

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	No software was used for data collection.
Data analysis	We developed custom code using Python 3.9.10. for data analysis, which is available at https://github.com/ByunghweeLee-IU/Belief-Embedding .

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

We used a publicly accessible dataset from Debate.org (<https://esdurmus.github.io/ddo.html>). A processed version of this dataset, including pre-processed user-level debate records and the fine-tuned models used in our analyses, is available at <https://github.com/ByunghweeLee-IU/Belief-Embedding>.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	n/a
Reporting on race, ethnicity, or other socially relevant groupings	n/a
Population characteristics	n/a
Recruitment	n/a
Ethics oversight	n/a

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study is quantitative and utilizes data from an online debate forum (Debate.org) to examine the interplay between thousands of beliefs. A fine-tuned large language model (LLM) was employed to create an embedding space that represents human beliefs. This study explores how the belief space captures the interconnectedness and polarization of diverse beliefs, and tests whether this embedding space can be used to predict new beliefs of individuals.
Research sample	The research sample consists of user participation records from the Debate.org dataset, including a corpus of 78,376 debates (68,900 unique debate titles excluding duplicates) by 42,906 debaters from October 15, 2007, to September 19, 2018. While the dataset covers a broad range of social and political topics and captures diverse perspectives, it is primarily based on user-generated content from the United States, and thus may not fully represent global beliefs or cultural diversity. Nonetheless, we selected the Debate.org dataset because it offers a large-scale, structured collection of debate participation records and belief expressions of users across diverse social and political topics. Its format allows for direct mapping of users' positions on debates, making it well suited for modeling belief structures and studying patterns of polarization.
Sampling strategy	We utilized the entire Debate.org dataset (from October 15, 2007, to September 19, 2018).
Data collection	We utilized the existing publicly accessible Debate.org dataset (version 2, available at https://esdurmus.github.io/ddo.html), which includes user voting records on debates. Since the study used an existing public dataset, the researchers were not blinded to the hypothesis. However, the data were collected independently of this study and prior to hypothesis formulation.
Timing	We utilized the publicly accessible Debate.org dataset, which was downloaded on March 18, 2023, from https://esdurmus.github.io/ddo.html . The dataset covers a period from October 15, 2007, to September 19, 2018.
Data exclusions	Debate titles that could not be interpreted as belief statements were excluded from the dataset. For example, titles consisting of a single word or vague versus comparisons were filtered out using GPT-4 to ensure consistency in belief statement formulation. Of the 68,900 unique debate titles, 8,914 were removed during this filtering process, resulting in 59,986 debate titles retained for analysis.
Non-participation	No participants were involved in this study.
Randomization	Our study doesn't involve random-control design.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involvement in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involvement in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

n/a

Novel plant genotypes

n/a

Authentication

n/a