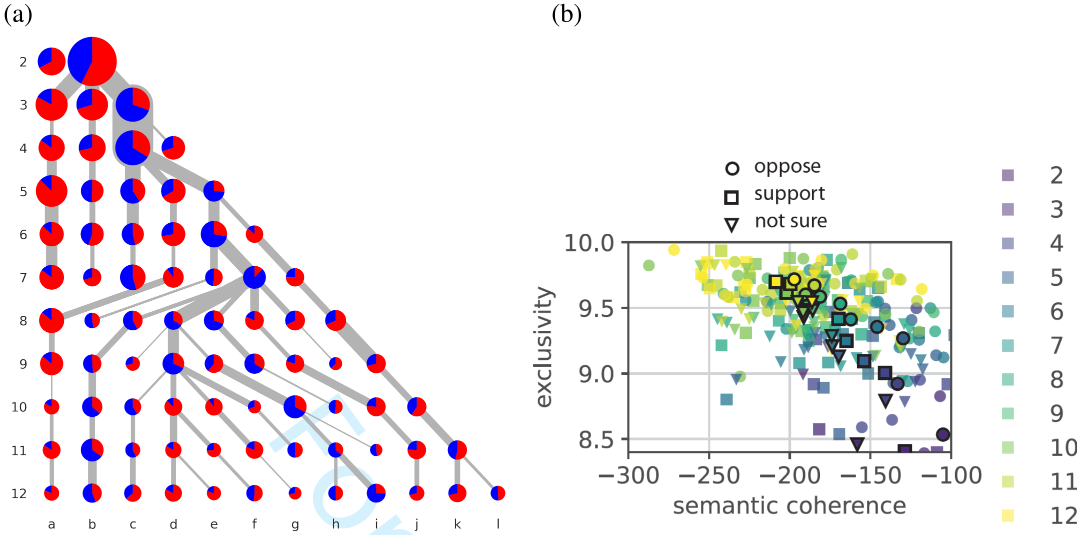

Figure 4. Topic structure over topic number, *K. (a) Topic trees from varying the number of topics, K, in the combined model. A row corresponds to a given value of K starting at the top with K = 2 down to K = 12. The relative support (blue) and oppose (red) contribution to each response-averaged mixture weight component is shown as a piechart. Link thickness is inverse Euclidean distance between topics. (b) Topic quality for response type models. Topic-averaged exclusivity plotted against topic-averaged semantic coherence for topic number K = 2, . . . , 12 (colors) for the three response types (marker type). Black outlined markers denote the unweighted average over K.*

aforementioned ranking is only apparent after averaging over topics. Semantic coherence bottomed out at around $K = 10$ topics. Topic quality for the composite model was qualitatively similar.

How predictive of response type are the way topics are recruited in the responses? In fig. 5a, we plot the empirical distribution of mixture weights, projected into its first two principal components, and colored coded with each response type. High separability (especially for $K = 7$) is observed across different topic number K, suggesting high discriminability. We performed a classification analysis on the support/oppose label similar to that using word use and presented at the beginning of our results (see fig. 3(b)), but now instead using the inferred topic mixture weights. Note that this representation is able to recruit information in the correlations among words that distinguishes the two response types. In fig. 5(b), we again show performance as a function of using the most predictive features (here not words, but topic mixture weights). Consistent with the high observed separability, we find accuracies approaching and more than 99% across all values of $K > 2$, demonstrating the power of this latent structure to predict support or opposition to the tax. Accuracies were only marginally lower at 95% or more for $\sigma = 0$ (not shown). For $K = 7$, near 100% accuracy is possible even using only a single topic and perfect accuracy for only two topics. As an example of more fine-grained analysis made possible by our approach, we focus in on these two topics. In fig. 5(c), we show the projection of the set of best-estimate mixtures across responses, colored by response type, into the plane in mixture space spanned by the components associated with these two topics. We see that the perfect classification results from topic 7 (topic 5) being strongly suppressed in support (oppose) responses. We further assessed this association by inspecting top responses for each topic (one from each is shown in fig. 5(c)). The responses in topic 7 describe the tax as a malicious injustice imposed on citizens, while those of topic 5 describe the effectiveness of the monetary incentive. We note that while the term ‘tax grab’ is the core idea in Topic 7 (‘grab’ is in 8.5% of oppose responses compared to only 1.5% of support responses), the idea recruits a much wider set of terms operationalized as the support of topic 7 on the

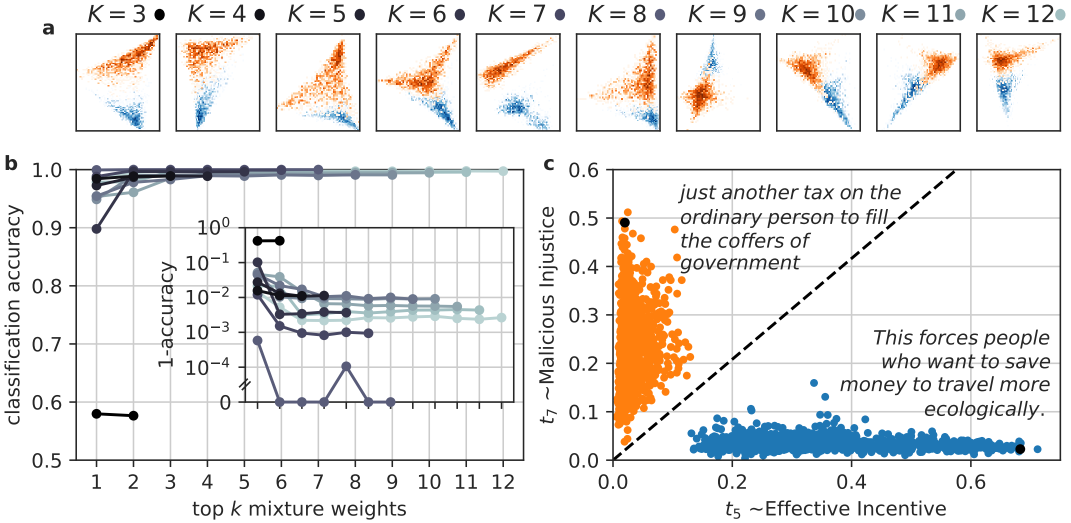

***Figure 5. Topic-level Representations**. (a) Projection of inferred mixture weights from the combined model onto the first two PCA components for a range of values of K. A histogram of counts on a fine-grained grid is shown using a white-to-blue and white-to-red color scale for support and oppose responses, respectively. (b) Classification test accuracy for using the best k of the mixture weights ranked by largest frequency-averaged effect size(black to gray is different values of K from 2 to 12). (c) The mixture estimates projected into the two top topics associated with the top k = 2 mixture weights for K = 7. The blacked-dashed line is the optimal classifier that achieves 100% accuracy. The text and mixture position of a top response for each topic is shown as a block dot.*

vocabulary. It is this wider set of words that confers the discriminability (in this case terms that are not used in high frequency by support responses).

We now turn to the topic-topic correlations to better understand the origin of the topic representation's higher classification performance over word frequency representations. We also arrive at analyses appropriate to investigating the structure of the mixture statistics that putatively hold distinctive features of the underlying ideology. We used the four measures of data cloud geometry summarized in table 1 to assess the topic-topic correlation structure using $\boldsymbol{\mu}$-representation as explained in the Methods. For the combined response type models, we can directly compare the support and oppose responses at each value of K, so we represent the results as the difference in the values of each measure over the two response types for a range of K around the maximum in the heldout Likelihood (positive values indicate oppose responses exceed support responses on that metric; fig. 6a-d). For the separate response type models, for which the best fitting value of K differs, we simply report both measures over a larger range of K (fig. 6f-i). By showing the heldout likelihood (fig. 6e,j), a comparison at respective maximums could be made. We find consistent results for both comparisons. Namely, oppose responses are more specific, somewhat more variable, much less expressive, and more aligned across a wide range of K around where the held-out likelihood is maximized. All three of the non-weak results are consistent with our hypothesis of a rigid ideology underlying carbon tax opposition.

4. Discussion

In this study, we presented a set of principled, quantitative measures and analyses to pin down topic structure in the learned parameters of the STM. We applied them to understand the effects of ideology behind carbon tax opinion in Canada. We find topic mixture weights derived from the learned models

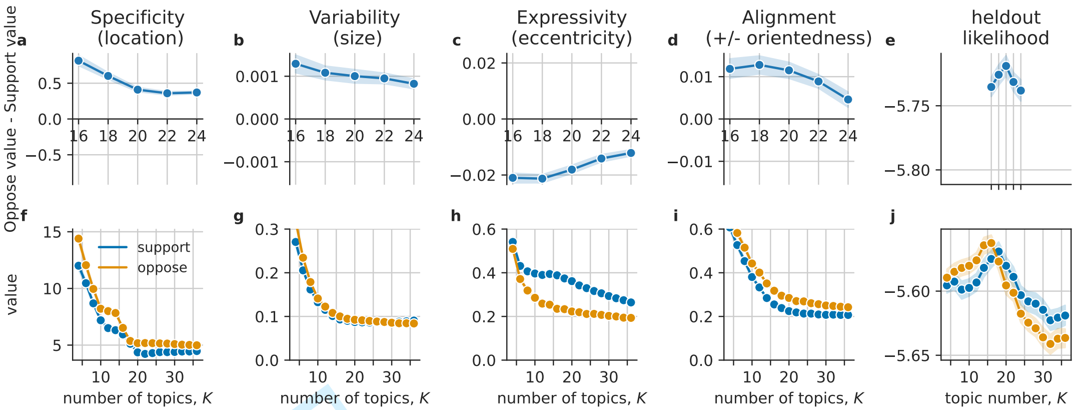

__Figure 6. Evaluations of Statistical Measures of Mixture Statistics__. Top (a-d): The average difference of the oppose value and the support value in each respective measure from the combined model (eq. (4)). Bottom (f-i): The value of the measure for each of the support (blue) and oppose (orange) models (eq. (3)). Lines are mean estimates and error bars are standard deviation of 100 (200) posterior samples for top (bottom). Respective held-out model likelihoods are shown in (e) and (j).

are highly predictive of opposition to or support of the carbon tax and we presented a topic tree analysis similar to phylogenetics to show that this performance arises when the model has a sufficient number of topics that together are discriminating on each response. Finally, we proposed a set of statistical measures on topic mixtures and evaluated them to find that the oppose responses had higher levels of specificity and orientedness, and were less expressive. This suggests carbon tax opposition is explained by its proponents through a more well-worn, coherent ideology. Together, these results on carbon tax are thus suggestive of our initial hypothesis.

We anticipate the methods we present here can be useful to social scientists who are interested in inferring social and political phenomenon from open-ended survey responses. Most obviously, our methods could be used to further quantify the structure of collective beliefs. In a longitudinal approach, they could also reveal the dynamics of these collective beliefs, and in particular their response characteristics, for example from large-scale interventions or events (e.g. the COVID19 pandemic, Russia-Ukraine war, etc.). A study of the timescales over which the effects of interventions decay would inform the recent work on the notion of inoculation to misinformation. To push the metaphor, the characteristics of the memory it induces have not yet been deeply probed (Roozenbeek et al., 2022). Our work suggests that the topic neighborhood around the focus of any given intervention may play a role via the topic-topic correlation structure. This could also impact policy design, for example in how the carbon tax is promoted. Typical communications advice is to focus on a single issue/message to not overload the target audience. However, neighborhood effects that are strongly restoring (Hall and Bialek, 2019) could rapidly erase the effects of single issue interventions. Instead, a communications strategy aimed at a local neighborhood of target issues in some appropriately defined issues space may be more effective at loosening the grip that ideology has on opinion.

There are many potential follow-up areas to take this work. For one, the existing set of metrics of mixture statistics only cover first and second order statistical moments. Moreover, there even aggregate low order statistics missing from this set, *e.g.* how concentrated is the distribution around its center (the expectation of $||\boldsymbol{\mu}||$ might provide signal into how hollow is the data cloud). In another direction, the efficiency of the topic representations suggests it could serve as an input embedding space for use in deep learning approaches to large-scale social computations (e.g. those run on large social media platforms). For example, recent work on reinforcement learning with human feedback for large language models has focussed on generating consensus statements from a set of human-generated opinions (Bakker et al., 2022). A currently limiting constraint of this approach is how many human-generated text inputs can be used. A topic space learned from many opinions could circumscribe this constraint.

We probed model behaviour over a range of K around the maximum in held-out model likelihood. We were able to make conclusions from the fact that the sign of the difference of the values of the metrics was unchanged over these ranges. This will not be true in general, in which case one approach is to marginalize K out of the problem. In a fully Bayesian approach, given a prior over K, one would infer a posterior over K and take the posterior average over any K-dependent quantities, e.g. average topic quality or the statistical metrics. Given that models with different K have a different number of parameters, a complexity-penalty term such as the AIC could be added to the posterior to account for variable model complexity.

One direction we did not pursue here was exploring how different meta-data models affect the results. Are there specific factor models of fewer covariates that best explain the data? As we motivated our choice of the all-in model, this is a delicate analysis requiring disentangling various cases. For example, is car use a confounder or collider of the effect of residence environment? Such distinctions will influence estimated effect sizes (Wysocki et al., 2022). A first step along this direction is to further analyze the results in the case of a sparsity prior on Γ. Including meta-data model dependence in topic content is a related direction for future work. Note that in absence of rich metadata, the STM model reduces roughly to the Correlated Topic Model (Blei and Lafferty, 2005), in which case topic correlations can be studied via the mixture covariance matrix parameter, Σ_{θ}.

Finally, there is a broader question of the validity of the generative model class. In particular, are topics models with latent structure suitable mathematical representations of semantic content of networks of beliefs? How much is lost when representing language as text, without syntax and without positional encoding? This question broaches broader philosophical issues around meaning in natural human language. Given the recent evidence of (position-encoding) large language models that achieve in-context learning (*c.f.* Mollo and Millière (2023)), we believe text is a sufficiently rich representation of natural language from which to extract semantics. Stripping text of its positional encoding must result in some degree of loss of meaning. Topic models are nevertheless highly interpretable and we believe can serve as useful complements to less easily interpretable large language models. While we often think of syntax and word order as crucial to meaning, its importance to ecological communication may be overstated. That said, ecological communication also presents challenges to using text to infer intended meaning: a word can have one of many different meanings depending on who is speaking (Marti et al., 2023).

In sum, we see a bright future for topic models in quantitative social science and beyond. We hope that this work inspires their development, refinement and application as a research direction along which more computational data science approaches contribute to the social aspects of climate change.

Acknowledgments. MPT would like to acknowledge helpful discussions with Matto Mildenberger, Kathryn Harrison, and Dhanya Sridhar.

Funding Statement. MPT was funded under the CERC for Autonomous AI awarded to Irina Rish.

Competing Interests. The authors declare none.

Data Availability Statement. All supporting data was obtained from an existing Harvard Dataverse replication archive at https://doi.org/10.7910/DVN/3WBCH9.

Ethical Standards. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Author Contributions. MPT: Conceptualization (Equal), Formal analysis (Lead), Investigation (Lead), Methodology (Lead), Software (Lead), Writing - original draft (Lead), Writing - review & editing (Lead). EL: Conceptualization (Equal), Data curation (Lead), Supervision (Lead), Writing – review & editing (Supporting).

Supplementary Material. Supplementary Material is available here.

References

J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177, 1982. ISSN 00359246. URL http://www.jstor.org/stable/2345821.

J. E. Aldy and R. N. Stavins. The promise and problems of pricing carbon: Theory and experience. *The Journal of Environment & Development*, 21(2):152–180, 2012.

M. A. Bakker, M. J. Chadwick, H. R. Sheahan, M. H. Tessler, L. Campbell-Gillingham, J. Balaguer, N. McAleese, A. Glaese, J. Aslanides, M. M. Botvinick, and C. Summerfield. Fine-tuning language models to find agreement among humans with diverse preferences, 2022. URL https://arxiv.org/abs/2211.15006.

D. Baldassarri and A. Gelman. Partisans without constraint: Political polarization and trends in american public opinion. *American Journal of Sociology*, 114(2):408–446, 2008.

A. Baranzini, J. Goldemberg, and S. Speck. A future for carbon taxes. *Ecological Economics*, 32(3):395–412, 2000. ISSN 0921-8009. doi: https://doi.org/10.1016/S0921-8009(99)00122-6. URL https://www.sciencedirect.com/science/article/pii/S0921800999001226.

S. Birch. Political polarization and environmental attitudes: A cross-national analysis. *Environmental Politics*, 29(4):697–718, 2020.

D. M. Blei and J. D. Lafferty. Correlated topic models. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS'05, page 147–154, Cambridge, MA, USA, 2005. MIT Press.

R. J. Dalton and M. P. Wattenberg. *Parties Without Partisans: Political Change in Advanced Industrial Democracies*. Oxford University Press, 03 2002. ISBN 9780199253098. doi: 10.1093/0199253099.001.0001. URL https://doi.org/10.1093/0199253099.001.0001.

S. Drews and J. C. Van den Bergh. What explains public support for climate policies? a review of empirical and experimental studies. *Climate policy*, 16(7):855–876, 2016.

J. Eisenstein, A. Ahmed, and E. P. Xing. Sparse additive generative models of text. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1041–1048, 2011.

S. J. Gershman. How to never be wrong. *Psychonomic Bulletin and Review*, 26(1):13–28, 2019. ISSN 15315320. doi: 10.3758/s13423-018-1488-8.

J. Grimmer, M. Roberts, and B. Stewart. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press, 2022. ISBN 9780691207544. URL https://books.google.ca/books?id=cL40EAAAQBAJ.

E. Haites. Carbon taxes and greenhouse gas emissions trading systems: what have we learned? *Climate policy*, 18(8):955–966, 2018.

G. Hall and W. Bialek. The statistical mechanics of twitter communities. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(9):093406, sep 2019. doi: 10.1088/1742-5468/ab3af0. URL https://dx.doi.org/10.1088/1742-5468/ab3af0.

N. P. Kalmoe. Uses and abuses of ideology in political psychology. *Political Psychology*, 41(4):771–793, 2020.

D. Klenert, L. Mattauch, E. Combet, O. Edenhofer, C. Hepburn, R. Rafaty, and N. Stern. Making carbon pricing work for citizens. *Nature Climate Change*, 8(8):669–677, Aug 2018. ISSN 1758-6798. doi: 10.1038/s41558-018-0201-2. URL https://doi.org/10.1038/s41558-018-0201-2.

E. Lachapelle. Communicating about carbon taxes and emissions trading programs. In *Oxford research encyclopedia of climate science*. Oxford University Press, 2017.

G. Lakoff. Why it matters how we frame the environment. *Environmental Communication*, 4(1):70–81, 2010. doi: 10.1080/17524030903529749. URL https://doi.org/10.1080/17524030903529749.

L. Marti, S. Wu, S. T. Piantadosi, and C. Kidd. Latent Diversity in Human Concepts. *Open Mind*, 7:79–92, 03 2023. ISSN 2470-2986. doi: 10.1162/opmi_a_00072. URL https://doi.org/10.1162/opmi_a_00072.

A. M. McCright and R. E. Dunlap. The politicization of climate change and polarization in the american public's views of global warming, 2001–2010. *The Sociological Quarterly*, 52(2):155–194, 2011.

A. M. McCright, R. E. Dunlap, and S. T. Marquart-Pyatt. Political ideology and views about climate change in the european union. *Environmental Politics*, 25(2):338–358, 2016.

C. McGlade and P. Ekins. The geographical distribution of fossil fuels unused when limiting global warming to 2 c. *Nature*, 517 (7533):187–190, 2015.

M. Mildenberger, E. Lachapelle, K. Harrison, and I. Stadelmann-Steffen. Limited impacts of carbon tax rebate programmes on public support for carbon pricing. *Nature Climate Change*, 12(2):141–147, 2022. ISSN 17586798. doi: 10.1038/s41558-021-01268-3.

D. C. Mollo and R. Millière. The vector grounding problem, 2023.

V. Pawlowsky-Glahn and J. J. Egozcue. Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, 15(5):384–398, Oct 2001. ISSN 1436-3259. doi: 10.1007/s004770100077. URL https://doi.org/10.1007/s004770100077.

M. Povitkina, S. Carlsson Jagers, S. Matti, and J. Martinsson. Why are carbon taxes unfair? Disentangling public perceptions of fairness. *Global Environmental Change*, 70(September), 2021. ISSN 09593780. doi: 10.1016/j.gloenvcha.2021.102356.

B. G. Rabe. *Can we price carbon?* MIT Press, 2018.

S. Recanatesi, S. Bradde, V. Balasubramanian, N. A. Steinmetz, and E. Shea-Brown. A scale-dependent measure of system dimensionality. *Patterns*, 3(8), Aug 2022. ISSN 2666-3899. doi: 10.1016/j.patter.2022.100555. URL https://doi.org/10.1016/

j.patter.2022.100555.

M. E. Roberts, B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082, 2014. ISSN 15405907. doi: 10.1111/ajps.12103. URL http://dvn.iq.harvard.edu/dvn/dv/ajps.

M. E. Roberts, B. M. Stewart, and D. Tingley. Stm: An R package for structural topic models. *Journal of Statistical Software*, 91:1–40, 2019.

J. Roozenbeek, S. van der Linden, B. Goldberg, S. Rathje, and S. Lewandowsky. Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, 8(34):eabo6254, 2022. doi: 10.1126/sciadv.abo6254. URL https://www.science.org/doi/abs/10.1126/sciadv.abo6254.

I. Savin, S. Drews, S. Maestre-Andrés, and J. van den Bergh. Public views on carbon taxation and its fairness: a computational-linguistics analysis. *Climatic Change*, 162(4):2107–2138, 2020. ISSN 15731480. doi: 10.1007/s10584-020-02842-y.

E. Tvinnereim, K. Fløttum, Ø. Gjerstad, M. P. Johannesson, and Å. D. Nordø. Citizens' preferences for tackling climate change. Quantitative and qualitative analyses of their freely formulated solutions. *Global Environmental Change*, 46(May):34–41, 2017. ISSN 09593780. doi: 10.1016/j.gloenvcha.2017.06.005.

D. Welsby, J. Price, S. Pye, and P. Ekins. Unextractable fossil fuels in a 1.5° c world. *Nature*, 597(7875):230–234, 2021.

A. C. Wysocki, K. M. Lawson, and M. Rhemtulla. Statistical control requires causal justification. *Advances in Methods and Practices in Psychological Science*, 5(2):25152459221095823, 2022. doi: 10.1177/25152459221095823. URL https://doi.org/10.1177/25152459221095823.