New conceptions of truth foster misinformation in online public political discourse

Jana Lasser[1,2], Segun Taofeek Aroyehun[1], Fabio Carrella[3], Almog Simchon[3], David Garcia[1,2] and Stephan Lewandowsky[3,4,5*]

[1]Institute for Interactive Systems and Data Science, Graz University of Technology, Inffeldgasse 16C, Graz, 8010, Austria.
[2]Complexity Science Hub Vienna, Josefstädterstr. 39, Vienna, 1080, Austria.
[3] University of Bristol, , Bristol, UK.
[4]University of Western Australia, Crawley, 6009, W.A., Australia.
[5]University of Potsdam, Potsdam, Germany.

*Corresponding author(s). E-mail(s): stephan.lewandowsky@bristol.ac.uk;

Abstract

The spread of online misinformation is increasingly perceived as a major problem for societal cohesion and democracy [1, 2]. Much attention has focused on the role of social media as a vector of misinformation [3]. The role of political leaders has attracted less research attention, even though leaders demonstrably influence media coverage [4] and public opinion [5], and even though politicians who "speak their mind" are perceived by segments of the public as authentic and honest even if their statements are unsupported by evidence or facts [6–8]. Here we show that in the last decade, politicians' concept of truth has undergone a distinct shift, with authentic but evidence-free belief-speaking becoming more prominent and more differentiated from evidence-based truth seeking. We analyze communications by members of the U.S. Congress on Twitter between 2011 and 2022 and show that political speech has fractured into two distinct components related to belief-speaking and evidence-based truth-seeking, respectively, and that belief-speaking is related to spreading of untrustworthy information. We show that in tweets by conservative members of Congress, an increase in belief-speaking of 10% is associated with a decrease of 6.8 points of quality (using the NewsGuard scoring system) in the sources shared in a tweet. In addition, we find that an increase of belief-speaking language by 10% in the shared articles themselves is associated with a drop in NewsGuard score of 4.3 points for members of both parties. By contrast, increase in truth-seeking language is associated with a slight increase in quality of sources. The results support the hypothesis that the current flood of misinformation in political discourse is in part driven by a new understanding of truth and honesty that has replaced reliance on evidence with the invocation of subjective belief.

1 Main

Numerous indicators suggest that democracy is under threat worldwide [e.g., 9, 10]. Although symptoms and causes of democratic backsliding are difficult to tease apart, a flood of misinformation—on social media, in hyperpartisan news sites, and in political discourse—is undoubtedly a core challenge to democracies [1]. Misinformation has been causally linked to populist voting [11] and ethnic hate crimes [2].

A troubling aspect of misinformation is that it lingers in memory even if people acknowledge, believe, and try to adhere to a correction [12]. That is, even though people may adjust their factual beliefs in response to corrections [e.g., 13], their political behaviors and attitudes may be largely unaffected [e.g., 6, 7]. Perhaps most concerningly, in some circumstances people may even come to *value* overt dishonesty as a signal of authenticity [8]. A politician who routinely and blatantly lies is overtly violating the "establishment" norm of being accurate and honest. Within a populist logic this identifies the politician as an authentic champion of "the people"—dishonesty thus becomes a feature and a sign of distinction. For example, polls have shown that around 75% of Republicans considered President Trump to be "honest" (e.g., NBC poll, April 2018), which is at odds with the records kept by independent fact checkers and the media, which have identified more than 30,000 false or misleading statements during Trump's presidency (*Washington Post* fact checker).

The disconnect between accuracy and politicians' attractiveness to voters has also been established in behavioral experiments involving the American public [6, 7].

This discrepancy between factual accuracy and perceived honesty is, however, entirely understandable if "speaking one's mind" on behalf of a constituency is considered a better marker of honesty than veracity. The idea that untrue statements can be "honest", provided they arise from authentic belief speaking, points to a new ontology of truth that no longer relies on the notion of evidence but on a radically constructivist appeal to an intuitive shared experience as a new "truth" [1].

To date, there have been several attempts to highlight the changing ontology of truth and honesty, and how these changes are accompanied by different streams of misinformation [e.g., 1, 14, 15]. A particularly promising recent development is the conceptual model of honesty proposed by [16]. The model differentiates between three components of honesty: belief-speaking, truth-seeking, and fostering understanding. The "belief-speaking" aspect of honesty relates only to the speaker's beliefs, thoughts, and feelings, without regard to factual accuracy or downstream consequences on beliefs of the listeners. The truth-seeking component relates to the search for truthful information and an updating of one's beliefs based on that information. The fostering understanding component relates to communication that tries to foster true beliefs in an audience.

This model of honesty was developed in the context of corporate communication, but the first two proposed components of honesty, belief-speaking and truth-seeking, map onto a recent analysis of two distinct ontologies of political truth [17]. The first of these two ontologies has roots in fascism and refers to a radical constructivist "truth" based on intuition and belief speaking [e.g., 18]. This conception of truth is typically suffused with emotion and feelings and its adherents sometimes reject the role of evidence outright. Instead, Nazi ideology postulated the existence of an "organic truth" based on personal experience and intuition that can only be revealed through inner reflection but not external evidence [e.g., 18, 19]. The second ontology, based on truth-seeking, is firmly anchored in democracy and seeks a shared evidence-based reality. This conception of truth is typically dispassionate and does not admit appeals to emotion as a valid tool to adjudicate evidence, although it does not preclude truth-finding from being highly contested and messy ([20] vs. [21]).

For a democracy, a conception of truth that is based on "belief-speaking" alone can have painful consequences as democracy requires a body of common political knowledge in order to enable societal coordination [22]. For example, people in a democracy must share the knowledge that the electoral system is fair and that a defeat in one election does not prevent future wins. Without that common knowledge, democracy is at risk. The attempts by Donald Trump and his supporters to nullify the 2020 election results have brought that risk into sharp focus [23].

To achieve a common body of knowledge, democratic discourse must go beyond belief-speaking. Democratic politics requires truth-seeking by leaders—otherwise, they may choose to remain wilfully ignorant of embarrassing information, for example, by refusing briefings from experts that are critical of their favoured public-health policy. In addition, democratic politics must seek to foster true beliefs in the public—otherwise even truthful statements ("There is not a sexual relationship"; Bill Clinton, 21 January 1998) can be corrosive when used to create a mistaken impression (Clinton was in an illicit sexual relationship, albeit not at the time of the statement).

To date, there has been much concern but limited evidence about the increasing prevalence of belief-speaking at the expense of truth-seeking in

American public and political life. We aim to explore this presumed shift in conceptions of truth and honesty by focusing on Twitter activity by members of both houses of the U.S. Congress. The choice of Twitter is driven by the fact that public outreach on Twitter has become one of the most important avenues of public-facing discourse by U.S. politicians in the last decade [24] and is frequently used by politicians for agenda-setting purposes [25].

Our analysis addressed several research questions: Can we identify aspects of belief-speaking, truth-seeking, and fostering of true beliefs in public-facing statements by members of Congress? And if so, how do these three components evolve over time? What partisan differences, if any, are there? Is the quality of shared information linked to the aspects of honesty? To answer these questions, we performed a computational analysis of an exhaustive dataset of tweets posted by US politicians, detecting links to misinformation sources and analyzing text of tweets and news sources.

2 Identifying honesty components in political speech

We first sought to identify the three components of honesty — belief-speaking, truth-seeking and fostering understanding — in public-facing political speech by elected U.S. officials. We showed that the identified honesty components are different from existing measures that capture related text features and we quantitatively explored how different components of honesty are related to a variety of topics prevalent in political discourse.

For our analyses, we collected a corpus of tweets from members of the U.S. Congress between January 1, 2011 and March 16, 2022. After removing retweets and duplicates, our corpus contained a total of 1,824,022 tweets (see Section "U.S. Congress Member tweet corpus" in the Methods).

To measure the three components of honesty in text, we created three dictionaries of words associated with each of the concepts. We followed a computational grounded theory approach [26] to incorporate both expert knowledge and computational pattern recognition. We started with a list of seed words for each component, followed by computational expansion and iterative pruning and refinement through human input, as illustrated in Extended Data Figure 1 (see also Section "Honesty component keywords and validation" in Methods). The full list of keywords is supplied in Extended Data Table 1. We subjected the final list of keywords to validation by three trained raters by first using the keyword lists to label 200 posts for each component from the tweet corpus. We then asked raters to confirm whether the component assigned to that sub-sample was effectively present in the tweet. We found substantial agreement between raters and computational identification of honesty components for belief-speaking (Krippendorff's $\alpha = 0.42$ [0.30; 0.55]) and truth-seeking (0.53 [0.31; 0.70]), but not for fostering understanding (-0.57 [-0.68; -0.45]), suggesting inverse agreement [27]. While generally a value of $\alpha \geq 0.8$ is accepted as standard for high agreement, for exploratory and ground-breaking research, lower values may be acceptable [28, 29]. A possible reason for the difficulty of reliably identifying the "fostering understanding" component is the fact that the initially proposed three honesty components by [16] were grounded in the analysis of conversations in a work-related context, which is substantially different from the public-facing political speech contained in our corpus. We concluded that we can only identify belief-speaking and truth-seeking in political conversation, not fostering understanding. Therefore, going forward, we removed the seeking-understanding component from further analysis.

By identifying the presence of words in each component and computing the most frequent component, we categorized each tweet in the corpus as "belief-speaking" or "truth-seeking". We refer to tweets that contain neither component as "neutral". In addition, we categorized each Twitter account in the corpus by party affiliation to investigate differences and similarities between Democrats and Republicans.

Of the tweets contained in our corpus, 7.2% were categorized as belief-speaking, 15.0% as truth-seeking and 1.1% as both (i.e., the number of belief-speaking and truth-seeking associated words in the tweet was the same). We note that the difference between the frequency of belief-speaking and truth-seeking tweets has to be interpreted with caution, since the word lists of the two concepts do not have the same size (38 for belief-speaking and 60 for truth-seeking), which at least partly explains the over-representation of truth-seeking.

For comparison, we also identified the two honesty components in historic articles from the *New York Times* for three text categories: "opinion", "politics" and "science" (see Section "New York Times corpus" in the Methods for details). We found that truth-seeking is over-represented in the "science" cluster with 30.1% as compared to 20.3% and 20.9% in the "politics" and "opinion" categories, respectively. Belief-speaking, on the other hand, is over-represented in the "opinion" category with 13.2%, as compared to 7.1% in the "science" and 6.7% in the "politics" category, respectively. The analysis of *New York Times* content helped to test the validity of our dictionaries because truth-seeking and belief-speaking predominated in science and opinion, respectively – exactly as would be expected.

To contextualise the honesty components identified in the tweets with existing similar measures, we investigated their relation to text features such as authenticity [30], analytic language [31] and a moral component reflecting judgemental language [32], as well as emotion valence, each measured using LIWC 2022 [33] (see Section "LIWC text analysis" in the Methods for details). We calculated scores for each of these components for every tweet in the corpus. A summary of average values broken down by honesty component is given in Extended Data Table 2. Belief-speaking showed higher sentiment averages than truth-seeking, especially for negative emotions, confirming that belief-speaking discourse is often affectively more charged in comparison to truth-seeking. The "analytic" category was present in a larger share of truth-seeking than belief-speaking tweets, but even more prevalent in neutral tweets that contain neither belief-speaking nor truth-seeking language. Scores for the "authentic" category were slightly higher for the belief-speaking component while scores for the "moral" category were highest in neutral speech, and lowest in the truth-seeking component. Generally, variation of the prevalence of analytic, authentic, moral, positive emotional and negative emotional language within each honesty component was much higher than between honesty components. In summary, these analyses show that belief-speaking and truth-seeking are uniquely identifiable and do not overlap greatly with existing related measures of text features.

To explore the context in which belief-speaking and truth-seeking language is used in public facing conversations of politicians, we analyzed the corpus of Tweets of U.S. Congress members to produce textual scatterplots [34] (see Sections "Topic modelling" and "Word and topic keyness analysis" in the Methods for details). Figure 1 **A** depicts distinguishing terms on a two-dimensional plot, with the x- and y-axes representing "party" and "component" respectively. Each dot is a unigram from the corpus, and their colour is associated with party keyness. Finally, the closer to a corner a dot is, the more it characterizes that particular component and party dimension.

We see that Republican belief-speaking keywords, situated in the top-left corner, often refer to political opponents ("joe", "biden", "obama", "pelosi"), whereas truth-seeking keywords from the same party are linked to controversial matters ("impeachment", "china") or economic issues ("growth", "tax", "unemployment"). On the right hand side of the figure, we find that Democrat belief-speaking tweets often regard people ("woman", "friend", "child") or everyday politics ("community", "vote", "register"), whereas truth-seeking texts almost exclusively concern scientific topics ("climate", "health", "testing", "covid-19"). Notably, both parties present topics scattered across the two components. For Republicans, this is the case for political orientations ("liberal", "dems", "conservative" in belief-speaking tweets, "democrats" in truth-seeking tweets). Democrats, on the other hand, use both components to discuss social issues such as violence and racism ("gun", "justice" in belief-speaking texts, "violence", "black" in truth-seeking texts).

3 Partisan and temporal dynamics of honesty components

Building on the topic modelling, we investigated the difference between belief-speaking and truth-seeking in communication about controversial topics in U.S. politics, such as abortion, climate change, or the death penalty, and how this differs by party. The selection of controversial topics presented here is inspired by other research in the same area, e.g. [35] and current research topics of non-partisan think-tanks, e.g. [36]. In Fig. 1 **B** and **C** we show the relative frequency $f = N(\text{component}|\text{party}|\text{topic})/N(\text{party}|\text{topic})$ of belief-speaking and

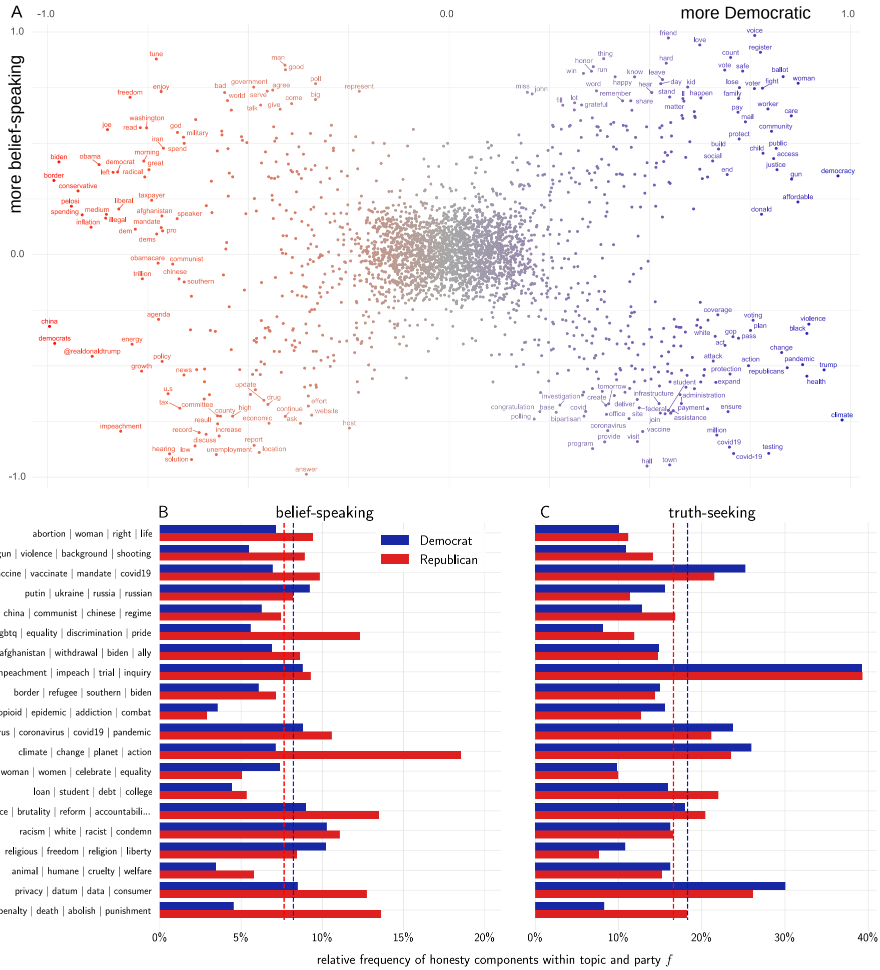


Figure 1 Panel **A** depicts the distribution of keywords on a two-dimensional scatterplot. Every term is a dot with two coordinates associated with party (x-coordinate) and component (y-coordinate) keyness. Each coordinate represents a Scaled F-Score (SFS) value ranging from -1 to 1. The word color is associated with the party keyness. We only show word labels where $SFS > 0.65 | SFS < -0.65$ for readability reasons. Panels **B** and **C** show belief-speaking and truth-seeking in Democratic (blue) and Republican (red) tweets for a range of hand-picked controversial topics. The bars indicate the relative frequency f of tweets associated with the two honesty components within each topic. Dashed lines indicate the mean prevalence of belief-speaking and truth-seeking for each party in the full corpus. We note that direct comparisons between the frequency of belief-speaking and the frequency of truth-seeking within a topic are not warranted, because of the different sizes of the dictionaries that measure each component.

truth-seeking within a given topic for members of the Democratic and Republican parties, respectively. There is a large variance in the amount of belief-speaking and truth-seeking used between the topics: Topics such as impeachment, racism and Putin / Ukraine show a large amount of belief-speaking in both parties, whereas topics such as the opioid epidemic show little. Similary, for truth-seeking the topics vaccines, impeachment and privacy show a large share of this component for both parties whereas abortion shows little.

There are also marked differences in the balance of belief-speaking and truth-seeking within a topic and between the parties. The topics of climate change and the death penalty have the largest difference in belief-speaking, with tweets by Republicans containing more than double the amount of belief-speaking than those by Democrats. On the other hand, for the topic of women / equality the Democrats show more belief-speaking. Differences in the prevalence of truth-seeking between the parties are generally smaller than differences in belief-speaking, with the exception of the death penalty, where tweets by Republicans also contain more than double the amount of truth-seeking as compared to tweets by Democrats.

We next examined the temporal trends of the two honesty components. To arrive at a finer-grained picture of the variability of these components between individual politicians, we calculated the share of belief-speaking and truth-seeking tweets for each individual politician. Fig. 2 **A** and **B** shows how the distribution of the share of belief-speaking tweets shifted between the first (2011-2013) and last (2019-2022) four years of tweets contained in the corpus.

For both parties, the share of belief-speaking considerably increased from mean 4.3% [0%; 10.4%] 95% CI and 4.9% [0%; 13.1%] for Democrats and Republicans, respectively to 6.9% [2.0%; 15.0%] and 6.0% [1.4%; 13.6%]. Similarly, we see an increase of the share of truth-seeking tweets from 8.0% [0%; 16.3%] and 8.6% [0%; 20.0%] for Democrats and Republicans, respectively, to 13.6% [5.8%, 24.7%] and 12.2% [4.4%; 25.7%]. This overall increase in both belief-speaking and truth-seeking also becomes apparent in Fig. 2 **E** and **F**, and is especially pronounced after the presidential election in late 2016.

This parallel increase for both belief-speaking and truth-seeking could reflect the fact that in recent years, topics concerning fake news have become increasingly central to political discourse [37], resulting in opposing claims and counterclaims [e.g., Donald Trump routinely accused mainstream media such as the New York Times of spreading "fake news", 4]. Whereas those claims represented mainly belief speaking, they were accompanied by increasing attempts by the media, and other actors, to correct misinformation through truth-seeking discourse.

4 Relation of honesty components to information trustworthiness

To test our hypothesis that belief-speaking is associated with dissemination of misinformation, we analyzed the relation of belief-speaking and truth-seeking to the quality of the information that is being relayed.

To assess information quality, we examined links to websites external to Twitter that were shared by the accounts. We followed an approach employed

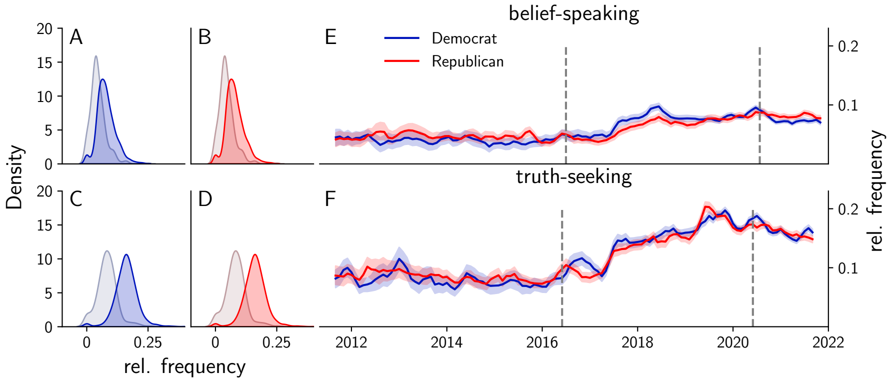


Figure 2 Number of tweets by members of the U.S. Congress for the period 2011 to 2021 shown separately for members of each party. **A** and **B** distributions of the relative frequency of belief speaking tweets of individual members of the democratic and republican parties for the years 2011 to 2013 (grey) and 2019 to 2022, respectively. **C** and **D** distributions of the relative frequency of truth seeking tweets. **E** and **F** relative frequency of belief speaking and truth seeking tweets over time. Timelines have been smoothed with a rolling average of three months. The 95% confidence intervals were computed with bootstrap sampling over 1,000 iterations. Dashed vertical lines indicate dates of presidential elections in 2016 and 2020.

by similar research in this domain [38, 39] and used a trustworthiness assessment by professional fact checkers of the *domain* a link points to. We used the NewsGuard information nutrition data base [40] as well as an independently compiled data base of domain trustworthiness labels [41] (see Sections "News-Guard nutrition labels" and "Independent list of untrustworhty sources" in Methods for details).

The NewsGuard data base as of the beginning of March 2022 indexed 6,860 English language domains. Each domain is scored on a total of 9 criteria, ranging from "doesn't label advertising" to "repeatedly publishes false information". Each category awards a varying number of points for a total of 100. Domains with less than 60 points are considered "not trustworthy". The majority of indexed domains (63%) are considered trustworthy. After excluding links to other social media platforms (e.g., twitter.com, facebook.com, youtube.com and instagram.com) as well as links to search services (google.com, yahoo.com), the database covered between 20% and 60% of the links posted by members of the U.S. Congress, with a steadily increasing share of links covered over time and no difference in coverage between the parties — see also Extended Data Fig. 3.

For each account, we calculated the average NewsGuard score of links shared in posts by that account $\langle S_{\mathrm{NG}} \rangle$, as well as the share of tweets that contain belief-speaking and truth-seeking words T_b and T_t. Fig. 3 **A** and **B** show $\langle S_{\mathrm{NG}} \rangle$ over T_b and T_t, respectively, for each account associated with the Democratic and Republican party. The scatterplot indicates a correlation of a high share of belief-speaking tweets with a lower mean NewsGuard score for Republicans but not for Democrats.

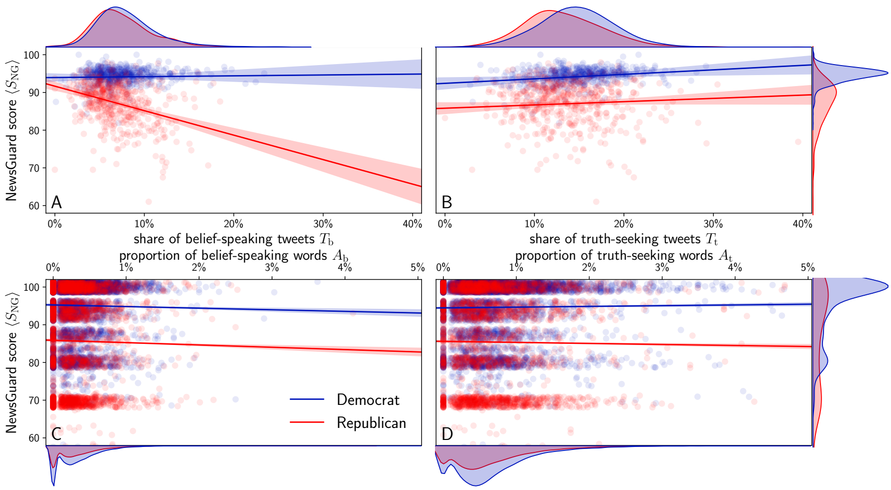


Figure 3 Relation of information quality with belief-speaking and truth-seeking. **A** and **B** Average NewsGuard score $\langle S_{\mathrm{NG}} \rangle$ of links posted by individual U.S. Congress members over the proportion of tweets that are categorized as belief-speaking (T_b) and truth-seeking (T_t), respectively. The lines and shaded areas indicate NewsGuard score predictions and 95% confidence intervals from a linear regression model. **C** and **D** NewsGuard score S_{NG} over the proportion of belief-speaking (A_b) and truth-seeking (A_t) words found in texts scraped from the links posted by U.S. Congress members. The lines and shaded areas indicate NewsGuard score predictions and 95% confidence intervals from an ordinary least squares regression model (see Eq.(2)). The scatter plot shows only 5% of the available data and vertical jitter was applied to visually separate data points. Note that we truncated the y-axis in panels **C** and **D** at $S_{\mathrm{NG}} = 60$. The full data is shown in Extended Data Fig. 4. Marginal distributions on the sides show the kernel density estimation over the full data on the respective axes, separated by party.

To further investigate this relationship, we fitted a linear regression model following Eq.(1) to predict $\langle S_{\mathrm{NG}} \rangle$ depending on T_{b}, T_{t}, party P and their interaction terms, adjusted for an account's followers and overall number of tweets (see Section "Regression" in the Methods for details).

The analysis yielded a significant effect for the interaction between Republican and belief-speaking (coefficient -68.3 [-85.9; -50.7], $p < 0.001$, $t = -7.6$), and the share of truth-seeking tweets independent of party (coefficient 12.1 [2.9; 21.3], $p = 0.01$, $t = 2.6$) — see Extended Data Table 5 for the full regression statistics. Therefore an increase in T_{b} of 10% predicted a decrease in $\langle S_{\mathrm{NG}} \rangle$ of 6.8, but only for members of the Republican party. Similarly, an increase in T_{t} of 10% predicted an increase in $\langle S_{\mathrm{NG}} \rangle$ of 1.2 for both parties. Predictions of the NewsGuard score depending on the share of belief-speaking and truth-seeking separated between the parties are shown as lines in panels **A** and **B** of Fig. 3, respectively.

To exclude a dependence of these results on use of the NewsGuard data base, we validated this analysis with an independently collected list of news outlet reliability from academic and fact-checking sources. Results are reported in the SI and are consistent with results reported in the main text.

Finally, we wanted to know whether the content of belief-speaking and truth-seeking words in the texts found at the websites the Congress members linked to was also indicative of low information quality. To this end, we attempted to scrape the text of all linked websites (see Section "News article collection" in the Methods). We successfully collected text from about 78% of links, with no difference in coverage for links with a low or high NewsGuard score. We excluded texts with less than 100 words and only retained one copy of the text in the case that multiple tweets contained links to the same website. In addition, we excluded all links that were posted by members of both parties, such that every link had a unique party designation. This resulted in a total of 162,508 unique news texts.

We investigated the dependence of the NewsGuard score associated with the domain the text was scraped from (S_{NG}) on the proportion of words contained in the belief-speaking dictionary (A_b) and the truth-seeking dictionary (A_t) that were found in the text (rather than in the original tweet). Again, we fitted an ordinary least squares model to predict S_{NG} depending on A_b, A_t, party and the interaction terms (see Eq.(2) and Section "Regression" in the Methods).

We show both the data for individual links and the model predictions for A_{b} and A_{t} in Fig. 3 **C** and **D**, respectively. We found a significant inverse relationship between A_b and S_{NG} for both parties (coefficient -42.56 [-65.19; -19.94], $p < 0.001$, $t = -3.69$). The interaction between party (Republican) and S_{NG} was not significant. For Democrats, we found a significant relationship between A_t and S_{NG} (coefficient 19.37 [7.39; 31.34], $p = 0.002$, $t = 3.17$), whereas for Republicans we found a significant inverse relationship (coefficient -46.70 [-66.01; -27.38], $p < 0.001$, $t = -4.74$). See Extended Data Table 6 for the full regression statistics.

Compared to the results from the analysis of the honesty components in tweet texts in relation to information quality, in the article text analysis we reproduced the negative relationship between belief-speaking and NewsGuard score for Republicans. Different from the previous analysis, we also found a negative relationship for belief-speaking for Democrats. For truth-seeking, we reproduce the positive relationship between more truth-seeking and a higher NewsGuard score - but only for Democrats. For Republicans, we find a significant inverse relationship.

5 Conclusions

We curated two dictionaries that captured the distinction between an evidence-based conception of honesty (truth-seeking) and a conception based on intuition, subjective impressions, and feelings (belief-speaking). We confirmed the diagnosticity of the dictionaries by showing that belief-speaking prevailed in opinion pieces in the New York Times but not in their science section, whereas the reverse occurred for truth-seeking. Our analysis of public political discourse of members of the U.S. Congress, represented by their tweets, linked greater use of belief-speaking to an increased likelihood of sharing low-quality content. This relationship was particularly pronounced for Republicans' tweets, although it was present for both parties when considering the target text that was shared [42]. While belief-speaking does not necessarily imply dishonesty and therefore not every expression of belief is dishonest, on average, the more a person resorts to belief-speaking, the more likely they are to be dishonest.

Several recent analyses of the American public's information diet have shown that conservatives are more likely to encounter and share untrustworthy information than their counterparts on the political left [38, 43, 44]. Several reasons have been put forward for this apparent asymmetry, for example that partisans are motivated to share derogatory content towards the political outgroup [45]. Because greater negativity towards Democrats is mostly found in lower-quality outlets, conservatives may disproportionately share untrustworthy information because it is satisfying a need for outgroup derogation [46]. Our analysis offers another explanation, namely that the public is sensitive to cues provided by the political elites which, as we have shown here, also differ considerably in the accuracy of content that they share on social media.

Our analysis furthermore identified belief-speaking as a "gateway" rhetorical technique for the sharing of low-quality information. The more politicians appeal to beliefs and intuitions, rather than evidence, the more likely they are to share low-quality information. This, however, begs the question *why* belief-speaking is a preferred means to spread low-quality information. Given that belief speaking was found to be associated with greater negative emotion (see Section 2), and given that lower-quality information tends to be slanted towards negativity [47], one option may be that belief-speaking is the result of Republican politicians' desire to derogate Democrats, as suggested by [46]. On that view, negative emotional content should be a mediator of the association between belief-speaking and low quality of shared content. Conversely, if belief-speaking were instrumental in the sharing of low-quality content for other reasons, then it should mediate the association involving negative emotionality. We report two competing mediation models in the SI (Section "Mediation Analysis"). While they cannot rule out any of the theoretical frameworks, the analyses suggest the latter hypothesis is in a better position to explain the mediating effect on the spread of low-quality news. Within this framework, belief-speaking is not subordinate to the need to express negative emotion but is an exogenous gateway to low-quality information. This supports the hypothesis that the current onslaught of online misinformation is in part driven by a new ontology of truth and honesty that has replaced reliance on evidence with the invocation of subjective belief.

6 Methods

6.1 U.S. Congress Member tweet corpus

A corpus of contemporary text with political content in English was created by scraping tweets by members of both houses of the U.S. Congress on March 15 and March 16 2022. To build the corpus, lists of Twitter handles of members of congress were collected for the 114th, 115th, 116th, and 117th Congress. For the 114th and 115th Congress, only handles of senators were available. For the 116th and 117th Congress, Twitter handles were available for both houses of Congress. This resulted in a total of 1,143 unique Twitter handles, which includes Congress member staff and Congress member campaign accounts.

For each of the Twitter handles, metadata of the Twitter account was collected on March 16, 2022 via the Twitter API v2 using the Python package twarc [48]. Metadata included the account's handle, user name, creation date, location, user description, number of followers, number of accounts followed, and tweet count. Out of the 1,143 accounts, 108 were not accessible because they had been deleted, suspended, or set to "private".

To build the text corpus, all tweets posted by the collected Twitter accounts starting from November 6, 2010 and up to March 16, 2022 were collected, using academic access to the Twitter API. Earlier tweets all the way back to 2006 could be retrieved, but we chose 2010 as the earliest date due to changes in the design of retweeting in the Twitter platform at that time. The retweet button was introduced in November 2009 (previously retweeting was done by hand), and it took approximately a year for users to start using it consistently. Furthermore, the prominence of Twitter in U.S. politics emerged later, especially since 2012. The language estimate facility of the Twitter API was used to remove non-English posts (58,326 tweets). The resulting corpus consisted of a total of 2,588,559 tweets, of which 1,523,050 were original tweets, 266,737 were quote tweets, 248,511 were replies and 552,892 were retweets. Note that quoting, replying and retweeting are not exclusive categories. We removed retweets from the corpus because they do not constitute original content. The number of tweets was low ($< 200,000$) in the years 2010 to 2017 and increased sharply in the later years until it exceeded 700,000 tweets in 2021. We removed exact matches and included only tweets with more than 10 words (210,867 tweets). Furthermore, we excluded tweets from the year 2010 from subsequent analyses because their number was relatively small (778 tweets). The final corpus contained 1,824,022 tweets. Next to the tweet text, the corpus contained the tweet creation date as well as a unique identifier of the account that posted the tweet. The identifier permitted linkage to the metadata collected about the user accounts, such as party affiliation.

In addition to the perspective of individual tweets taken in the analysis presented in Sec. 3, we also considered the perspective of individual links taken in the analysis presented in Sec. 4. For this analysis, we only considered tweets that contained at least one link (1,339,442 tweets). Because a single tweet can contain more than one link, we expanded the dataset such that every entry referred to a single link, transferring the tweet-level honesty-component labels to the individual links. This resulted in a total of 1,437,973 links. From each link, we extracted the domain the link pointed to. If the link was shortened using a link-shortening service such as bit.ly, we followed the link to retrieve the full domain name. The domains were then matched against the NewsGuard domain trustworthiness data base as well as the independently compiled list of trustworthiness labels (described in Section 6.8 and Section "Independent list of untrustworthy sources" in the SI).

6.2 Honesty component keywords and validation

We relied on keywords to identify the relevant subsets of tweets that involved the presumed three components of honesty. The steps followed to generate and validate keywords regarding the components are illustrated in Extended Data Figure 1.

Initially, three lists of keywords, one for each honesty component, were generated by the researchers involved in this article. The aim was to capture linguistic cues whose presence might signal that one of the components has been enacted by the speaker. To illustrate, initial keywords for truth-seeking included terms such as "reality", "assess" "examine", "evidence", "fact", "truth", "proof", and so on. For belief-speaking, initial keywords were terms such as "believe", "opinion", "consider", "feel", "intuition", or "common sense". Finally, for fostering understanding, the initial choices included "explain", "inform", "insight", "learn", "realize", "understand", and so on. The lists were expanded computationally using a combination of the *fasttext* library [49] and colexification networks [50, 51]. Colexification networks connect words in a language based on their common translations to other languages, thus signalling words that can be used to express multiple concepts. For example, the words "air" and "breath" are considered to be colexifications because they both translate into the same word in multiple languages ["sukdun" in Manchu, "vu:jnas" in Kildin Sami, "jind"in Nenets 52]. Colexification networks have been used recently to study emotion structures in language [53] and are predictors of word meaning ratings [50]. Including colexification networks in lexicon expansion gives word lists with a better trade-off between precision and recall [51] than previous approaches using *wordnet* or word embeddings, such as *empath*.

We subsequently filtered the expanded lists to remove duplicates, overlapping terms appearing in more than one list, and lemma inflections (i.e., "convey", "conveys", "conveyed"). The keywords were then used to identify texts relevant to the three presumed components of honesty.

To verify the keywords, we first scanned the Corpus of Historical American English (COHA) [54] looking for sentences that contained at least one keyword from the three lists. We collected 200 random matches for each component, for a total of 600 sentences that spanned the period from 1810 to 2009. Because the presence of a component is not guaranteed by the simple occurrence of a relevant keyword, one of the authors manually validated their existence in each of the texts. To do this, we first divided the sample into three different subsets, one for each component, based on the keyword(s) they contained. Because each sentence might have contained different keywords belonging to different components, it was possible for a sentence to appear in more than one subset.

Due to the generally low frequency of the three components, we decided to refine the keyword lists by removing less pertinent tokens (i.e., those yielding few matches). Moreover, we introduced linguistic patterns formed by a combination of parts-of-speech tagging and lemmas. This increased the precision of the filtering process so that, when the corpora were scanned for keywords matches, we could receive more component-related texts.

Next, we used the refined keyword lists to filter the Twitter corpus and generate a sample for manual validation, using the spaCy package in Python [55], which searched for keywords and linguistic pattern matches. The latter was represented by combinations of POS tags and lemmas (i.e. "I/We + VERB"). The resulting sample was split into three different subsets, one for each component of honesty, each of which contained thousands of tweets. To avoid authorial as well as chronological bias and to reduce the time allotted to manual text validation, we first sampled 200 tweets from each component.

Based on the validation results, we further improved the precision of the keyword filtering process. Keywords that yielded few matches were deleted, thus shortening the general length of the lists. Terms that returned numerous matches were further disambiguated to eliminate noise. As an example, the script was instructed to avoid instances where "look" was followed by "forward" because that particular bigram was frequent but not relevant to any of the components of interest. The final list of keywords used can be found in Table 1.

We then obtained a second sample from the Twitter corpus using the new keywords and patterns in spaCy. Similar to the first sample, the second sample was split into three different subsets, one per component, and randomized to avoid possible biases. Then, a sub-sample of 200 tweets for each component was collected for a further round of manual validation by one of the authors. Between the first and second tweet sample, the presence of the components roughly doubled.

In a final validation step, we examined inter-rater agreement. The three subsets extracted from the second Twitter round were each assigned to three external coders for annotation, meaning that each coder validated one component. Coders indicated for each item whether or not the texts contained the honesty component of interest. The annotation guide used by the annotators can be found in the online supplement. The same three subsets were also annotated by one of the authors. We then measured inter-rater agreement using the krippendorffsalpha package [56] in R, which also provides 95% CI ranges.

6.3 Identification of honesty components in text

Text gathered from social media may contain noise. This necessitates a cleaning step to filter out undesirable but frequent occurrences of some text elements before proceeding with the identification of different honesty components in text. Specifically, we first removed URLs and replaced user handles on Twitter with the word *user*.

To identify honesty components in tweet texts, we split tweet texts into individual tokens (words), which are matched against the belief-speaking and truth-seeking dictionaries. Matches are counted for each tweet and a tweet is assigned a label according to the honesty component which has the most matches. In case of a tie, the tweet is assigned both labels. A total of 131,626 tweets are labelled as "belief-speaking", 273,192 tweets are labelled as "truth-seeking" and 19,653 Tweets are assigned to both components.

6.4 New York Times corpus

We retrieved data from the New York Times (NYT) through their archive API (https://developer.nytimes.com/docs/archive-product/1/overview). By iterating over the months since the founding of the newspaper in the 19^{th} century, we retrieved information on every article in the archive. The information returned by the API included the article title and an abstract that summarizes the article content, as well as additional metadata such as publication date and section of the paper. This approach is different to earlier research that used the NYT API to obtain a number of articles over time that contain certain terms, which does not yield any further text or ways to filter the data [57]. Because we needed text to identify honesty components in articles, the archive endpoint was more suitable than the term search function of the NYT API, despite not giving us the full text of all articles but only returning a summary. We extracted three distinct categories of content from the NYT corpus based on the sections identified in the metadata: (i) An *opinion* category which comprises opinion pieces such as "OpEds"; (ii) a *politics* category consisting of articles in the sections U.S., Washington, and World; and (iii) a *science* category which includes health, science, education, and climate articles. We chose these three clusters because we expected opinion articles to contain more belief-speaking, whereas we expected science articles to contain more truth-seeking. We expected articles in the politics cluster to fall in between. We retrieved a total of 809,271 articles consisting of 240,567 opinion articles, 518,123 politics articles, and 50,581 science articles.

6.5 Topic modelling

We performed the topic modelling using the Python package Bertopic [58]. Following a three-step approach, the package uses the Sentence-BERT (SBERT) framework to create the embeddings for each document, then uses the Uniform Manifold Approximation and Projection (UMAP) technique [59] to decrease the dimensionality of embeddings and identify clusters through HDBSCAN [60]. Finally, it creates topic representations using class-based term-frequency inverse-document-frequency (TF-IDF). We opted for Bertopic rather than more established techniques such as Latent Dirichlet Allocation

(LDA) because the former performs better when modelling short and unstructured texts as in the case of Twitter data when compared to the latter [61, 62]. In addition, the package also supports topic modelling variations; in our case, class-based topic modelling was used to check topic distribution across parties and honesty components. Since Bertopic relies on an embedding approach, data was only minimally preprocessed to keep the original sentence structure. This means we lemmatized the entire dataset to produce cleaner topic representations, and only removed URLs from the texts.

To prepare the data for class-based topic modelling, each tweet was categorized according to two variables: one representing the honesty component (belief-speaking, truth-seeking or neutral), the other representing both the parties and the components ("Dem-belief", "Rep-belief", and so on). Considering that the dataset contains more than 1.5 million tweets, we also applied thresholds to the topic modelling. The document minimum frequency was set to 200 in order to reduce the number of small topics. The number of neighboring sample points used when making the manifold approximation was set to 100 to produce a more global view of the embedding structure. Finally, the minimum document frequency for the c-TF-IDF was set to 50 to reduce the topic-term matrix size and decrease memory-related issues during the computation. With these settings, the model was able to identify 377 topics.

To check whether this was an optimal number of topics, we used ldatuning [63], an R package that trains multiple models and calculates validation metrics. Despite the fact that *ldatuning* does not employ embeddings but Latent Dirichlet allocation and that the data it modelled was preprocessed by removing stopwords and irrelevant text (numbers, unknown characters, URLs, Twitter handles), it indicated 300 as an optimal number of topics for the dataset, thus converging towards the Bertopic results.

6.6 Word and topic keyness analysis

The scatterplot in panel **A** of Figure 1 was produced following the same approach used in Scattertext [34], a Python package designed to illustrate words and phrases that are more characteristic of a category than others. Starting from raw frequencies, we calculated for each word both the relative frequency across categories (i.e. parties) as well as the relative frequency within a category. These values are defined by the package author as precision and recall, respectively. The former represents the discriminative power of a word regardless of its frequency in a certain category. The latter is the percentage frequency with which a word appears in a certain category. We then transformed these two values using a normal cumulative distribution function to scale and standardize the scores. Next, we calculated the harmonic mean of the normal CDF-transformed scores, obtaining a Scaled F-Score (SFS), which ultimately is the term scoring metric used to identify distinguishing words. Since we were comparing different categories (i.e., Democrats vs. Republicans), we calculated SFS of words for both of them, obtaining two values, SFS^x and SFS^y. Lastly, we extracted a final SFS that ranges from -1 to 1 using the following formula:

$$SFS = 2 \cdot \left(-0.5 + \begin{cases} SFS^x & \text{if } SFS^x > SFS^y, \\ 1 - SFS^y & \text{if } SFS^x < SFS^y, \\ 0 & \text{otherwise} \end{cases} \right).$$

We repeated the same operation on word frequencies across honesty components, so that each term finally had two SFS: one for party distribution and one for component distribution. These two values were used as coordinates for the scatterplot shown in panel **A** of Fig. 1.

On the other hand, the class-based topic modelling allowed us to identify how topics are represented across parties and honesty components. By default, Bertopic assigns each document to a single topic, thus producing frequency tables of topics within pre-determined classes. Relying on the class-based topic modelling, we calculated how many belief-speaking or truth-seeking tweets were assigned to a particular controversial topic within a party, as shown in panels **C** and **D** of Figure 1. Finally, we produced panels **E** and **F** of Fig. 1 by calculating the relative frequency of specific controversial topics within each component and party. To do this, we obtained a component share value for each topic by dividing its frequency in a certain component by its overall frequency in the entire dataset. Then, we repeated the same procedure including party as a variable in order to have the component share value of a topic within the two parties. Lastly, we subtracted the two values to highlight how parties differ in honesty-speech when talking about controversial matters.

6.7 LIWC text analysis

We explored the content of the tweet texts within the two honesty components using the Linguistic Inquiry and Word Count (LIWC) program [33]. LIWC is a text processing software that has been continuously developed for more than two decades and computes several indicator variables from text based on word lists generated by psychologists and validated in various experiments — similar to our approach in generating the word lists for the belief-speaking and truth-seeking word lists.

With the Beta version of LIWC-2022 software (https://www.liwc.app/), we computed the scores for each tweet text for the following LIWC categories: emotion (divided into positive and negative), authenticity, analytic, and moral. Positive and negative emotion scores are related to the sentiment polarity of a text. Authenticity indicates to what extent the language used is perceived as honest and genuine [30]. Analytic is linked to logical and formal thinking [31]. Finally, moral reflects the judgmental language expressed by positive or negative evaluation of someone's behavior or character [32]. The scores provide an efficient summary of those attributes in each text.

Scores broken down by the two honesty components "belief-speaking" and "truth-seeking" as well as for "neutral" tweets are given in Extended Data Tab. 2. In addition, we show the time-development of the scores broken down by honesty component for positive and negative emotions in Extended Data Fig. 2 and for the "analytic", "authentic" and "moral" components in supplementary figure 1.

6.8 NewsGuard nutrition labels

Following precedent [38, 39], we use source trustworthiness as an estimator for the trustworthiness of an individual piece of shared information. We use nutrition labels provided by NewsGuard, a company that offers professional fact checking as a service and curates a large data base of domains. The trustworthiness of a domain is assessed in nine categories, each of which awards a number of points: Does not repeatedly publish false content (22), gathers and presents information responsibly (18), regularly corrects or clarifies errors (12.5), handles the difference between news and opinion responsibly (12.5), avoids deceptive headlines (10), website discloses ownership and financing (7.5), clearly labels advertising (7.5), reveals who is in charge, including any possible conflicts of interest (5), the site provides names of content creators, along with either contact or biographical information (5).

NewsGuard categorizes domains with a score of 60 or higher as "generally adheres to basic standards of credibility and transparency" [40]. Similar to [64], we use this value as a threshold below which we categorize a domain and the link pointing to it as "not trustworthy".

After excluding links to other social media platforms (e.g., twitter.com, facebook.com, youtube.com, and instagram.com) as well as links to search services (google.com, yahoo.com), the database covers between 20% and 60% of the links posted by members of the U.S. Congress, with a steadily increasing share of links covered over time — see also Extended Data Fig. 3 **A**.

6.9 Regression

We performed a range of regression analyses to quantify the relationship between various manifestations of honesty components and information quality. For the predictions shown in Fig. 3 **A** and **B** we fitted the following linear model for individual members of the U.S. Congress:

$$\langle S_{\mathrm{NG}} \rangle \sim T_b + T_t + P \times T_b + P \times T_t + \log(N_f) + \log(N_t) \ . \qquad (1)$$

Here, $\langle S_{\mathrm{NG}} \rangle$ is the average NewsGuard nutrition score of a member of the U.S. Congress, calculated by averaging all NewsGuard scores of domains a given Congress member linked to in a post on Twitter. T_b and T_t are the proportion of tweets by individual Congress members that contain belief-speaking or truth-seeking, respectively. P is the party designation which can be "Republican" or "Democrat". We correct for the number of followers N_f and overall number of tweets at the time of data collection N_t by the given account.

For the predictions shown in Fig. 3 **C** and **D**, we fitted the following model for individual links

$$S_{\mathrm{NG}} \sim A_b + A_t + P \times A_b + P \times A_t \ . \tag{2}$$

Here, S_{NG} is the NewsGuard score of the domain the link points to. A_b and A_t are the proportion of belief-speaking and truth-seeking words within the article retrieved from the link.

We fitted each regression model using an ordinary least squares fitting approach from the Python package statsmodels [65]. Regression results are reported in Extended Data Tables 5 and 6.

6.10 News article collection

Our corpus contained 219,787 unique links to news articles that were shared by members of Congress. We scraped the text using Newspaper3k [66], a python package for scraping and curating news articles. Some links were broken, restricted, or could not be scraped by the package, resulting in 78% of total scraping coverage. In addition, 1,914 links did not contain a valid NewsGuard score. We removed from the analysis articles that were shared by members of more than one political party (i.e., a link was shared either by Republicans or Democrats, but not both). This resulted in the removal of 1,046 links (0.78% of all remaining links). When broken down by trustworthiness (NewsGuard score < 60), the coverage for trustworthy links (N = 212,983) was 82%, and 79% for untrustworthy links (N = 4,892). A test of independence between the share of articles scraped for trustworthy and untrustworthy links did not reach statistical significance $\chi^2(1) = 2.81$, $p = 0.094$. The distribution of NewsGuard scores as well as the proportion of belief-speaking and truth-seeking words in each article is shown in Extended Data Fig. 4.

7 Data availability

The tweet IDs of the tweet texts and URLs of the articles analysed in this study are deposited in OSF under accession code https://doi.org/10.17605/OSF.IO/VNY8K. We provide code to download tweets from tweet IDs and article texts from article URLs and process the data in the code repository that accompanies this article https://doi.org/10.5281/zenodo.6826515.

Dictionaries of keywords associated with the three honesty components are deposited in OSF under accession code https://doi.org/10.17605/OSF.IO/VNY8K.

The independently compiled list of domain accuracy and transparency scores is deposited on GitHub under accession code https://doi.org/10.5281/zenodo.6536692.

The NewsGuard data base used to asses domain trustworthiness is commercially available from NewsGuard.

Aggregated values for information trustworthiness and honesty components for tweets and articles used to produce all figures in this article are deposited in OSF under accession code https://doi.org/10.17605/OSF.IO/VNY8K.

8 Code availability

Python 3.8.5 and R 4.2 were used to collect the data and perform the data analysis presented in this study. Data collection and analysis code is available under MIT license in a GitHub repository under accession code https://doi.org/10.5281/zenodo.6826515.

References

[1] S. Lewandowsky, in *Deliberate Ignorance: Choosing Not to Know*, ed. by R. Hertwig, C. Engel (MIT Press, 2020), pp. 101–117

[2] K. Müller, C. Schwarz, Fanning the flames of hate: Social media and hate crime. Journal of the European Economic Association **19**(4), 2131–2167 (2021)

[3] S. Lewandowsky, L. Smillie, D. Garcia, R. Hertwig, J. Weatherall, S. Egidy, R.E. Robertson, C. O'Connor, A. Kozyreva, P. Lorenz-Spreen, Y. Blaschke, M. Leiser, Technology and democracy: understanding the influence of online technologies on political behaviour and decision making. Tech. rep. (2020). https://doi.org/10.2760/709177

[4] S. Lewandowsky, M. Jetter, U.K.H. Ecker, Using the president's tweets to understand political diversion in the age of social media. Nature Communications **11**, 5764 (2020). https://doi.org/10.1038/s41467-020-19644-6

[5] J.T. Carmichael, R.J. Brulle, Elite cues, media coverage, and public concern: an integrated path analysis of public opinion on climate change, 2001–2013. Environmental Politics **26**(2), 232–252 (2017)

[6] B. Swire, A.J. Berinsky, S. Lewandowsky, U.K.H. Ecker, Processing political misinformation: comprehending the Trump phenomenon. Royal Society Open Science **4**, 160,802 (2017). https://doi.org/10.1098/rsos.160802

[7] B. Swire-Thompson, U.K.H. Ecker, S. Lewandowsky, A.J. Berinsky, They might be a liar but they're my liar: Source evaluation and the prevalence of misinformation. Political Psychology **41**, 21–34 (2020). https://doi.org/10.1111/pops.12586

[8] O. Hahl, M. Kim, E.W.Z. Sivan, The authentic appeal of the lying demagogue: Proclaiming the deeper truth about political illegitimacy.

American Sociological Review **83**, 1–33 (2018). https://doi.org/10.1177/0003122417749632

[9] Freedom House, Freedom in the world 2020. A leaderless struggle for democracy. Tech. rep. (2020)

[10] Freedom House, Freedom in the world 2021. democracy under siege. Tech. rep. (2021)

[11] M. Cantarella, N. Fraccaroli, R.G. Volpe, Does fake news affect voting behaviour? SSRN Electronic Journal (2020). https://doi.org/10.2139/ssrn.3629666

[12] S. Lewandowsky, U.K.H. Ecker, J. Cook, Beyond misinformation: Understanding and coping with the post-truth era. Journal of Applied Research in Memory and Cognition **6**, 353–369 (2017). https://doi.org/10.1016/j.jarmac.2017.07.008

[13] T. Wood, E. Porter, The elusive backfire effect: Mass attitudes' steadfast factual adherence. Political Behavior (2018). https://doi.org/10.1007/s11109-018-9443-y

[14] S. Lewandowsky, E.A. Lloyd, S. Brophy, When THUNCing Trumps thinking: What distant alternative worlds can tell us about the real world. Argumenta **3**, 217–231 (2018). https://doi.org/10.23811/52.arg2017.lew.llo.bro

[15] A.M. McCright, R.E. Dunlap, Combatting misinformation requires recognizing its types and the factors that facilitate its spread and resonance. Journal of Applied Research in Memory and Cognition (2017). https://doi.org/10.1016/j.jarmac.2017.09.005

[16] B. Cooper, T.R. Cohen, E. Huppert, E. Levine, W. Fleeson, Honesty in organizations: A systematic review and new conceptual framework. Academy of Management Annals (2021). https://doi.org/10.17605/OSF.IO/PCG7M

[17] S. Lewandowsky, in *Deliberate Ignorance: Choosing Not to Know*, vol. 29, ed. by S.F. Reports (MIT Press, 2020)

[18] A. Varshizky, Alfred Rosenberg: The Nazi *Weltanschauung* as modern gnosis. Politics, Religion & Ideology **13**, 311–331 (2012). https://doi.org/10.1080/21567689.2012.698977

[19] E. Voegelin, in *Thc collected works of Eric Voegelin (Volume 5)* (University of Missouri Press, 2000), pp. 21–73

[20] J.E. Uscinski, R.W. Butler, The epistemology of fact checking. Critical Review **25**, 162–180 (2013)

[21] M.A. Amazeen, Revisiting the epistemology of fact-checking. Critical Review **27**, 1–22 (2015). https://doi.org/10.1080/08913811.2014.993890

[22] H. Farrell, B. Schneier, Common-knowledge attacks on democracy. Tech. rep., Berkman Klein Center for Internet & Society (2018)

[23] G.C. Jacobson, Driven to Extremes: Donald Trump's Extraordinary Impact on the 2020 Elections. Presidential Studies Quarterly **51**, 492–521 (2021). https://doi.org/10.1111/psq.12724

[24] P. Barberá, A. Casas, J. Nagler, P.J. Egan, R. Bonneau, J.T. Jost, J.A. Tucker, Who leads? who follows? measuring issue attention and agenda setting by legislators and the mass public using social media data. American Political Science Review **113**(4), 883–901 (2019)

[25] S. Lewandowsky, M. Jetter, U.K. Ecker, Using the president's tweets to understand political diversion in the age of social media. Nature communications **11**(1), 1–12 (2020)

[26] L.K. Nelson, Computational grounded theory: A methodological framework. Sociological Methods & Research **49**(1), 3–42 (2020)

[27] A. Zapf, S. Castell, L. Morawietz, A. Karch, Measuring inter-rater reliability for nominal data–which coefficients and confidence intervals are appropriate? BMC medical research methodology **16**(1), 1–10 (2016)

[28] D. Riff, S. Lacy, F. Fico, *Analyzing media messages: Using quantitative content analysis in research* (Routledge, 2014)

[29] S. Geiß, Statistical power in content analysis designs: How effect size, sample size and coding accuracy jointly affect hypothesis testing–a monte carlo simulation approach. Computational Communication Research **3**(1), 61–89 (2021)

[30] M.L. Newman, J.W. Pennebaker, D.S. Berry, J.M. Richards, Lying words: Predicting deception from linguistic styles. Personality and Social Psychology Bulletin **29**(5), 665–675 (2003). https://doi.org/10.1177/0146167203029005010. PMID: 15272998

[31] J.W. Pennebaker, C.K. Chung, J. Frazee, G.M. Lavergne, D.I. Beaver, When small words foretell academic success: The case of college admissions essays. PloS one **9**(12), e115,844 (2014)

[32] W.J. Brady, M.J. Crockett, J.J.V. Bavel, The mad model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. Perspectives on Psychological Science **15**(4), 978–1010 (2020). https://doi.org/10.1177/1745691620917336. PMID: 32511060

[33] R.L. Boyd, A. Ashokkumar, S. Seraj, J.W. Pennebaker, The development and psychometric properties of LIWC-22. Tech. rep., University of Texas at Austin, Austin, TX (2022)

[34] J.S. Kessler, in *Proceedings of ACL-2017 System Demonstrations* (Association for Computational Linguistics, Vancouver, Canada, 2017)

[35] M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, M. Starnini, The echo chamber effect on social media. Proceedings of the National Academy of Sciences **118**(9), e2023301,118 (2021)

[36] https://www.pewresearch.org/ (2022). Accessed 2022-06-15

[37] A. Kozyreva, S. Lewandowsky, R. Hertwig, Citizens Versus the Internet: Confronting Digital Challenges With Cognitive Tools. Psychological Science in the Public Interest **21**, 103–156 (2020). https://doi.org/10.1177/1529100620946707

[38] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, Fake news on twitter during the 2016 us presidential election. Science **363**(6425), 374–378 (2019)

[39] G. Pennycook, Z. Epstein, M. Mosleh, A.A. Arechar, D. Eckles, D.G. Rand, Shifting attention to accuracy can reduce misinformation online. Nature **592**(7855), 590–595 (2021)

[40] I. NewsGuard. Rating process and criteria. Internet Archive, https://web.archive.org/web/20200630151704/https://www.newsguardtech.com/ratings/rating-process-criteria/ (2020). Accessed: 2022-04-20

[41] J. Lasser. List of domain accuracy and transparency scores v1.1 (2022). https://doi.org/10.5281/ZENODO.6536692. URL https://zenodo.org/record/6536692

[42] J. Lasser, S.T. Aroyehun, A. Simchon, F. Carrella, D. Garcia, S. Lewandowsky, Social media sharing by political elites: An asymmetric american exceptionalism. arXiv (2022). https://doi.org/10.48550/ARXIV.2207.06313. URL https://arxiv.org/abs/2207.06313

[43] A.M. Guess, B. Nyhan, J. Reifler, Exposure to untrustworthy websites in the 2016 U.S. election. Nature Human Behavior **4**, 472–480 (2020).

https://doi.org/10.1038/s41562-020-0833-x

[44] A.M. Guess, J. Nagler, J. Tucker, Less than you think: Prevalence and predictors of fake news dissemination on Facebook. Science Advances **5**, eaau4586 (2019). https://doi.org/10.1126/sciadv.aau4586

[45] S. Rathje, J.J. Van Bavel, S. van der Linden, Out-group animosity drives engagement on social media. Proc. Natl. Acad. Sci. U. S. A. **118**(26) (2021)

[46] M. Osmundsen, A. Bor, P.B. Vahlstrup, A. Bechmann, M.B. Petersen, Partisan polarization is the primary psychological motivation behind political fake news sharing on twitter. Am. Polit. Sci. Rev. **115**(3), 999–1015 (2021)

[47] S. Soroka, P. Fournier, L. Nir, Cross-national evidence of a negativity bias in psychophysiological reactions to news. Proceedings of the National Academy of Sciences **116**, 18,888–18,892 (2019). https://doi.org/10.1073/pnas.1908369116

[48] E. Summers, I. Brigadir, S. Hames, H. Van Kemenade, P. Binkley, Tinafigueroa, N. Ruest, Walmir, D. Chudnov, Recrm, , Celeste, Hause Lin, A. Chosak, R. Miles McCain, I. Milligan, A. Segerberg, Daniyal Shahrokhian, M. Walsh, L. Lausen, N. Woodward, F.V. Münch, Eggplants, Ashwin Ramaswami, D. Hereñú, D. Milajevs, F. Elwert, K. Westerling, Rongpenl, S. Costa, , Shawn. Docnow/twarc: v2.10.4 (2022). https://doi.org/10.5281/ZENODO.6503180. URL https://zenodo.org/record/6503180

[49] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)

[50] A. Di Natale, M. Pellert, D. Garcia, Colexification Networks Encode Affective Meaning. Affective Science **2**, 99–111 (2021). https://doi.org/10.1007/s42761-021-00033-1

[51] A. Di Natale, D. Garcia, Lexpander: applying colexification networks to automated lexicon expansion. arXiv (2022). https://doi.org/10.48550/ARXIV.2205.15850. URL https://arxiv.org/abs/2205.15850

[52] A. François, in *From polysemy to semantic change: Towards a typology of lexical semantic associations*, 106 (John Benjamins Publishing Company, Philadelphia, PA, 2008), p. 163

[53] J.C. Jackson, J. Watts, T.R. Henry, J.M. List, R. Forkel, P.J. Mucha, S.J. Greenhill, R.D. Gray, K.A. Lindquist, Emotion semantics show both cultural variation and universal structure. Science **366**(6472), 1517–1522

(2019)

[54] M. Davies. The corpus of historical american english: 400 million words, 1810-2009. http://corpus.byu.edu/coha/ (2010)

[55] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017). To appear

[56] J. Hughes, *krippendorffsalpha: Measuring Agreement Using Krippendorff's Alpha Coefficient* (2021). URL https://CRAN.R-project.org/package=krippendorffsalpha. R package version 1.1-1

[57] M. Scheffer, I. van de Leemput, E. Weinans, J. Bollen, The rise and fall of rationality in language. Proceedings of the National Academy of Sciences **118**(51) (2021). https://doi.org/10.1073/pnas.2107848118. URL https://www.pnas.org/content/118/51/e2107848118. https://www.pnas.org/content/118/51/e2107848118.full.pdf

[58] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794 (2022)

[59] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)

[60] R.J. Campello, D. Moulavi, J. Sander, in *Pacific-Asia conference on knowledge discovery and data mining* (Springer, 2013), pp. 160–172

[61] R. Egger, J. Yu, A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. Front. Sociol. 7: 886498. doi: 10.3389/fsoc (2022)

[62] F. Alhaj, A. Al-Haj, A. Sharieh, R. Jabri, Improving arabic cognitive distortion classification in twitter using bertopic. International Journal of Advanced Computer Science and Applications **13**(1), 854–860 (2022)

[63] M. Nikita, M.M. Nikita. Package 'ldatuning' (2016)

[64] S. Bhadani, S. Yamaya, A. Flammini, F. Menczer, G.L. Ciampaglia, B. Nyhan, Political audience diversity and news reliability in algorithmic ranking. Nature Human Behaviour pp. 1–11 (2022)

[65] S. Seabold, J. Perktold, in *9th Python in Science Conference* (2010)

[66] L. Ou-Yang. newspaper: News, full-text, and article metadata extraction in python 3. advanced docs:

[67] M. Mosleh, D. Rand, Falsehood in, falsehood out: A tool for measuring exposure to elite misinformation on twitter. PsyArXiv (2021)

[68] Politifact. Politifact's guide to fake news websites and what they peddle. https://www.politifact.com/article/2017/apr/20/politifacts-guide-fake-news-websites-and-what-they/. Accessed: 2022-05-01

[69] R. Gallotti, F. Valle, N. Castaldo, P. Sacco, M. De Domenico, Assessing the risks of 'infodemics' in response to covid-19 epidemics. Nature Human Behaviour **4**(12), 1285–1293 (2020)

[70] Bufale. The black list. https://www.bufale.net/. Accessed: 2022-05-01

[71] Bufalopedia. Un catalogo di indagini e risorse antibufala. https://bufalopedia. blogspot.com/p/siti-creatori-di-bufale.html. Accessed: 2022-05-01

[72] Butac. The black list. https://www.butac.it/the-black-list/. Accessed: 2022-05-01

[73] B. News. Inside the partisan fight for your news feed. https://www.buzzfeednews.com/article/craigsilverman/ inside-the-partisan-fight-for-your-news-feed. Accessed: 2022-05-01

[74] C.J. Review. CJR index of fake-news, clickbait, and hate. http://web.archive.org/web/20210720140548/https://www. cjr.org/fake-beta. Accessed: 2022-05-01

[75] F.N. Watch. Fake news watch. https://web.archive.org/web/20180213181029/, http://www.fakenewswatch.com/. Accessed: 2022-05-01

[76] M.B.F. Check. Media bias fact check. https://mediabiasfactcheck.com/. Accessed: 2022-05-01

[77] M. Zimdars. My "fake news list" went viral. but made-up stories are only part of the problem. https://www.washingtonpost.com/posteverything/wp/2016/11/18/my-fake-news-list-went-viral-but-made-up-stories-are-only-part-of-the-problem/. Accessed: 2022-05-01

[78] J. Lasser. List of domain accuracy and transparency scores v1.1 (2022). https://doi.org/10.5281/ZENODO.6536692. URL https://zenodo.org/record/6536692

[79] D. Tingley, T. Yamamoto, K. Hirose, L. Keele, K. Imai, Mediation: R package for causal mediation analysis (2014)

9 Acknowledgments

10 Funding

This report was funded by the Templeton Foundation through a grant awarded to Wake Forest University. SL was also supported by funding from the Humboldt Foundation in Germany, and SL and DG are beneficiaries of the ERC Advanced Grant PRODEMINFO. JL was supported by the Marie Skłodowska-Curie grant No. 101026507.

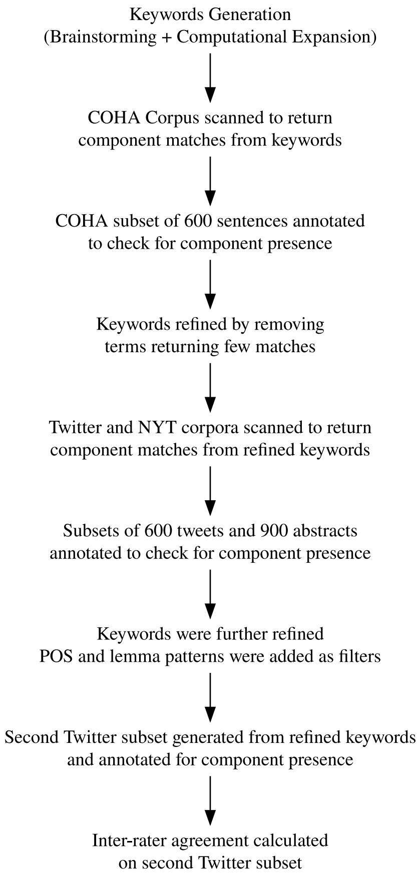
11 Author information

12 Ethics declarations

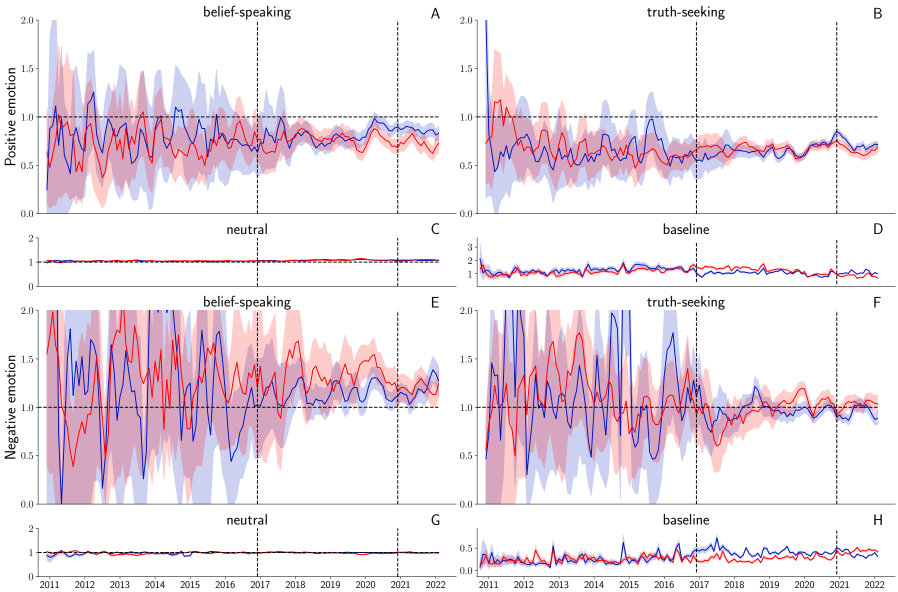
12.1 Competing interests

The authors declare no competing interests.

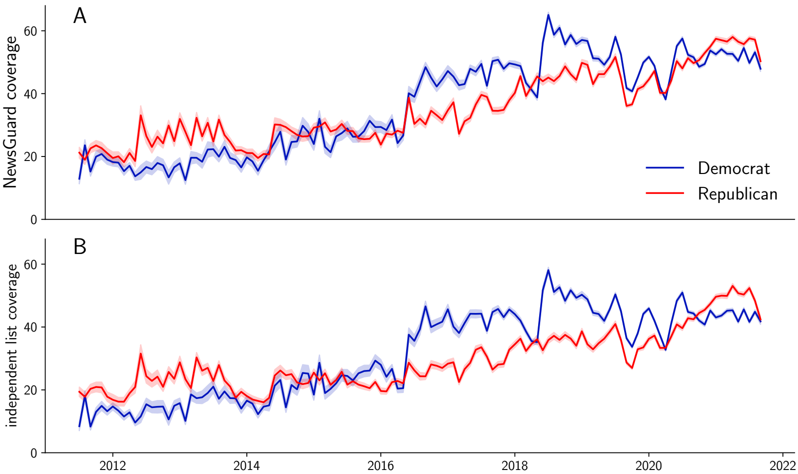
13 Extended data figures and tables



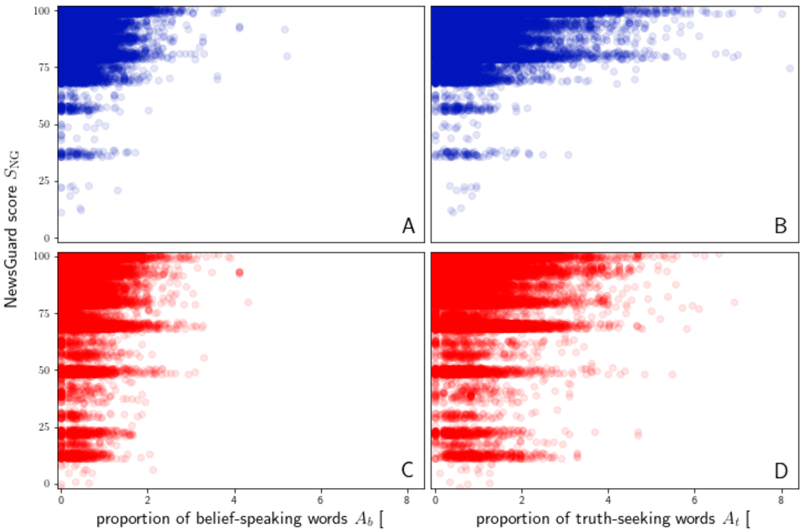
Extended Data Figure 1 Flowchart for keywords generation and validation



Extended Data Figure 2 Time-development of LIWC scores of positive and negative emotions in tweets of members of the U.S. Congress. Panels **A** and **B** show the score for positive emotions within the belief-speaking and truth-seeking components normalized by the overall score (baseline), respectively. The dashed horizontal line at 1.0 corresponds to prevalence equal to baseline. Red and blue lines correspond to tweets by Republicans and Democrats, respectively. Panel **C** shows the score within "neutral" (not belief-speaking or truth-seeking) tweets normalized by the baseline, and panel **D** shows the baseline. Panels **E**, **F**, **G** and **H** show the same information as panels **A**, **B**, **C** and **D**, but for the negative emotion score instead of the positive emotion score. The 95% confidence intervals (indicated by shading) were computed with bootstrap sampling over 1,000 iterations. Dashed vertical lines indicate dates of presidential elections in 2016 and 2020. Timelines are smoothed, using a rolling average over three months.



Extended Data Figure 3 Share of links posted by accounts belonging to members of the U.S. Congress pointing to domains indexed in **A** the NewsGuard data base and **B** the independently compiled list.



Extended Data Figure 4 **A** and **B** NewsGuard score S_{NG} over proportion of belief-speaking words A_b and truth-seeking words A_t in articles scraped from links posted by Democrat members of the U.S. Congress, respectively. **C** and **D** show the same information for Republican members of the U.S. Congress.

Belief-speaking	Truth-seeking	Fostering understanding
admittedly	actually	acknowledge
basically	analyze	admit
believe	assert	affirm
certainly	assertion	agree
clearly	assess	appreciate
confide	claim	approve
consider	contemplate	argue
definitely	contention	ask
doubtless	correct	certify
envisage	correction	concede
feel	determine	confirm
frankly	estimate	consider
guess	evaluate	convey
honestly	evidence	convince
indeed	examine	declare
know	exploration	deem
no doubt	explore	demonstrate
observe	fact	discover
obvious	find	discuss
obviously	genuinely	dispute
of course	hint	endorse
opinion	improvement	exhibit
plainly	information	explain
position	inspect	expose
presume	investigate	find out
probably	judge	gather
really	look	hand over
seem	overhaul	illustrate
sensation	ponder	indicate
sentiment	proof	induce
suggestion	prove	inform
suppose	question	interpret
sure	quiz	introduce
surely	rate	learn
think	real	learn of
truly	reality	manifest
trust	rectify	notice
undoubtedly	reflect	observe
view	research	pay attention to
	revise	perceive
	sample	persuade
	science	point
	scrutinize	realize
	search	reason
	signal	recognize
	specify	reveal
	suggest	say
	supervise	show
	tentative	take note
	test	tell
	testimony	think
	trace	think about
	track	tolerate
	trial	uncover
	truth	understand
	try	unfold
	validate	unveil
	verify	warn
	virtually	
	witness	

Extended Data Table 1: Keyword lists for the three honesty components belief-speaking, truth-seeking and fostering understanding.

	Analytic		Authentic		Moral		Pos. emotion		Neg. emotion	
Text Category	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Belief-speaking	63.52	30.82	46.16	33.93	0.68	1.77	0.86	2.06	0.43	1.39
Truth-seeking	74.14	26.13	43.49	32.82	0.62	1.67	0.71	1.81	0.36	1.22
Neutral	76.80	27.03	42.05	33.45	0.74	1.94	1.15	2.54	0.35	1.29

Extended Data Table 2 LIWC score averages measuring the prevalence of "analytic", "authentic" and "moral" language, as well as positive and negative emotions stratified by honesty component.

Score	Label	Description
1	False Information	No or very little accuracy (e.g. fake news, conspiracy, satire)
2	Clickbait	Might contain smatterings of facts but is mostly misleading or clickbait
3	Biased	Mixed accuracy, half-truths, left/right bias
4	Mainstream	Low bias, mainstream media
5	Scientific	No reporting bias, scientific information

Extended Data Table 3 Description of accuracy scores.

Score	Label	Description
1	No Transparency	Intentionally misleading or no information about editorial process (e.g. fake news, conspiracy)
2	Mixed Transparency	Sites with (partially) transparent intention, but can still be misunderstood because of the way articles are written (e.g. bias, clickbait, satire)
3	Transparent	Sites with a transparent editorial process and legal notice (e.g. mainstream, scientific news)

Extended Data Table 4 Description of transparency scores.

	Estimate	std. err.	t	P	[0.025	0.975]
Intercept	91.256	1.048	87.107	0.000	89.200	93.312
Republican	-1.679	0.884	-1.898	0.058	-3.414	0.057
$\log(N_f)$	-0.222	0.102	-2.178	0.030	-0.422	-0.022
$\log(N_t)$	0.429	0.155	2.761	0.006	0.124	0.734
T_b	2.141	5.756	0.372	0.710	-9.155	13.437
Republican $\times$ T_b	-68.298	8.957	-7.625	0.000	-85.875	-50.722
T_t	12.086	4.682	2.581	0.010	2.898	21.273
Republican $\times$ T_t	-3.999	6.330	-0.632	0.528	-16.420	8.422

R-squared	0.492	Mean dependent var	90.980
Adjusted R-squared	0.488	S.D. dependent var	5.439
Model MSE	2111.592	AIC	5657.731
Sum squared resid	15277.858	BIC	5697.128
Log-likelihood	-2820.865	F-statistic	139.456
Durbin-Watson stat	1.955	Prob(F-statistic)	0.000

Extended Data Table 5 Results of an ordinary least-squares regression for the average NewsGuard score S_{NG} of each U.S. Congress member following Eq.(1). 1017 observations were included. Regression was performed with the function `ols` from the Python package statsmodels [65], version 0.13.2.

	Estimate	std. err.	t	P	[0.025	0.975]
Intercept	94.773	0.068	1396.422	0.000	94.640	94.906
Republican	-9.194	0.108	-84.746	0.000	-9.406	-8.981
A_b	-42.563	11.542	-3.688	0.000	-65.186	-19.940
Republican × A_b	-19.107	17.684	-1.080	0.280	-53.768	15.553
A_t	19.365	6.108	3.170	0.002	7.393	31.337
Republican × A_t	-46.697	9.855	-4.739	0.000	-66.012	-27.383

R-squared	0.131	Mean dependent var	91.093
Adjusted R-squared	0.131	S.D. dependent var	12.906
Model MSE	701684.186	AIC	1255933.876
Sum squared resid	23268646.824	BIC	1255993.801
Log-likelihood	-627960.938	F-statistic	4847.361
Durbin-Watson stat	1.162	Prob(F-statistic)	0.000

Extended Data Table 6 Results of an ordinary least-squares regression for the NewsGuard score of each link S_{NG} following Eq.(2). 160,750 observations were included. Regression was performed with the function `ols` from the Python package statsmodels [65], version 0.13.2.

14 Supplementary information

Annotation guide

What follows is a copy of the guidelines provided to the three external annotators for the component validation on the Twitter corpus.

What is the definition and value of honesty?

Introduction

According to Cooper et al. (2021), the notion of being "honest" can be understood through three different lenses: belief-speaking, truth-seeking and fostering understanding. The first relates to communicating opinions, thoughts and feelings which are believed to be true by the speaker. The second involves gathering truthful information prior to stating one's belief, as well as updating one's point of view when collecting new facts. The third regards the willingness to foster understanding of true beliefs in someone's audience.

We are interested in classifying a series of texts regarding these three components of honesty, and we need your help.

You will be presented with texts related to only one of the three components at a time, meaning that you will only read texts that may show belief-speaking, truth-seeking or fostering understanding, and we ask you to ascertain the presence of these components in the text.

If you know what component you have been assigned to, you can jump to the related paragraph and learn how to annotate the texts.

Belief-speaking

If the texts you are asked to classify are related to belief-speaking, you should ask yourself the following question:

- Does the person communicate what they believe to be true?

Based on the answer you give, click on the cell on the right of the text, click on the arrow and select "Yes" or "No" accordingly.

Examples of texts that contains belief-speaking behaviour are the following:

- That statement was clearly made without the advice of counsel.
- We don't know what all will be in the Democrats' tax-and-spending spree, but we know their intent: a several trillion-dollar injection of federal government into every aspect of our lives and our economy.

We are not interested in third-person reporting of beliefs (i.e. "Trump believes the Earth is flat"), nor in negations of beliefs (i.e. "I don't believe that vaccines are safe"), therefore this kind of texts should be flagged as not showing the behaviour ("No"). Also, we are not judging the veracity of the statements, therefore a sentence such as "I believe the Earth is flat" must be flagged as showing the behaviour ("Yes"), regardless of its (un)truthfulness.

Truth-seeking

If the texts you are asked to classify are related to truth-seeking, you should ask yourself the following question:

- Does the person seek out truthful information and update their beliefs based on this information?

Based on the answer you give, click on the cell on the right of the text, click on the arrow and select "Yes" or "No" accordingly. Examples of texts that contains truth-seeking behaviour are the following:

- It's called a circumtriple planet, and evidence that one exists suggests that planet formation is less unusual than once believed.
- Today, I sent a letter to the Department of Homeland Security to inquire about the concerning increase in rabbit testing and the extent of the department's animal research.

We are not interested in third-person reporting of truth-seeking (i.e. "He is looking for the truth."), nor in negations of truth-seeking (i.e. "It's clear Trump is not looking for the truth."), therefore this kind of texts should be flagged as not showing the behaviour ("No").

Fostering understanding

If the texts you are asked to classify are related to fostering understanding, you should ask yourself the following question:

- Does the person attempt to foster true beliefs in the audience?

Based on the answer you give, click on the cell on the right of the text, click on the arrow and select "Yes" or "No" accordingly. Examples of texts that contains fostering understanding behaviour are the following:

- There's a lot of vaccine disinformation out there. Please consider posting on your own social media encouraging your network to get vaccinated! Your friends and family probably trust you more than something bogus they see online!
- With the attacks in Paris & San Bernardino, a good read from @60Minutes if you want to understand the #encryption debate.

14.1 LIWC scores for "authentic", "analytic" and "moral" text components

In Extended Data Fig. 2 we show the timelines of LIWC scores for positive and negative emotions broken down by honesty component. We performed the same analysis for "authentic", "analytic" and "moral" language, using LIWC dictionaries as described in the Methods Section "LIWC text analysis". The time development of "analytic" language broken down by honesty component is shown in Fig. S1, panels **A** to **D**, the time development of "authentic"

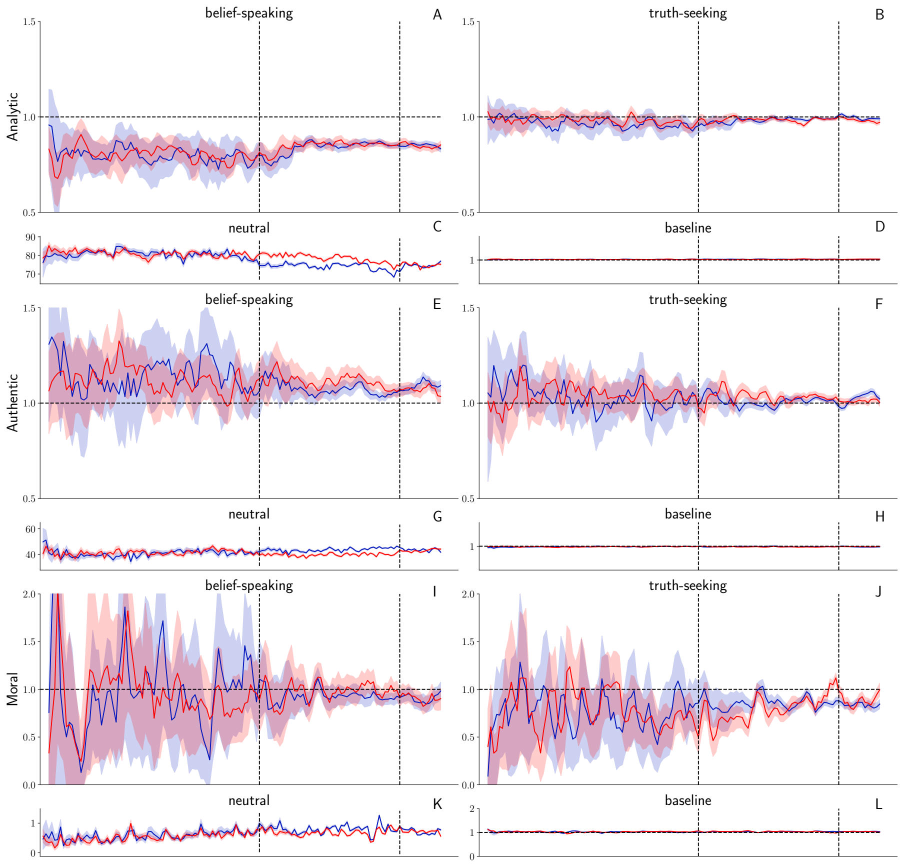


Figure S1 Time-development of LIWC scores of "analytic", "authentic" and "moral" language in tweets of members of the U.S. Congress. Panels **A** and **B** show the score for analytic language within the belief-speaking and truth-seeking components normalized by the overall score (baseline), respectively. The dashed horizontal line at 1.0 corresponds to prevalence equal to baseline. Red and blue lines correspond to tweets by Republicans and Democrats, respectively. Panel **C** shows the score within "neutral" (not belief-speaking or truth-seeking) tweets normalized by the baseline, and panel **D** shows the baseline. Panels **E**, **F**, **G** and **H** show the same information as panels **A**, **B**, **C** and **D**, but for "authentic" language instead of "analytic" language. Panels **I**, **J**, **K** and **L** show the same information but for "moral" language. The 95% confidence intervals (indicated by shading) were computed with bootstrap sampling over 1,000 iterations. Dashed vertical lines indicate dates of presidential elections in 2016 and 2020. Timelines are smoothed, using a rolling average over three months.

language is shown in Fig. S1 panels **E** to **H** and the time development of "moral" language is shown in Fig. S1 panels **I** to **L**.

Validation using an independently compiled list of unreliable news sources

To exclude a dependence of the main results reported in Section "Relation of honesty components to information trustworthiness" on use of the News-Guard data base, we validated this analysis with an independently collected list of news outlet reliability from academic and fact-checking sources. Details on how this list was compiled are reported in Section "Independent list of untrustworthy sources" below. Using this list, we can assign an accuracy score S_a ranging from 1 to 5 as well as a transparency score S_t, ranging from 1 to 3 to each domain. In addition, a domain with an accuracy score of ≤ 2 and/or a transparency score of 1 will be labelled as "unreliable". Similar to the analysis above, we assign an average accuracy score $\langle S_\mathrm{a} \rangle$ and average transparency score $\langle S_\mathrm{t} \rangle$ to each account, based on the individual accuracy and transparency scores of all links posted by that account.

Again, we fit an ordinary least squares regression model for each of the two scores:

$$\langle S_\mathrm{a} \rangle \sim T_b + T_t + P \times T_b + P \times T_t + \log(N_f) + \log(N_t)\text{ , and} \quad (3)$$

$$\langle S_\mathrm{t} \rangle \sim T_b + T_t + P \times T_b + P \times T_t + \log(N_f) + \log(N_t)\,. \quad (4)$$

Here, $\langle S_\mathrm{a} \rangle$ and $\langle S_\mathrm{t} \rangle$ are the independently determined accuracy and transparency scores of domains averaged for each Congress member, respectively.

Again, we found a significant relation between the interaction of T_b and Republican, and a low $\langle S_\mathrm{a} \rangle$ and $\langle S_\mathrm{t} \rangle$ (accuracy: coefficient -4.64 [-6.07; -3.22], $p < 0.001$, $t = -6.39$, transparency: coefficient -2.91 [-3.70; -2.12], $p < 0.001$, $t = -7.21$). We also found a significant inverse relationship between the main effect of T_b and $\langle S_\mathrm{a} \rangle$ (coefficient -1.32 [-2.23; -0.40], $p = 0.005$, $t = -2.83$) as well as $\langle S_\mathrm{t} \rangle$ (coefficient -0.54 [-1.05; -0.03], $p = 0.037$, $t = -2.09$). The dependence on $P \times T_\mathrm{t}$ is non-significant for both $\langle S_\mathrm{a} \rangle$ and $\langle S_\mathrm{t} \rangle$ (accuracy: coefficient 0.56 [-0.45; 1.57], $p = 0.274$, transparency: coefficient 0.23 [-0.33; 0.79], $p = 0.425$, $t = 0.80$). Full regression statistics are reported in Tables S1 and S2. We note that there is a strong correlation between the average NewsGuard score S_NG and the average accuracy and transparency scores S_a and S_t of user accounts, respectively ($R^2 = 0.74$ [0.69; 0.77] and $R^2 = 0.76$ [0.72; 0.80], 95% CI from 10,000 bootstrapping iterations, as shown in Fig. S2) **A** and **B**. An account that is labelled "untrustworthy" in the NewsGuard data base has a high chance of being labelled "unreliable" in the alternative database as well (Krippendorff's α of 0.84).

In addition and as shown in Fig. S2 **C**, there is a moderate correlation between a politician's PolitiFact score [67] and $\langle S_\mathrm{NG} \rangle$ (0.44 [0.35; 0.52]). PolitiFact [68] is a fact-checking website that monitors the claims made by politicians and other public figures, and rates them for their accuracy. [67] have curated a dataset of 1005 public figures and calculated the average veracity scores of documented statements on PolitiFact. This dataset contained 302 Members of Congress that overlapped with our data.

Independent list of untrustworthy sources

We compiled a list of trustworthiness ratings from a range of academic sources and fact-checking sites. Most of these sources were also used by [69]. The list includes Bufale [70], Bufalopedia [71], Butac [72], Buzzfeed News [73],

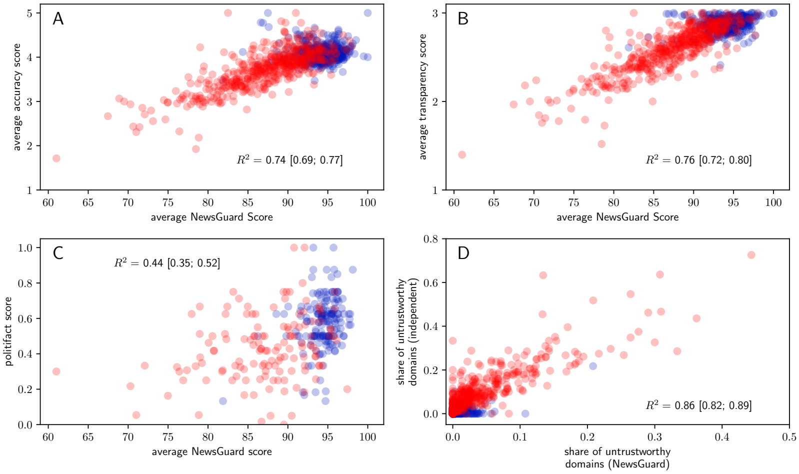


Figure S2 Correlations between different measures of trustworthiness of information posted on Twitter by members of the U.S. Congress. **A** average accuracy score over average NewsGuard score. **B** average transparency score over average NewsGuard score. **C** PolitiFact score [68] over average NewsGuard score. **D** share of untrustworthy domains (NewsGuard) over share of unreliable domains (accuracy ≤ 2 and/or transparency $= 1$). All values were averaged over all URLs contained in tweets for every twitter account associated with a member of the U.S. Congress. Accounts associated with Republicans and Democrats are indicated in red and blue, respectively. Independent politicians are omitted.

	Estimate	std. err.	t	P	[0.025	0.975]
Intercept	4.241	0.085	50.091	0.000	4.075	4.407
Republican	-0.118	0.072	-1.647	0.100	-0.260	0.023
$\log(N_f)$	-0.022	0.008	-2.608	0.009	-0.038	-0.005
$\log(N_t)$	0.015	0.013	1.221	0.222	-0.009	0.040
T_b	-1.317	0.466	-2.827	0.005	-2.231	-0.403
Republican $\times$ T_b	-4.645	0.727	-6.389	0.000	-6.072	-3.218
T_t	0.242	0.377	0.643	0.521	-0.497	0.981
Republican $\times$ T_t	0.561	0.512	1.094	0.274	-0.445	1.566

R-squared	0.318	Mean dependent var	3.910
Adjusted R-squared	0.313	S.D. dependent var	0.381
Model MSE	6.687	AIC	547.840
Sum squared resid	100.452	BIC	587.237
Log-likelihood	-265.920	F-statistic	67.167
Durbin-Watson stat	1.972	Prob(F-statistic)	0.000

Table S1 Results of an ordinary least-squares regression for the average accuracy score S_a of each U.S. Congress member as described in Eq.(3). 1017 observations were included. Regression was performed with the function `ols` from the Python package statsmodels [65], version 0.13.2.

Columbia Journalism Review [74], Fake News Watch [75], Media Bias Fact Check [76], Politifact [68], and Melissa Zimdars [77]. After removing duplicates, our list contained 4,767 domains, 1,677 of which were also contained in the NewsGuard data base, as of March 1, 2022.

The main challenge in combining lists from different fact checkers lies in unifying the labels the fact checkers assign to the domains. To address this, we devised a scheme where we rated each domain on two dimension that we

	Estimate	std. err.	t	P	[0.025	0.975]
Intercept	2.819	0.047	60.021	0.000	2.727	2.911
Republican	-0.111	0.040	-2.775	0.006	-0.189	-0.032
$\log(N_f)$	-0.012	0.005	-2.536	0.011	-0.021	-0.003
$\log(N_t)$	0.021	0.007	3.013	0.003	0.007	0.035
T_b	-0.540	0.258	-2.089	0.037	-1.047	-0.033
Republican $\times$ T_b	-2.906	0.403	-7.208	0.000	-3.698	-2.115
T_t	0.323	0.209	1.545	0.123	-0.087	0.733
Republican $\times$ T_t	0.227	0.284	0.798	0.425	-0.331	0.785

R-squared	0.448	Mean dependent var	2.741
Adjusted R-squared	0.444	S.D. dependent var	0.235
Model MSE	3.586	AIC	-651.090
Sum squared resid	30.901	BIC	-611.693
Log-likelihood	333.545	F-statistic	117.092
Durbin-Watson stat	1.969	Prob(F-statistic)	0.000

Table S2 Results of an ordinary least-squares regression for the average transparency score S_t of each U.S. Congress member as described in Eq.(4). 1017 observations were included. Regression was performed with the function `ols` from the Python package statsmodels [65], version 0.13.2.

consider to be important to assess reliability and trustworthiness of information: "accuracy" and "transparency". We devise an accuracy score S_a that varies from 1 (false information) to 5 (scientific) and a transparency score S_t that varies from 1 (no transparency) to 3 (transparent). We provide a more detailed description of the five accuracy and three transparency levels in Extended Data Tables 3 and 4. Mappings of the labels of individual fact checking sites to accuracy and transparency scores as well as the full list of domains are provided at [78].

After mapping all individual lists to the accuracy and transparency dimensions, we label every domain that has an accuracy score of 1 (False Information) or 2 (Clickbait) and/or a transparency score of 1 (No Transparency) as "unreliable". This results in a total of 2,170 domains being labelled as "unreliable" and 2,597 as "reliable". For the 1,677 domains that are contained in both data bases, the Krippendorff's α between "untrustworthy" (score < 60 in NewsGuard) and "unreliable" in the independently compiled data base is 0.84, which shows a very high agreement between the two databases. The independently compiled domain list including the unified labels is openly accessible at https://doi.org/10.5281/zenodo.6536692.

After excluding links to other social media platforms (e.g., twitter.com, facebook.com, youtube.com, and instagram.com) as well as links to search services (google.com, yahoo.com), the database covers a very similar share of links as the NewsGuard data base (between 20% and 60%) — see also Extended Data Fig. 3 **B**.

14.2 Mediation analysis

Why is it the case that belief speaking is the preferred means to spread low-quality information? One possibility is that belief-speaking is the result of Republican politicians' desire to disparage Democrats, as suggested by [46], given that belief speaking was found to be associated with greater negative

		Estimate	P	[0.025	0.975]
Republicans	ACME	0.055	<.001	0.021	0.103
	ADE	0.031	0.043	0.001	0.081
	Total Effect	0.086	<.001	0.037	0.161
	Prop. Mediated	0.635	<.001	0.407	0.985
Democrats	ACME	0.002	0.008	0.001	0.001
	ADE	-0.008	0.009	-0.016	-0.002
	Total Effect	-0.006	0.044	-0.013	-0.001
	Prop. Mediated	-0.383	0.052	-2.372	0.011

Table S3 Mediation analysis with belief speaking as mediator. ACME = average causal mediation effect; ADE = average direct effect. 493 observations were included for Republicans and 525 for Democrats. Mediation was performed using the function `mediate` from the R package mediation, version 4.5.0.

emotion (see Section 2), and given that lower-quality information tends to be biased towards negativity [47]. According to this theory, the relationship between belief-speaking and low-quality shared information should be mediated by negative emotional content. On the other hand, if believe speaking were involved in the dissemination of poor quality content for other reasons, it should mediate the association involving unpleasant emotions.

To test these opposing predictions, we examined separately for Democrats and Republicans whether (1) negative emotion mediated the effects of belief speaking on sharing low-quality information, or (2) belief speaking mediated the effects of negative emotion on sharing low-quality information. For each user, we computed mean scores of negative emotion (measured via LIWC), belief speaking, and prevalence of sharing low-quality news (average proportion of sharing articles with a NewsGuard score of < 60). We conducted a causal mediation analysis using the ‘mediation‘ R package [79] and a bootstrap method with 10,000 iterations. Among Republicans, when considering negative emotion as a mediator, the effect of the direct path was statistically significant (mean direct effect = 0.77, 95% CI of bootstrapped samples = [0.52, 1.01], $p < .0001$). The mediation was also significant (average causal mediation effect = 0.07, 95% CI = [0.00, 0.20], $p = .035$), accounting for 8% of the total effect. When considering belief speaking as a mediator, the pattern was similar: the direct was statistically significant (mean direct effect = 0.03 95% CI of bootstrapped samples = [0.00, 0.08], $p = .043$), so did the average causal mediation effect, (average causal mediation effect = 0.05, 95% CI = [0.02, 0.10], $p < .0001$). Notably, belief speaking as a mediator explained 63% of the total effect of negative emotion on sharing low-quality news. See tables S3 and S4 for the full details.

		Estimate	P	[0.025	0.975]
Republicans	ACME	0.068	0.035	0.005	0.204
	ADE	0.765	<.001	0.515	1.006
	Total Effect	0.833	<.001	0.606	1.075
	Prop. Mediated	0.081	0.035	0.006	0.249
Democrats	ACME	-0.015	0.008	-0.029	-0.004
	ADE	0.034	0.008	0.009	0.063
	Total Effect	0.019	0.112	-0.004	0.046
	Prop. Mediated	-0.764	0.120	-6.938	5.200

Table S4 Mediation analysis with negative emotion as mediator. ACME = average causal mediation effect; ADE = average direct effect. 493 observations were included for Republicans and 525 for Democrats. Mediation was performed using the function `mediate` from the R package mediation, version 4.5.0.