Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-023-02106-x OPEN

Negativity bias in the spread of voter fraud conspiracy theory tweets during the 2020 US election

Mason Youngblood [1,2]✉, Joseph M. Stubbersfield [3], Olivier Morin [1,4], Ryan Glassman[5] & Alberto Acerbi[6]

During the 2020 US presidential election, conspiracy theories about large-scale voter fraud were widely circulated on social media platforms. Given their scale, persistence, and impact, it is critically important to understand the mechanisms that caused these theories to spread. The aim of this preregistered study was to investigate whether retweet frequencies among proponents of voter fraud conspiracy theories on Twitter during the 2020 US election are consistent with frequency bias and/or content bias. To do this, we conducted generative inference using an agent-based model of cultural transmission on Twitter and the *VoterFraud2020* dataset. The results show that the observed retweet distribution is consistent with a strong content bias causing users to preferentially retweet tweets with negative emotional valence. Frequency information appears to be largely irrelevant to future retweet count. Follower count strongly predicts retweet count in a simpler linear model but does not appear to drive the overall retweet distribution after temporal dynamics are accounted for. Future studies could apply our methodology in a comparative framework to assess whether content bias for emotional valence in conspiracy theory messages differs from other forms of information on social media.

[1] Minds and Traditions Research Group, Max Planck Institute for Geoanthropology, Jena, Germany. [2] Institute for Advanced Computational Science, Stony Brook University, Stony Brook, NY, USA. [3] Department of Psychology, University of Winchester, Winchester, UK. [4] Institut Jean Nicod, ENS, EHESS, PSL University, CNRS, Paris, France. [5] IBM Watson, Yorktown, VA, USA. [6] Department of Sociology and Social Research, University of Trento, Trento, Italy. ✉email: masonyoungblood@gmail.com

Introduction

Allegations of malign acts, carried out in secret by powerful groups, have been offered as explanations for major events throughout history, from ancient Rome, through the medieval period, to the present day (Brotherton, 2015; Pagán, 2020; Zwierlein, 2020). People across the globe have shared these conspiracy theories (Butter & Knight, 2020; West & Sanders, 2003). Conspiracy theories have been a part of North American culture since the colonial period, with beliefs about conspiring, "un-American" groups of witches, enslaved Africans, Masons, Catholics, and Jews dominating early versions, before shifting to the US government itself as the source of conspiring agents in the nineteenth and twentieth centuries (Goldberg, 2003; Olmsted, 2018).

Conspiracy theories are typically defined as explanations of important events that allege secret plots by powerful actors as salient causes (Douglas et al., 2019; Goertzel, 1994; Keeley, 1999). Belief is not inherently irrational (as conspiracies do occur; see (Dentith, 2014; Pigden, 1995)), but conspiracy *theories* (as opposed to simply conspiracies) are allegations that survive and spread despite a lack of reliable evidence (Douglas et al., 2019; Keeley, 1999). Belief in conspiracy theories is associated with reduced engagement with mainstream politics (Imhoff et al., 2020; Jolley & Douglas, 2014), increased support for political violence and extremism (Imhoff et al., 2020; Uscinski & Parent, 2014), and increased prejudice towards minority groups (Jolley et al., 2020; Kofta et al., 2020).

A range of recent and ongoing conspiracy theories allege that the result of the 2020 US presidential election was achieved through large-scale electoral fraud (Enders et al., 2021). Building on allegations of voter fraud made prior to the 2016 election (Cottrell et al., 2018) and years of Republican messaging about electoral fraud and illegal voting (Edelson et al., 2017), these conspiracy theories were widely circulated on social media platforms like Twitter. Major political and public figures, including US President Donald Trump, boosted these theories using hashtags like #stopthesteal (Sardarizadeh & Lussenhop, 2021) and eventually had their accounts suspended for incitement of violence following the January 6th attack on the US Capitol (Conger & Isaac, 2021). More specific claims, such as hacked voting machines being programmed in favor of then-Presidential Candidate Joe Biden and large numbers of ballots being thrown out in trash bags (Cohen, 2021; Spring, 2020) have been used to justify election audits and tighter voting laws in states like Arizona (Cooper & Christie, 2021) and Georgia (Corasaniti & Epstein, 2021). The Justice Department has found "no evidence of widespread voter fraud" (Balsamo, 2020), and the Cybersecurity and Infrastructure Security Agency concluded that 2020 was "the most secure election ever" (Tucker & Bajak, 2020). Despite this, polls suggest that up to a third of Americans (Cillizza, 2021) and the majority of Republicans (Skelley, 2021) believe that Biden won the election illegitimately through voter fraud. Exposure to such claims has been shown to reduce confidence in democratic institutions (Albertson & Guiler, 2020) and is thought to have contributed to motivating the US Capitol attack (Beckett, 2021). Given the scale, persistence, and impact of voter fraud conspiracy theories, it is critically important to understand the mechanisms that caused them to spread.

While conspiracy theories, like everything else, are disseminated through social media, the nature of the association between social media usage and conspiracy theory belief is an open question (Enders et al., 2021; Hall Jamieson & Albarracín, 2020; Min, 2021; Stempel et al., 2007). Social media does provide, however, a source of data that can be used to test theories about their spreading. Most studies of the spread of conspiracy theory messages on social media have focused on the content of posts in general, highlighting the importance of negative content (Schöne et al., 2021), emotional content (Brady et al., 2017), or out-group derogation (Osmundsen et al., 2021; Rathje et al., 2021). However, content is only one of the possible features that influence the success of a social media post. In what follows, we use a framework inspired by cultural evolution that allows us to distinguish among various features and assess their relative importance. It is important to highlight up-front that this is not a comparative study, so we are unable to make conclusions about whether detected processes are unique to conspiracy theories relative to other forms of information. Instead, our goal is simply to characterize what transmission processes are present in a particular high-profile case of conspiracy theory spread on social media.

Broadly, cultural evolution adopts an evolutionary framework to research the stability, change, and diffusion of cultural traits (Mesoudi, 2011). Transmission biases—biases in social learning that cause individuals to adopt some cultural variants over others—are thought to be some of the most important factors driving cultural evolutionary patterns (Kendal et al., 2018). According to this perspective, the probability that a behavior will be adopted is influenced by various cues. Frequency bias, which includes conformity and anticonformity bias, is when the frequency of a variant in the population disproportionately affects its probability of adoption (e.g., users are more likely to retweet something viral). Content bias is when the inherent characteristics of a variant affect its probability of adoption (e.g., users are more likely to retweet content with higher emotional valence). Demonstrator bias is when some characteristic of the individuals expressing a variant affects its probability of adoption (e.g., users more likely to retweet content from verified users) (see review in (Kendal et al., 2018)). Importantly, transmission biases can lead to discernible changes in the cultural frequency distributions of populations (Lachlan et al., 2018). For example, in the context of Twitter, a positive frequency bias (i.e., conformity) would cause users to be more likely to retweet content that has already been heavily retweeted by other users, thus increasing the right skew of the overall retweet distribution. This framework allows us to consider both individual susceptibility and the influence of social context on wider population-level patterns.

Using generative inference, it is possible to infer the underlying cognitive biases of individuals in a population from the cultural frequency distribution that they generate. Generative inference is a statistical procedure in which a model is run many times with varying parameter values to generate large quantities of simulated data. This simulated data is then compared to real data using approximate Bayesian computation (ABC) to infer the parameter values that likely generated it (Kandler & Powell, 2018). ABC is most often used when likelihood functions are intractable, as is often the case when studying population-level patterns with incomplete data. While there are pseudo-experimental approaches that allow direct measurement of biases on social media, they require a level of access that is very difficult for large-scale phenomena (e.g., users providing researchers with full access to their Twitter accounts) (Butler et al., 2023; Milli et al., 2023). Carrignon et al. (Carrignon et al., 2019) recently applied generative inference to the spread of confirmed and debunked information on Twitter and found that the retweet distributions of both were more consistent with random copying than with conformity. However, their model did not include parameters for the influence of follower count and did not explore the influence of content bias due to computational limitations (Carrignon et al., 2019).

The aim of this study is to investigate whether retweet frequencies among proponents of voter fraud conspiracy theories on Twitter during the 2020 US election are consistent with frequency bias and/or content bias. To do this, we conducted generative inference using an agent-based model (ABM) of cultural transmission on Twitter that combines elements from Carrignon et al. (Carrignon et al., 2019), Lachlan et al. (Lachlan et al., 2018), and Youngblood and Lahti (Youngblood & Lahti, 2022) (see Methods for details). Twitter (recently renamed "X") is a microblogging platform where users can post original content (i.e., "tweets"), read an algorithmically generated timeline of content that prioritizes recent posts from people that they follow, and search for new content using keywords. Users can engage with tweets by replying to them, liking them, and sharing them with their own followers (i.e., "retweeting"). All tweets include the date and time they were posted, the name and verification status of the author (see below), and their engagement (e.g., how many and which users replied, liked, and retweeted). Our ABM simulates a population of Twitter users with follower counts, activity levels, and probabilities of tweeting original tweets from real users in the observed data. Every six hours, a subset of users become active and either compose a new tweet or retweet an existing tweet. The probability of an existing tweet being retweeted is based on four factors: (1) the attractiveness of the content in the tweet (e.g., emotional valence and/or intensity), (2) the follower count of the user who tweeted it, (3) how many times it has already been retweeted, and (4) the age of the tweet. The influence of each of these factors on retweet probability is controlled by separate parameters which correspond to content bias (c), follower influence (d), frequency bias (a), and age dependency (g), which we fitted to real data using the random forest version of ABC (Raynal et al., 2019). This ABM follows Carrignon et al. (Carrignon et al., 2019) in assuming a fully connected population so that, under neutral conditions, retweet probability is independent of follower count. Note that retweet probability being independent of follower count is an unrealistic scenario, as tweets from users with more followers will be seen by more users. As such, the parameter d simulates departure from this baseline, where follower count has an increasing influence on retweet probability. Our original intent was to use d to assess demonstrator bias, but a more parsimonious explanation for an effect of follower count is network structure—tweets from people with more followers appear on more users' timelines. It is very difficult to disentangle these two factors without access to the full follower network of users, so we will describe d with the more general term "follower influence".

The data used in this study comes from a team of researchers at Cornell Tech, who retrieved millions of tweets and retweets relating to voter fraud conspiracy theories between October 23 and December 16, 2020 (Abilov et al., 2021). After iteratively building a set of search terms from the seeds "voter fraud" and #voterfraud and using them to collect data in real-time, they estimate that they collected ~60% of tweets about voter fraud conspiracy theories during that period. An anonymized version of the *VoterFraud2020* dataset is publicly available (https://github.com/sTechLab/VoterFraud2020), and Abilov et al. (Abilov et al., 2021) generously provided us with access to their full disambiguated dataset. Importantly, this dataset includes tweets from users who were "purged" from Twitter following the US Capitol attack (Romm & Dwoskin, 2021). Beyond the quality of the data and its intrinsic historical importance, the choice of this dataset was motivated by its exceptional scale and focus. Few other datasets gather such a large number of social media actors engaging in one conversation topic over a period of several months.

Additionally, we conducted secondary analyses with general linear mixed models (GLMM) to assess the potential targets of content bias, while accounting for other factors such as follower count and whether the account holder was verified. Twitter verifies some accounts to make sure they are authorized by the person they claim to represent but only undertakes this costly verification for high-profile accounts whose status is signaled by a "blue check mark" icon (the data from this study predate Twitter's paid verification policy that began in the fall of 2022). The emotional valence of tweets was measured using the valence-aware dictionary and sentiment reasoner (VADER), a sentiment analysis model trained for use with Twitter and other social media data (Hutto & Gilbert, 2014). The output of VADER includes a compound score of the overall emotional valence and intensity from strongly negative to positive, in addition to the proportion of words that are negative, positive, and neutral. A large body of research suggests that content with negative valence has an advantage over content with positive valence across several domains (Baumeister et al., 2001; Rozin & Royzman, 2001). In digital media, evidence of negativity bias has been suggested within online "echo chambers" (Asatani et al., 2021; Del Vicario et al., 2016) and for tweets about political events both from individual users (Schöne et al., 2021) and institutions (Bellovary et al., 2021). False rumors, on the other hand, seem to be subject to a positivity bias (Pröllochs et al., 2021), despite the fact that "fake news" articles tend to contain more negative language (Acerbi, 2019). Other studies have suggested that just the strength of emotion influences the transmission of content on social media (Brady et al., 2017; Stieglitz & Dang-Xuan, 2013) (but see critiques in (Burton et al., 2021)), and, in two experimental studies, the appeal of conspiracy theories was associated with the intensity of emotion evoked, rather than the valence of that emotion (van Prooijen et al., 2021).

Based on this research, if content bias is detected, then we hypothesize that it will be targeted towards more emotional content (i.e., more positive or negative according to VADER, which collapses valence and intensity into one indicator), but we make no specific prediction about the direction of valence.

Methods

Our methods, models, and predictions were preregistered in advance of data analysis (https://osf.io/jnvyf), except for the post hoc comparison between tweets and quote tweets. Departures from our preregistration that arose during peer review can be found in the SI.

The data for this study comes from the *VoterFraud2020* dataset, collected between October 23 and December 16, 2020, by Abilov et al. (Abilov et al., 2021). This dataset includes 7.6 million tweets and 25.6 million retweets that were collected in real-time using Twitter's streaming API. The *VoterFraud2020* dataset was collected according to Twitter's Terms of Service and is consistent with established academic guidelines for ethical social media data use (Abilov et al., 2021). Abilov et al. (Abilov et al., 2021) started out with a set of keywords and hashtags that co-occurred with "voter fraud" and #voterfraud between July 21 and October 22, and expanded their search with additional keywords and hashtags as they emerged (e.g., #discardedballots and #stopthesteal). They estimate that their dataset includes at least 60% of tweets that included their search terms. Abilov et al. (Abilov et al., 2021) also applied the infomap clustering algorithm to the directed retweet network to identify different communities that engaged with voter fraud conspiracy theories. We ran our analysis using only the user and tweet data from cluster #2, the "proponent" community that tweets primarily in English and does not have significant connections to members of the "detractor" community. We restricted our analysis to cluster #2 so that retweets would be indicative of the spread of the conspiracy theories among proponents, as opposed to discourse and debate between both proponents and detractors. The activity levels and original tweet probabilities from these data only reflect users' interactions with conspiracy theory content. By using these as input to our agent-based model, we are only simulating the subset of users' behavior that pertains to the voter fraud conspiracy theory and is relevant for producing the observed retweet distribution.

The agent-based model (ABM) we used has elements from Carrignon et al. (Carrignon et al., 2019), Lachlan et al. (Lachlan et al., 2018), and Youngblood and Lahti (Youngblood & Lahti, 2022), and is available as an R package on GitHub (https://github.com/masonyoungblood/TwitterABM). The ABM is initialized with a fully connected population of N users and is run for 216 timesteps, each of which corresponds to a 6-h interval in the real dataset (the highest resolution possible given computational limits). Each simulated user is assigned a follower count (T), an activity level (r), and a probability of tweeting an original tweet (μ) from a real user in the observed data. In this way, we retain the correlation structure of follower count, activity, and retweet probability to reflect real variation in users' behavior on Twitter (e.g., users with few followers who exclusively retweet). μ is the proportion of a user's total tweets and retweets that are original tweets (population mean of ~0.45). T is scaled with a mean of 1 and a standard deviation of 1. The ABM is also initialized with a set of tweets with retweet frequencies drawn randomly from the first timestep in the observed data. Each tweet is assigned an attractiveness (M). At the start of each timestep, a pseudo-random subset of users becomes active (weighted by their values of r) and tweets according to the observed overall level of activity in the same timestep. Each active user either tweets original tweets or retweets existing tweets based on their unique value of μ. New original tweets are assigned an attractiveness of M, while retweets occur with probability $P(x)$:

$$P(x) = \frac{F_x^a \cdot T_x^d \cdot M_x^c \cdot \frac{1}{age_x^g}}{\sum_{x=1}^{n} F_x^a \cdot T_x^d \cdot M_x^c \cdot \frac{1}{age_x^g}}$$

The denominator simply normalizes the probability of retweeting x against the probability of retweeting all other tweets in the population. F is the number of times that a tweet has been previously retweeted, and is raised by the level of frequency bias (a). a is the same across all agents, where values > 1 simulate conformity bias and values < 1 simulate anticonformity bias. T is raised by the level of follower influence (d). d is the same across all agents, where values of 0 simulate neutrality by removing variation in follower count and values > 0 simulate increasing levels of follower influence. M is the attractiveness of the tweet and is drawn from a truncated normal distribution with a mean of 1, a standard deviation of 1, and a lower bound of 0. Note that our measures of tweet content follow a variety of different forms (e.g., the compound score is zero-inflated Gaussian, positive/negative words are proportions, and presence/absence of media is binary). In the agent-based model, we made the simplifying assumption that attractiveness fits a normal distribution because it is common in natural systems and useful for approximation when information about generative processes is limited (Frank, 2009). M is raised by the level of content bias (c). c is the same across all agents, where values of 0 simulate neutrality by removing variation in the attractiveness of content and values > 0 simulate increasing levels of content bias. The final term simulates the decreasing probability that a tweet is retweeted as it ages, where g controls the rate of decay. Once the active users are done each tweet increases in age by 1 and the next timestep begins. Lastly, we should note that we chose to exclude "top n" dynamics (e.g., trending topics) from our ABM, because they did not improve the fit of neutral models of cultural transmission on Twitter in Carrignon et al. (Carrignon et al., 2019) and hashtags/keywords related to voter fraud rarely made it into the top trending topics during our study period (https://www.exportdata.io/trends/united-states).

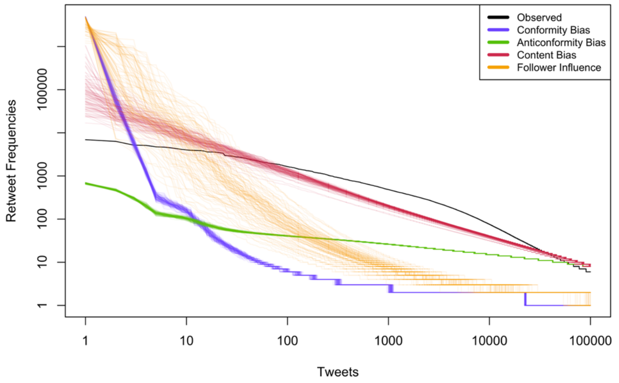

Fig. 1 The retweet distributions resulting from conformity, anticonformity, content, and follower influence using this ABM (100 iterations each), alongside the observed retweet distribution (in black). Biases were all modeled with a g of 0.25 and the following parameter values: $a = 1.4$ (conformity bias), $a = 0.6$ (anticonformity bias), $c = 1$ (content bias), and $d = 1$ (follower influence). The x-axis is the identity of each tweet ranked by descending retweet count, and the y-axis shows the number of times each of these tweets was retweeted. Both axes have been log-transformed.

In summary, the following are the dynamic parameters in this ABM that we estimated using approximate Bayesian computation (ABC):

- c—variation in the salience of the attractiveness of content
- d—variation in the salience of follower count
- a—level of frequency bias
- g—rate of decay in tweet aging

All other parameters in the ABM were assigned static values based on the real dataset. The output of this ABM is a distribution of retweet frequencies (see Fig. 1), which was used to calculate the following summary statistics: (1) the proportion of tweets that only appear once, (2) the proportion of the most common tweet, (3) the Hill number when $q = 1$ (which emphasizes more rare tweets), and (4) the Hill number when $q = 2$ (which emphasizes more common tweets). We used Hill numbers rather than their traditional diversity index counterparts (Shannon's and Simpson's diversity) because they are measured on the same scale and better account for relative abundance (Chao et al., 2014; Roswell et al., 2021).

The same summary statistics were calculated from the observed retweet distribution of the real dataset. For purposes of the summary statistic calculations, quote tweets were treated like original tweets, as they themselves can be retweeted. Then, the random forest version of ABC (Raynal et al., 2019) was conducted with the following steps:

- 500,000 iterations of the ABM were run to generate simulated summary statistics for different values of the parameters: c, d, a, and g.
- The output of these simulations was combined into a reference table with the simulated summary statistics as predictor variables, and the parameter values as outcome variables. The parameter values were logit-transformed prior to inference, as is recommended for bounded outcome variables in non-linear regression-based ABC (Blum & François, 2010; Sisson et al., 2018).
- A random forest of 1000 regression trees was constructed for each of the four parameters using bootstrap samples from the reference table. The number of sampled summary

statistics and the optimal minimum node size were tuned to minimize prediction error using the *tuneRanger* package in R (Probst et al., 2018) (see Table A in the SI).

- Each trained forest was provided with the observed summary statistics, and each regression tree was used to predict the parameter values that likely generated the data.

Uniform prior distributions were used for all four of the dynamic parameters: $c = \{0–12\}$, $d = \{0–4\}$, $a = \{0–2\}$, $g = \{0–8\}$. We plotted the output from 10,000 iterations to ensure that we were capturing enough of the parameter space before running the full analysis (see Fig A in the SI). We conducted posterior checks by running the agent-based model with parameter values drawn from the posterior distributions to see how closely the output matched the original data (see Fig A and Fig B in the SI). We also repeated the analysis using the basic rejection form of ABC to ensure that our main conclusions are robust to the choice of the random forest algorithm (see Fig C in the SI). Finally, we ran four additional rounds of the ABM, each with only a single term from the probability function included, to investigate the behavior of each parameter in isolation (see Fig D and Fig E in the SI).

Sentiment analysis was conducted using VADER from the natural language toolkit in Python, a model that performs similarly to human raters when applied to social media posts from platforms like Twitter ($r = 0.88$) (Hutto & Gilbert, 2014). VADER assigns a valence score to each word (and emoji or emoticon) in a tweet and weights those scores according to a set of rules (e.g., negation, capitalization, punctuation). The main output of VADER is a compound score that sums and normalizes the weighted valences of the words in a tweet to give an overall score of emotional valence and intensity between -1 (strongly negative) and $+1$ (strongly positive) (Fig F in the SI). VADER also outputs the proportion of words in a tweet that are identified as neutral, positive, or negative. VADER was specifically trained to handle emojis, URLs, hashtags, and tagged users during sentiment analysis so we did not remove those from our dataset. Up-to-date details about VADER can be found in the GitHub repository (https://github.com/cjhutto/vaderSentiment).

To determine the potential targets of content bias we conducted GLMM using the *lme4* package in R (Bates et al., 2015). Retweet frequency was used as the outcome variable. To determine which grouping variables would be suitable as random effects we ran separate null models with each and calculated the intraclass correlation coefficient (ICC), or the proportion of the variance in retweet frequency explained by the grouping levels of each variable. Once random effects were chosen we added predictor variables in three stages, using the Akaike information criterion (AIC) and likelihood-ratio test (LRT) to choose between competing models. First, we determined whether tweet length would be an appropriate control variable. Next, we determined whether a tweet that includes media (i.e., image or video) would be an appropriate control variable. Then, we added follower count and verification status to see which measure of platform size best improves the model. The 1.9% of tweets from users with missing verification statuses and follower counts were assigned verification statuses of "false" and follower counts of 0. Finally, we added the compound score, the proportion of negative words, and the proportion of positive words to see which measure of content best improves the model. All continuous predictor variables were scaled and centered prior to analysis. Model choice and residual diagnostic tests were conducted using a random 10% of observations, but the best-fitting model was run using the entire dataset. The Poisson family was used since our outcome variable was count data and did not appear to have over- or underdispersion issues (Fig H in the SI).

To ensure that our decision to treat quote tweets like original tweets did not bias our results related to the content, we did a second round of GLMM to determine whether quote tweets have different emotional valence compared to the original tweets that they are quoting. We refer to original tweets that are quoted as target tweets (i.e., targets), and the tweets that quote them as quote tweets (i.e., quotes). Here we only considered target and quote tweets from cluster #2. Whether a tweet was a target (0) or a quote (1) was used as the outcome variable, and the identity of each target tweet was used as a random effect. In other words, each target and all of its quotes were assigned the same random effect. Like above, we first added tweet length and the presence of media as control variables (for both targets and quotes). Then, we added the compound score, the absolute value of the compound score, the proportion of negative words, the proportion of positive words, and the proportion of neutral words as predictor variables to see which measure of content best improves the model. The absolute value of the compound score and the proportion of neutral words were included as indicators of a general reduction in the intensity of emotion independent of positive or negative valence. All continuous predictor variables were scaled and centered prior to analysis, and model choice and residual diagnostic tests were conducted using all observations (Fig I in the SI).

Results

The *VoterFraud2020* dataset is divided into several sub-communities, including both detractors and proponents of the conspiracy theories. We chose to focus on cluster #2, the "proponent" community that tweets and retweets content in English and does not have significant connections to the "detractor" community (see Methods). After subsetting the *VoterFraud2020* data to only include user and tweet data from cluster #2, we ended up with 3,982,990 tweets from 341,676 users. Note that we calculated the number of users as all unique users that either tweeted or retweeted content from cluster #2. The agent-based model was initialized with a population size (N) of 341,676. The model and methods were preregistered in advance of the analysis (see Methods).

The posterior distributions for content bias (c), follower influence (d), frequency bias (a), and age dependency (g) can be seen in Fig. 2 and Table 1. Higher values of c, d, and g are indicative of stronger effects of those parameters, where 0 is neutrality. Values of a that are lower and higher than 1 are indicative of anticonformity and conformity bias, respectively, where 1 is neutrality. The median estimate for content bias is 4.612, with a 95% credible interval (CI) that spans from 3.479 to 5.679. $c = 4.612$ causes a tweet with $M = 2$ (one standard deviation (SD) above the mean) to be ~24x more likely to be retweeted. This indicates that content bias plays a significant and strong role in driving retweet frequencies.

The posterior distributions for both follower influence (d) and frequency bias (a) are bimodal, with smaller peaks at intermediate values (~1.5 and ~0.6, respectively; Fig. 2). These peaks may correspond to alternate parameter combinations that partly reproduce the real data, but the simpler rejection form of ABC produced unimodal posteriors that converge towards zero and do not have significant probability mass at these smaller peaks (Fig C in the SI). As such, we focus our interpretation on the global maxima. Follower influence has a median estimate of 0.362 with a wide 95% CI and a right-skewed posterior that converges toward zero. This indicates that follower count has a weak effect on retweet frequencies that is either difficult to estimate or varies among users. The median estimate for frequency bias is 0.295 with a similarly wide and right-skewed posterior. If we assume that neutrality in frequency bias is $a = 1$, where retweet probability is perfectly proportional to the number of times a tweet has

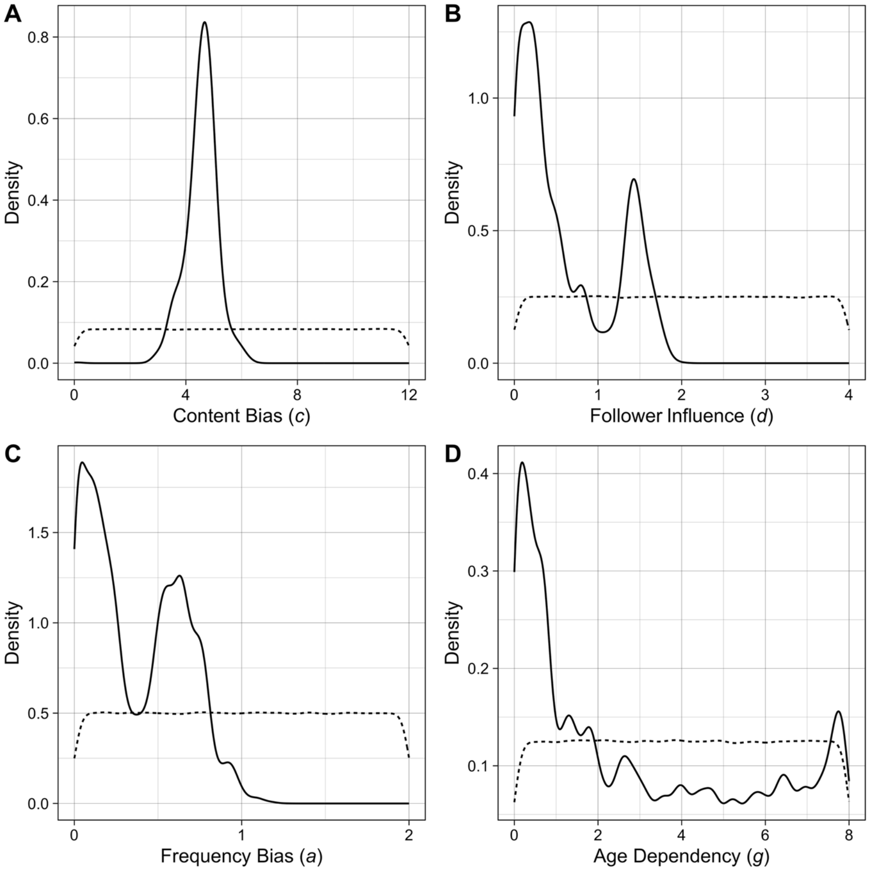

Fig. 2 The prior (dotted lines) and posterior (solid lines) distributions for each of the four dynamic parameters from the ABM that were estimated using ABC. Panel **A** corresponds to content bias (*c*), **B** corresponds to follower influence (*d*), **C** corresponds frequency bias (*a*), and **D** corresponds to age dependency (*g*).

Table 1 The median, 95% credible interval, and out-of-bag normalized mean absolute error of the posterior distribution for each dynamic parameter in the agent-based model.

	M	95% CI	NMAE
c: content bias	4.612	[3.479, 5.679]	0.530
d: follower influence	0.362	[0.007, 1.663]	0.741
a: frequency bias	0.295	[0.004, 0.916]	1.458
g: age dependency	2.043	[0.019, 7.824]	2.016

NMAE stands for normalized mean absolute error.

already been shared, then this result is indicative of an extremely strong anticonformity bias. However, we feel that an anticonformity bias of this magnitude is unrealistic and that this result instead suggests that retweet probability is mostly decoupled from the number of times a tweet has already been shared (so neutrality is $a = 0$). This is much more consistent with our personal experiences on Twitter, where the timeline includes a balance of new tweets from followers and trending tweets that have already been heavily retweeted. According to this interpretation, the wide 95% CI for a could reflect departures from neutrality resulting from conformity bias, variation in user behavior, or how Twitter's recommendation algorithm weights trending tweets.

The posterior distribution for age dependency is also converging towards zero but has high out-of-bag error and probability mass across the entire prior. We chose to make no conclusions about this parameter. Chaotic posterior distributions such as this one can sometimes result from misspecified priors, but we feel that this is unlikely to be the case here as higher values of g generate retweet distributions that are further away from the observed data (Fig E in the SI).

Twitter's recommendation algorithm has just been made publicly available and has not yet been systematically studied (Twitter, 2023), so it is very difficult to differentiate between a bias produced by the algorithm or user behavior. However, we do know that the algorithm primarily uses author and engagement data (Koumchatzky & Andryeyev, 2017) and does not include information about the emotional valence of tweets (https://github.com/twitter/the-algorithm-ml/blob/main/projects/home/recap/FEATURES.md), making this issue most relevant for the results related to follower influence and frequency bias.

Based on the null models for the GLMM, the user appears to be the only grouping variable that explains a high level of variance in the data ($\text{ICC}_{\text{user}} = 0.610$, $\text{ICC}_{\text{date}} = 0.109$, $\text{ICC}_{\text{hour}} = 0.039$). As such, we chose to include user as a random effect in our base model. Adding tweet length as a control variable improved model fit ($\Delta\text{AIC} = 40803$; LRT: $\chi^2 = 40{,}805$, $p < 0.0001$). Adding the presence of media (i.e., images or videos) as a control variable improved model fit ($\Delta\text{AIC} = 510$; LRT: $\chi^2 = 511$, $p < 0.0001$). Both follower count and verification status further improved model fit ($\Delta\text{AIC} > 2$), but the model with follower count was significantly better ($\Delta\text{AIC} = 6204$; LRT: $\chi^2 = 6205$, $p < 0.0001$) so we updated our base model accordingly. All three content measures further improved model fit ($\Delta\text{AIC} > 2$), but the model with the compound score was significantly better than the models with the proportion of negative words ($\Delta\text{AIC} = 831$; LRT: $\chi^2 = 830$, $p < 0.0001$) or positive words ($\Delta\text{AIC} = 1276$; LRT: $\chi^2 = 1276$,

Table 2 The incidence rate ratio (IRR) and 95% confidence interval for each predictor in the best-fitting model.

	IRR	95% CI
Tweet length	1.478	[1.477, 1.480]
Presence of media	1.515	[1.507, 1.523]
Follower count	1.505	[1.496, 1.515]
Compound score	0.959	[0.958, 0.960]

IRR, the exponentiated beta estimate, is interpreted as the rate at which the outcome variable is expected to change per unit increase in a predictor (one standard deviation for scaled and centered predictors). Wald confidence intervals were used due to the high sample size.

$p < 0.0001$). All model specifications and AIC values for the primary GLMM are in Table B in the SI, along with a partial specification curve analysis in Fig G in the SI. The best-fitting model included the user as a random intercept, and tweet length, the presence of media, follower count, and the compound score as fixed effects (see Table 2). Note that we chose to not report p-values for our GLMM results, given the unreliability of p-values at higher sample sizes (Halsey, 2019). Instead, we focused on the effect sizes (calculated here as incidence rate ratios by applying the exponential function to model coefficients) and 95% confidence intervals, where intervals that do not overlap with 1 are interpreted as evidence for a significant effect.

Tweet length, presence of media, follower count, and compound score all have significant effects on retweet frequency (Table 2). The incidence rate ratio (IRR) for tweet length is 1.478, indicating that if a tweet is one SD longer then it is 47.8% more likely to be retweeted. The IRR for the presence of media is 1.515, so tweets with images or videos are 51.5% more likely to be retweeted. Follower count has a similarly strong effect, where tweets from users with one SD more followers are 50.5% more likely to be retweeted. The IRR for the compound score is much lower but still significant. Tweets with a compound score that is one SD lower (more strongly negative) are 5.7% more likely to be retweeted. The models that include the proportion of words that are negative or positive instead of the compound score, fit during model choice and show similar effects: tweets with one SD with more negative words are 8.0% more likely to be retweeted, and tweets with one SD more positive words are 6.4% less likely to be retweeted. Pseudo R^2 values, calculated using log-normal approximation, indicate that the fixed effects alone account for about 10% of the variance in the data ($R^2 = 0.101$), whereas the fixed effects and random intercept together account for about 68% of the variance in the data ($R^2 = 0.676$) (Nakagawa et al., 2017). A variance inflation factor test indicates that there are no significant issues with multicollinearity between predictors (VIFs < 2). Residual diagnostics for the best fitting model indicate that, while there are some extreme low and high outliers, the Poisson family is appropriate and there are no significant problems with overdispersion (see Fig H in the SI).

An additional GLMM found that quote tweets tend to have reduced negative valence relative to the original tweets that they are quoting (see SI). This lends support to our generative inference results, given our decision to treat quote tweets like original tweets when computing retweet distributions (see Methods). If quote tweets tend to be less attractive than original tweets, then our estimate for content bias is likely more conservative than it would have been if we had treated quote tweets like retweets instead of original tweets.

Discussion

Based on the results of generative inference, the observed retweet distribution is consistent with a strong content bias. Tweets with a lower compound score are more likely to be retweeted, suggesting that this bias is targeted toward negative emotional valence. Follower count, on the other hand, has a weak effect in the agent-based model and a strong effect in the GLMM. This discrepancy is likely due to the fact that the former accounts for key temporal dynamics that cannot be included in the latter, such as activity levels over time, the effect of current retweet count on future retweet probability, and age dependency. For frequency bias, the results of generative inference could be interpreted as evidence for either an extremely strong anticonformity bias or for frequency information being irrelevant. We feel that an anticonformity bias of this magnitude is unrealistic, and we interpret this result as evidence that retweet probability is mostly decoupled from the number of times a tweet has already been retweeted. Interestingly, we also found that quote tweets tend to contain less negative emotional valence than their targets. This means that users do not tend to amplify negativity when commenting on a retweet, despite having a content bias for negative valence. Quote tweets are often thought to reflect criticism, so this reduced negativity may be unique to discussions between like-minded users.

Importantly, the results related to follower influence and frequency bias are extremely difficult to separate from the influence of Twitter's recommendation algorithm, which is heavily based on user characteristics and engagement (Koumchatzky & Andryeyev, 2017; Twitter, 2023). The algorithm has just been made public and has not yet been systematically studied (Twitter, 2023), so we do not know how much it prioritizes tweets that are popular or from followed users. In the case of follower count, this makes it impossible to disentangle the effect of network structure from a demonstrator bias. Luckily, we do know that the recommendation algorithm does not use information about the emotional valence of tweets (https://github.com/twitter/the-algorithm-ml/blob/main/projects/home/recap/FEATURES.md). We were unable to collect the follower network because of account suspensions following the January 6th attack (Abilov et al., 2021), but even if we had access to the network structure the relevance of it would depend upon user behavior. If users primarily share information that they see passively on their timeline, then network structure is much more important, whereas if users are searching with keywords and hashtags, then it is less so. In this case, we speculate that the weak effect of follower count is due to the latter. Proponents of conspiracy theories may actively seek out messages using keywords and hashtags, which are known to be important in the spread of disinformation (Hindman & Barash, 2018), whereas more generalist Twitter users may rely more on the timeline and be more impacted by network structure (Asatani et al., 2021). People may also vary in how selectively they retweet particular users or groups (e.g., only retweeting former President Trump). Based on the confounding effects of the algorithm and network structure, we are most confident in concluding that the spread of voter fraud claims among proponents of voter fraud conspiracy theories on Twitter during and after the 2020 US election was partly driven by a content bias causing users to preferentially retweet tweets with more negative emotional valence. Our other findings related to follower influence and frequency bias should be viewed as provisional and interpreted with caution—to be investigated further once the recommendation algorithm has been systematically studied.

Our results are consistent with previous work suggesting that emotionally negative content has an advantage on social media across a variety of domains, including news coverage and political discourse (Asatani et al., 2021; Bellovary et al., 2021; Del Vicario et al., 2016; Schöne et al., 2021). Other studies, though, have shown that positive messages spread more slowly but reach more people (Ferrara & Yang, 2015b), that exposure to both positive and negative tweets increases the probability of a user tweeting content with similar emotional valence (Ferrara & Yang, 2015a), and that tweets with greater emotional intensity (independent of valence) are more likely to be retweeted (Brady et al., 2017; Stieglitz & Dang-Xuan, 2013). In such cases, there seems to be significant variation across domains and individuals. Messages about same-sex marriage, for example, are more likely to be retweeted if they use positive language, whereas messages about climate change are more likely to be retweeted if they use negative language (Brady et al., 2017). We suspect that conspiracy theory content generally falls into the latter category. Similarly, another study found that there is variation in how users respond to emotional content, where some "highly susceptible" users are more likely to be influenced by positive messages (Ferrara & Yang, 2015a).

Beyond conspiracy theories and social media, negativity bias is a widespread psychological phenomenon (Baumeister et al., 2001) that is thought to be adaptive because negative events are more relevant to survival (Rozin & Royzman, 2001). Negative information (e.g., potential threats, contagions, injuries) is transmitted with higher fidelity in urban legends and stories (Acerbi, 2022; Bebbington et al., 2017; Eriksson & Coultas, 2014; Fay et al., 2021; Heath et al., 2001) and is overrepresented in supernatural beliefs around the world (Fessler et al., 2014). Interestingly, people are also more likely to interpret negative information as credible, presumably because the cost of erroneously ignoring real hazards can be very high (Fessler et al., 2014). This tendency towards "negatively biased credulity" appears to be higher in political conservatives in the US (Fessler et al., 2017), who tend to view the world as more dangerous than liberals, regardless of the party that is currently in power (Samore et al., 2018). People in heightened emotional states are also more likely to believe fake news stories (Martel et al., 2020). Future studies could collect survey data from social media users, possibly in a pseudo-experimental context (Butler et al., 2023), to explore these interactions between ideology, emotional state, information sharing, and belief in more detail.

Most recently, Pröllochs et al. (Pröllochs et al., 2021) found that false rumors with more positive language are more likely to go viral, despite the fact that articles espousing false rumors tend to have more negative language (Acerbi, 2019). The discrepancy between our findings and this evidence for positivity bias in "fake news" is puzzling, but we think it is due to differences in granularity and methodology. Pröllochs et al. worked with unique news-related rumors spreading through a general Twitter population (Pröllochs et al., 2021; Vosoughi et al., 2018) whereas we analyzed sub-conversations about the same topic within a single conspiracy theory community. Content biases for emotional valence may vary when messages are shared between like-minded individuals or embedded within a single conversation. Additionally, while Pröllochs et al.'s methodology is based on a well-established and cross-culturally validated emotion model, their primary valence measure has a low correlation with human raters when applied to social media data ($r = 0.11$) (Pröllochs et al., 2021) relative to VADER ($r = 0.88$) (Hutto & Gilbert, 2014) and other methods like SentiStrength (>60% accuracy) (Ferrara & Yang, 2015a).

The sentiment analysis in this study has several limitations that should be highlighted. First, some new phrases that emerged during the 2020 election (e.g., "stop the steal") may differ in emotional valence compared to their constituent words in VADER's lexicon (Hutto & Gilbert, 2014). Second, several of the common words and phrases used to create the *VoterFraud2020* dataset (e.g., "fraud") have negative valence in VADER's lexicon. Luckily, this only appears to introduce a slight skew in our valence measure (Fig F in the SI), which is common in studies of emotion and sentiment on social media (Burton et al., 2021; Ferrara & Yang, 2015b, 2015a; Pröllochs et al., 2021) and is unlikely to affect our results as GLMMs make no distributional assumptions about predictors (Stroup, 2013). To improve the robustness of modeling of emotional contagion on social media, Burton et al. (Burton et al., 2021) recently came up with three recommendations for future studies: going beyond correlational evidence, analyzing the effect of specification decisions on model estimates, and preregistration. We fully agree with these recommendations, and we hope that we adequately addressed them by using a preregistered generative inference framework to ensure that the data was consistent with transmission bias before conducting GLMM and ensuring that our estimates were robust across a reasonable range of modeling specifications.

Regarding the spread of conspiracy theories, previous research has proposed "herd behavior", in which rational individuals with limited information defer to the beliefs of the majority, to be a potential explanation (Sunstein, 2014a, 2014b). Our study addresses the sharing of conspiracist tweets among proponents, who presumably already believed some voter fraud claims before the election took place, but our lack of clear evidence for a frequency bias suggests that a disproportionate tendency to "follow the herd" may not be the primary driver of the spread of conspiracy theory messages. Rather, our study suggests that the content of conspiracy theory messages were, in the case we studied, more salient cues for cultural transmission. That being said, we were unable to incorporate many of the contextual and individual-level factors that have been associated with conspiracy theory beliefs in survey-based and experimental studies, such as anxiety (Radnitz & Underwood, 2017), age (Guess et al., 2019), demographic diversity among social ties (Min, 2021), and morbid curiosity (Scrivner & Stubbersfield, 2022). Such data are difficult, if not impossible, to infer from social media corpora, but could be included in future analyses based on more detailed profiles of individuals who have adopted conspiracy theory beliefs (as has been done in the extremism literature (Becker, 2019; Youngblood, 2020).

While recognized as important, the transmission processes involved in the spread of conspiracy theories have received relatively little attention in research and are not well understood (Bangerter et al., 2020). This study demonstrates the value of cultural evolutionary approaches for understanding the transmission processes at play on social networks, and it highlights the importance of considering the roles of both the content of conspiracy theories and the context in which they are shared. Identifying and characterizing the biases influencing the transmission of conspiracy theories can help us to generate potential methods for countering the spread of harmful conspiracy theories and promoting the spread of genuine information (Salali & Uysal, 2021).

A previous study using generative inference to investigate behavior on Twitter found that retweet patterns of both confirmed and debunked information were more consistent with unbiased random copying than with conformity (Carrignon et al., 2019). At first glance, our study seems to contradict this result, but Carrignon et al. (Carrignon et al., 2019) did not include a parameter for follower influence in their agent-based model, and they assumed neutrality for content bias due to computational limitations. When we ran our agent-based model with neutral values for both content bias and follower influence, we too found that the model best fit the observed data when copying was unbiased by frequency (see Fig D in the SI). The discrepancy between the results, when parameters are estimated together instead of individually, highlights the importance of considering equifinality—the fact that different processes can lead to similar patterns at the population level (Barrett, 2019). If different processes lead to only subtle differences in retweet frequencies, then the effect of one could be mistakenly attributed to another if both are not considered simultaneously. For example, in our study, we found that the observed retweet distribution was consistent with content bias when content bias was estimated alongside other parameters (Fig. 2 and Table 1), but not when it was estimated in isolation (Fig D in the SI).

One of our biggest takeaways from this study is how challenging it is to construct realistic null models of behavior on social media without detailed information about recommendation algorithms. We are very happy to see that Twitter has just released its algorithm to the public (Twitter, 2023), but there is only so much that can be inferred from a neural network without the original training data. Instead, the company could try to infer how the algorithm boosts different kinds of content by running natural experiments on the platform, as has been recently done for racial and gender bias in the image cropping algorithm (Agrawal & Davis, 2020) and right-leaning political bias in the recommendation algorithm (Huszár et al., 2021). Twitter could also publish the results of a model in which simulated users randomly share information from simulated timelines constructed by their algorithm to see how different kinds of content spread under neutral conditions. Such an analysis would make it possible, for example, to parse out what portion of frequency dependence in retweeting is due to the algorithm as opposed to conformity among users. Luckily, algorithmic transparency and accountability are increasingly prioritized by governments around the world (Koene et al., 2019; *Algorithmic Justice and Online Platform Transparency Act*, 2021). We hope that future studies can take advantage of improved transparency to develop more effective policy recommendations for fighting the spread of conspiracy theories and disinformation on social media platforms.

In conclusion, our methodology, based on a cultural evolution framework, allowed us to weigh the relative importance of different features influencing the spread of voter fraud claims among conspiracy theorists on Twitter. Most importantly, we found that retweet frequencies of voter fraud messages posted during and after the 2020 US election are consistent with a content bias for tweets with more negative emotional valence. While previous research focused a priori on the role of tweets' content, without evaluating other possibilities, we were able to show that content is indeed central when compared with other possible mechanisms of social influence. Future studies could apply our methodology in a comparative framework to assess whether content bias for emotional valence in conspiracy theory messages differs from other forms of information on social media.

Data availability

The agent-based model, analysis code, and processed data (everything required for replication) used in this study can be found on *Zenodo* (https://doi.org/10.5281/zenodo.8311560) and GitHub (https://github.com/masonyoungblood/TwitterABM). The full anonymized *VoterFraud2020* dataset can be found on Abilov et al.'s website (https://github.com/sTechLab/VoterFraud2020). The full disambiguated dataset with tweet text is available from Abilov et al. upon request (mor.naaman@cornell.edu) to respect users' privacy.

Received: 16 March 2023; Accepted: 6 September 2023;
Published online: 14 September 2023

References

Abilov A, Hua Y, Matatov H, et al. (2021) VoterFraud2020: a multi-modal dataset of election fraud claims on Twitter. arXiv. https://doi.org/10.48550/arXiv.2101.08210

Acerbi A (2019) Cognitive attraction and online misinformation. Palgrave Commun 5:15. https://doi.org/10.1057/s41599-019-0224-y

Acerbi A (2022) From storytelling to Facebook: content biases when retelling or sharing a story. Hum Nat 33:132–144. https://doi.org/10.1007/s12110-022-09423-1

Agrawal P, Davis D (2020) Transparency around image cropping and changes to come. In: Twitter's Prod Blog. https://blog.twitter.com/en_us/topics/product/2020/transparency-image-cropping

Albertson B, Guiler K (2020) Conspiracy theories, election rigging, and support for democratic norms. Res Polit 7:2053168020959859. https://doi.org/10.1177/2053168020959859

Asatani K, Yamano H, Sakaki T, Sakata I (2021) Dense and influential core promotion of daily viral information spread in political echo chambers. Sci Rep 11:7491. https://doi.org/10.1038/s41598-021-86750-w

Balsamo M (2020) Disputing Trump, Barr says no widespread election fraud. Assoc Press

Bangerter A, Wagner-Egger P, Delouvée S (2020) How conspiracy theories spread. In: Butter M, Knight P (eds) Routledge handbook of conspiracy theories. Routledge, New York, pp. 206–218

Barrett BJ (2019) Equifinality in empirical studies of cultural transmission. Behav Processes 161:129–138. https://doi.org/10.1016/j.beproc.2018.01.011

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Softw 67:1–48. https://doi.org/10.18637/jss.v067.i01

Baumeister RF, Bratslavsky E, Finkenauer C, Vohs KD (2001) Bad is stronger than good. Rev Gen Psychol 5:323–370. https://doi.org/10.1037/1089-2680.5.4.323

Bebbington K, MacLeod C, Ellison TM, Fay N (2017) The sky is falling: evidence of a negativity bias in the social transmission of information. Evol Hum Behav 38:92–101. https://doi.org/10.1016/j.evolhumbehav.2016.07.004

Becker MH (2019) When extremists become violent: examining the association between social control, social learning, and engagement in violent extremism. Stud Confl Terror 1–21. https://doi.org/10.1080/1057610X.2019.1626093

Beckett L (2021) Millions of Americans think the election was stolen. How worried should we be about more violence? Guard

Bellovary AK, Young NA, Goldenberg A (2021) Left- and right-leaning news organizations use negative emotional content and elicit user engagement similarly. Affect Sci. https://doi.org/10.1007/s42761-021-00046-w

Blum MGB, François O (2010) Non-linear regression models for Approximate Bayesian Computation. Stat Comput 20:63–73. https://doi.org/10.1007/s11222-009-9116-0

Brady WJ, Wills JA, Jost JT et al. (2017) Emotion shapes the diffusion of moralized content in social networks. Proc Natl Acad Sci USA 114:7313–7318. https://doi.org/10.1073/pnas.1618923114

Brotherton R (2015) Suspicious minds: why we believe conspiracy theories. Bloomsbury Sigma, New York

Burton JW, Cruz N, Hahn U (2021) Reconsidering evidence of moral contagion in online social networks. Nat Hum Behav 5:1629–1635. https://doi.org/10.1038/s41562-021-01133-5

Butler LH, Lamont PX, Wan DBLY, et al. (2023) The (Mis)Information Game: a social media simulator. Behav Res Methods. https://doi.org/10.3758/s13428-023-02153-x

Butter M, Knight P (2020) General introduction. In: Butter M, Knight P (eds) Routledge Handbook of Conspiracy Theories, 1st edn. Routledge, New York, pp. 1–8

Carrignon S, Bentley RA, Ruck D (2019) Modelling rapid online cultural transmission: evaluating neutral models on Twitter data with approximate Bayesian computation. Palgrave Commun 5. https://doi.org/10.1057/s41599-019-0295-9

Chao A, Gotelli NJ, Hsieh TC et al. (2014) Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecol Monogr 84:45–67. https://doi.org/10.1890/13-0133.1

Cillizza C (2021) 1 in 3 Americans believe the "Big Lie." CNN

Cohen L (2021) 6 conspiracy theories about the 2020 election—debunked. CBS News

Conger K, Isaac M (2021) Twitter permanently bans Trump, capping online revolt. New York Times

Cooper JJ, Christie B (2021) Election conspiracies live on with audit by Arizona GOP. Assoc Press

Corasaniti N, Epstein RJ (2021) What Georgia's voting law really does. New York Times

Cottrell D, Herron MC, Westwood SJ (2018) An exploration of Donald Trump's allegations of massive voter fraud in the 2016 General Election. Elect Stud 51:123–142. https://doi.org/10.1016/j.electstud.2017.09.002

Del Vicario M, Vivaldo G, Bessi A et al. (2016) Echo chambers: emotional contagion and group polarization on Facebook. Sci Rep 6:37825. https://doi.org/10.1038/srep37825

Dentith M (2014) The philosophy of conspiracy theories. Palgrave Macmillan

Douglas KM, Uscinski JE, Sutton RM et al. (2019) Understanding conspiracy theories. Polit Psychol 40:3–35. https://doi.org/10.1111/pops.12568

Edelson J, Alduncin A, Krewson C et al. (2017) The effect of conspiratorial thinking and motivated reasoning on belief in election fraud. Polit Res Q 70:933–946. https://doi.org/10.1177/1065912917721061

Enders AM, Uscinski JE, Klofstad CA et al. (2021) The 2020 presidential election and beliefs about fraud: continuity or change. Elect Stud 72:102366. https://doi.org/10.1016/j.electstud.2021.102366

Eriksson K, Coultas JC (2014) Corpses, maggots, poodles and rats: emotional eelection operating in three phases of cultural transmission of urban legends. J Cogn Cult 14:1–26. https://doi.org/10.1163/15685373-12342107

Fay N, Walker B, Kashima Y, Perfors A (2021) Socially situated transmission: The bias to transmit negative information is moderated by the social context. Cogn Sci 45:1–17. https://doi.org/10.1111/cogs.13033

Ferrara E, Yang Z (2015a) Quantifying the effect of sentiment on information diffusion in social media. PeerJ Comput Sci 1. https://doi.org/10.7717/peerj-cs.26

Ferrara E, Yang Z (2015b) Measuring emotional contagion in social media. PLoS ONE 10:1–14. https://doi.org/10.1371/journal.pone.0142390

Fessler DMT, Pisor AC, Holbrook C (2017) Political orientation predicts credulity regarding putative hazards. Psychol Sci 28:651–660. https://doi.org/10.1177/0956797617692108

Fessler DMT, Pisor AC, Navarrete CD (2014) Negatively-biased credulity and the cultural evolution of beliefs. PLoS One 9. https://doi.org/10.1371/journal.pone.0095167

Goertzel T (1994) Belief in conspiracy theories. Polit Psychol 15:731–742. https://doi.org/10.2307/3791630

Goldberg RA (2003) Conspiracy theories in America: a historical overview. Conspir Theor Am Hist An Encycl 1–13

Guess A, Nagler J, Tucker J (2019) Less than you think: prevalence and predictors of fake news dissemination on Facebook. Sci Adv 5:eaau4586. https://doi.org/10.1126/sciadv.aau4586

Hall Jamieson K, Albarracín D (2020) The relation between media consumption and misinformation at the outset of the SARS-CoV-2 pandemic in the US. Harvard Kennedy Sch Misinformation Rev 1. https://doi.org/10.37016/mr-2020-012

Halsey LG (2019) The reign of the p-value is over: What alternative analyses could we employ to fill the power vacuum? Biol Lett 15. https://doi.org/10.1098/rsbl.2019.0174

Heath C, Bell C, Sternberg E (2001) Emotional selection in memes: the case of urban legends. J Pers Soc Psychol 81:1028–1041. https://doi.org/10.1037/0022-3514.81.6.1028

Hindman M, Barash V (2018) Disinformation, "fake news" and influence campaigns on Twitter

Huszár F, Ktena SI, O'Brien C et al. (2021) Algorithmic amplification of politics on Twitter. Twitter

Hutto CJ, Gilbert E (2014) VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media. pp. 216–225

Imhoff R, Dieterle L, Lamberty P (2020) Resolving the puzzle of conspiracy worldview and political activism: belief in secret plots decreases normative but increases nonnormative political engagement. Soc Psychol Personal Sci 12:71–79. https://doi.org/10.1177/1948550619896491

Jolley D, Douglas KM (2014) The social consequences of conspiracism: exposure to conspiracy theories decreases intentions to engage in politics and to reduce one's carbon footprint. Br J Psychol 105:35–56. https://doi.org/10.1111/bjop.12018

Jolley D, Meleady R, Douglas KM (2020) Exposure to intergroup conspiracy theories promotes prejudice which spreads across groups. Br J Psychol 111:17–35. https://doi.org/10.1111/bjop.12385

Kandler A, Powell A (2018) Generative inference for cultural evolution. Philos Trans R Soc B Biol Sci 373. https://doi.org/10.1098/rstb.2017.0056

Keeley BL (1999) Of conspiracy theories. J Philos 96:109–126. https://doi.org/10.2307/2564659

Kendal RL, Boogert NJ, Rendell L et al. (2018) Social learning strategies: bridge-building between fields. Trends Cogn Sci 22:651–665. https://doi.org/10.1016/j.tics.2018.04.003

Koene A, Clifton C, Hatada Y et al. (2019) A governance framework for algorithmic accountability and transparency. European Parliamentary Research Service

Kofta M, Soral W, Bilewicz M (2020) What breeds conspiracy antisemitism? The role of political uncontrollability and uncertainty in the belief in Jewish conspiracy. J Pers Soc Psychol 118:900–918. https://doi.org/10.1037/pspa0000183

Koumchatzky N, Andryeyev A (2017) Using deep learning at scale in Twitter's timelines. In: Twitter's Eng Blog

Lachlan RF, Ratmann O, Nowicki S (2018) Cultural conformity generates extremely stable traditions in bird song. Nat Commun 9. https://doi.org/10.1038/s41467-018-04728-1

Martel C, Pennycook G, Rand DG (2020) Reliance on emotion promotes belief in fake news. Cogn Res Princ Implic 5. https://doi.org/10.1186/s41235-020-00252-3

Mesoudi A (2011) Cultural evolution: how Darwinian theory can explain human culture and synthesize the social sciences. University of Chicago Press

Milli S, Carroll M, Pandey S, et al. (2023) Twitter's algorithm: amplifying anger, animosity, and affective polarization. arXiv 1–14. https://doi.org/10.48550/arXiv.2305.16941

Min SJ (2021) Who believes in conspiracy theories? Network diversity, political discussion, and conservative conspiracy theories on social media. Am Polit Res 49:415–427. https://doi.org/10.1177/1532673X211013526

Nakagawa S, Johnson PCD, Schielzeth H (2017) The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. J R Soc Interface 14:20170213. https://doi.org/10.1098/rsif.2017.0213

Olmsted K (2018) Conspiracy theories in US history. In: Uscinski JE (ed.) Conspiracy theories and the people who believe them. Oxford University Press, pp. 285–297

Osmundsen M, Bor A, Vahstrup PB et al. (2021) Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. Am Polit Sci Rev 115:999–1015. https://doi.org/10.1017/S0003055421000290

Pagán VE (2020) Conspiracy theories in the Roman Empire. In: Butter M, Knight P (eds) Routledge handbook of conspiracy theories, 1st edn. Routledge, New York

Pigden C (1995) Popper revisited, or what Is wrong with conspiracy theories. Philos Soc Sci 25:3–34. https://doi.org/10.1177/004839319502500101

Probst P, Wright M, Boulesteix A-L (2018) Hyperparameters and tuning strategies for random forest. Wiley Interdiscip Rev Data Min Knowl Discov. https://doi.org/10.1002/widm.1301

Pröllochs N, Bär D, Feuerriegel S (2021) Emotions explain differences in the diffusion of true vs. false social media rumors. Sci Rep 11:22721. https://doi.org/10.1038/s41598-021-01813-2

Radnitz S, Underwood P (2017) Is belief in conspiracy theories pathological? A survey experiment on the cognitive roots of extreme suspicion. Br J Polit Sci 47:113–129. https://doi.org/10.1017/S0007123414000556

Rathje S, Van Bavel JJ, van der Linden S (2021) Out-group animosity drives engagement on social media. Proc Natl Acad Sci USA 118:e2024292118. https://doi.org/10.1073/pnas.2024292118

Raynal L, Marin J-M, Pudlo P et al. (2019) ABC random forests for Bayesian parameter inference. Bioinformatics 35:1720–1728. https://doi.org/10.1093/bioinformatics/bty867

Romm T, Dwoskin E (2021) Twitter purged more than 70,000 affiliated with QAnon following Capitol riot. Washington Post

Roswell M, Dushoff J, Winfree R (2021) A conceptual guide to measuring species diversity. Oikos 130:321–338. https://doi.org/10.1111/oik.07202

Rozin P, Royzman EB (2001) Negativity bias, negativity dominance, and contagion. Personal Soc Psychol Rev 5:296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Frank SA (2009) The common patterns of nature. J Evol Biol 22:1563–1585. https://doi.org/10.1111/j.1420-9101.2009.01775.x

Salali GD, Uysal MS (2021) Effective incentives for increasing COVID-19 vaccine uptake. Psychol Med 1–3. https://doi.org/10.1017/S0033291721004013

Samore T, Fessler DMT, Holbrook C, Sparks AM (2018) Electoral fortunes reverse, mindsets do not. PLoS ONE 13:1–15. https://doi.org/10.1371/journal.pone.0208653

Sardarizadeh S, Lussenhop J (2021) The 65 days that led to chaos at the Capitol. BBC News

Schöne JP, Parkinson B, Goldenberg A (2021) Negativity spreads more than positivity on Twitter after both positive and negative political situations. Affect Sci. https://doi.org/10.1007/s42761-021-00057-7

Scrivner C, Stubbersfield JM (2022) Curious about threats: Morbid curiosity and interest in conspiracy theories. OSF. https://doi.org/10.31219/osf.io/7fubx

Sisson SA, Fan Y, Beaumont MA (2018) Handbook of Approximate Bayesian Computation. CRC Press

Skelley G (2021) Most Republicans still won't accept that Biden won. FiveThirtyEight

Spring M (2020) "Stop the steal": The deep roots of Trump's "voter fraud" strategy. BBC New

Stempel C, Hargrove T, Stempel GH (2007) Media use, social structure, and belief in 9/11 conspiracy theories. J Mass Commun Q 84:353–372. https://doi.org/10.1177/107769900708400210

Stieglitz S, Dang-Xuan L (2013) Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. J Manag Inf Syst 29:217–248. https://doi.org/10.2753/MIS0742-1222290408

Stroup WW (2013) Generalized linear mixed models: modern concepts, methods and applications. CRC Press

Sunstein CR (2014a) Conspiracy theories and other dangerous ideas. Simon & Schuster, New York

Sunstein CR (2014b) On rumors: how falsehoods spread, why we believe them, and what can be done. Princeton University Press, Princeton

Tucker E, Bajak F (2020) Repudiating Trump, officials say election "most secure." Assoc. Press

Twitter (2023) Twitter's recommendation algorithm. In: Twitter's Eng. Blog. https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm. Accessed 6 Apr 2023

Uscinski JE, Parent JM (2014) American conspiracy theories. Oxford University Press, New York

van Prooijen J-W, Ligthart J, Rosema S, Xu Y (2021) The entertainment value of conspiracy theories. Br J Psychol. https://doi.org/10.1111/bjop.12522

Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359:1146–1151. https://doi.org/10.1126/science.aap9559

West HG, Sanders T (2003) Transparency and conspiracy: ethnographies of suspicion in the new world order. Duke University Press, Durham

Youngblood M (2020) Extremist ideology as a complex contagion: the spread of far-right radicalization in the United States between 2005-2017. Humanit Soc Sci Commun 7. https://doi.org/10.1057/s41599-020-00546-3

Youngblood M, Lahti D (2022) Content bias in the cultural evolution of house finch song. Anim Behav 185:37–48. https://doi.org/10.1016/j.anbehav.2021.12.012

Zwierlein C (2020) Conspiracy theories in the middle ages and the early modern period. In: Butter M, Knight P (eds) Routledge handbook of conspiracy theories, 1st edn. Routledge, New York, pp. 542–554

Acknowledgements

We would like to thank Anne Kandler for providing us with feedback on our agent-based modeling. This research was supported, in part, under National Science Foundation Grants CNS-0958379, CNS-0855217, ACI-1126113, and the City University of New York High Performance Computing Center at the College of Staten Island. This work has received funding from the "Frontiers in Cognition" EUR grant, ANR-17-EURE-0017 EUR.

Author contributions

MY developed the agent-based model and conducted all statistical modeling. RG conducted the sentiment analysis. AA and OM provided feedback on the agent-based model and statistical modeling. MY, JS, AA, and OM contributed to writing the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-023-02106-x.

Correspondence and requests for materials should be addressed to Mason Youngblood.

Reprints and permission information is available at http://www.nature.com/reprints

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