Trade-offs between reducing misinformation and politically-balanced enforcement on social media

Mohsen Mosleh[1,2*], Qi Yang[3*], Tauhid Zaman[4], Gordon Pennycook[5,6], David G. Rand[2,3,7‡]

[1]Management Department, University of Exeter Business School; [2]Sloan School of Management, Massachusetts Institute of Technology; [3]Institute for Data, Systems, and Society, Massachusetts Institute of Technology; [4]Yale School of Management, Yale University; [5]Hill/Levene Schools of Business, University of Regina; [6]Department of Psychology, University of Regina, [7]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

[*]Joint first authors

[‡]Corresponding Author: drand@mit.edu

This working paper has not yet undergone peer-review.
First version: Apr 8 2022
This version: Jan 25 2023

An analysis of Twitter data shows that the tendency for conservative users to be suspended at higher rates than liberal users can be largely explained by conservative users sharing more links to low quality news sites; this partisan asymmetry in sharing behavior creates a trade-off between reducing the spread of misinformation and maintaining political balance in enforcement.

Mass communication is a central feature of modern life, with social media playing an increasingly important role in the global distribution and consumption of information. These changes in the world's information ecosystem are being accompanied by a rapid co-evolution of technology, cultural norms, and public policy. Although social media companies are constrained by government regulations to some extent (e.g., the E.U.'s General Data Protection Regulation), they set their own internal policies across a wide range of content. Platforms have thus far largely had free reign, for example, over their content moderation policies and have developed their own institutions for managing such policies (e.g., the Facebook Oversight Board).

As a result of this freedom, social media companies have faced widespread charges of bias in the policies they have adopted. Some of the loudest such accusations have arisen in the context of concerns about disinformation, misinformation, and "fake news" (*1*). There is substantial public pressure on platforms to reduce the spread of inaccurate content. For example, both liberals and conservatives in the United States believe technology companies should take action against misinformation (*2*), as do many people across European Union member countries (*3*); and governments around the world have begun to regulate misinformation on social media (*4*). In response, social media companies have implemented a wide range of anti-misinformation policies in recent years, such as removing or flagging posts deemed to be false by professional fact-checkers, using ranking algorithms to reduce the likelihood that users see potentially inaccurate posts, and suspending users who spread misinformation (*5*).

These policies, however, have often led to social media companies being accused of political bias in their choices about who and what to take action against – in the United States, for example, that conservatives and Republicans are purposefully targeted for enforcement because of their political orientation (e.g., when Donald Trump said that Twitter "totally silences conservatives' voices" (*6*)). As misinformation is notoriously hard to define, there is substantial room for bias to creep into subjective judgments about what to sanction and when. Decades of research in psychology have shown the pervasive effects of partisan bias in clouding judgment, including specifically in the context of judgments about misinformation (*7*). Thus, given that the employees of social media companies - as well as the professional fact-checkers they partner with - are typically left-leaning (e.g., (*8*)), it is plausible that they may exhibit bias against conservatives when deciding what counts as misinformation or who to take action against.

At the same time, however, people *can* sometimes override intuitive biases through conscious effort. Social media companies are surely keenly aware of the accusations of political bias made against them, and therefore may correct (or even over-correct) such biases. In other words, the fact that the social media companies are largely liberal does not necessarily mean that their policies exhibit anti-conservative bias. The existence of such bias remains an open empirical question.

We argue that reducing misinformation, and appearing to have political bias, are inexorably intertwined. Although there is a bi-partisan desire for action against misinformation, there is considerable evidence of a partisan *asymmetry* in the sharing of misinformation. During the 2016 election in the United States, for example, news from websites that journalists deemed to be low-quality "fake news" sites were shared much more by conservatives than liberals on Facebook and Twitter (*9, 10*); and survey experiments that present participants with politically balanced sets of headlines (removing the supply-side confound present in many observational studies) typically find that conservatives share more articles deemed to be false by professional fact-checkers than liberals (*11*). Beyond the U.S., a survey experiment conducted in 16 countries found that conservatives shared more COVID-19 misinformation than liberals across many countries (*12*); and an examination of Twitter data found that political elites on the right shared links to lower quality news sites than political elites on the left in the United States, Germany, and the UK (*13*). Of course, a natural objection to these prior results is that they rely on the evaluation of journalists and professional fact-checkers, who may themselves have a liberal bias. Below, we present new data showing a similar pattern among U.S. participants when using the evaluations of politically-balanced crowds of laypeople, which cannot be accused of having liberal bias.

If there is indeed a political asymmetry in misinformation sharing, this means that politically neutral enforcement against misinformation by the platforms - aimed at satisfying the bipartisan demand for a reduction in online misinformation - could lead to political asymmetries in who faces sanctions. If misinformation is shared by conservatives more so than liberals, then suspending users who share falsehoods will lead even a politically-neutral set of enforcement policies to preferentially sanction conservatives.

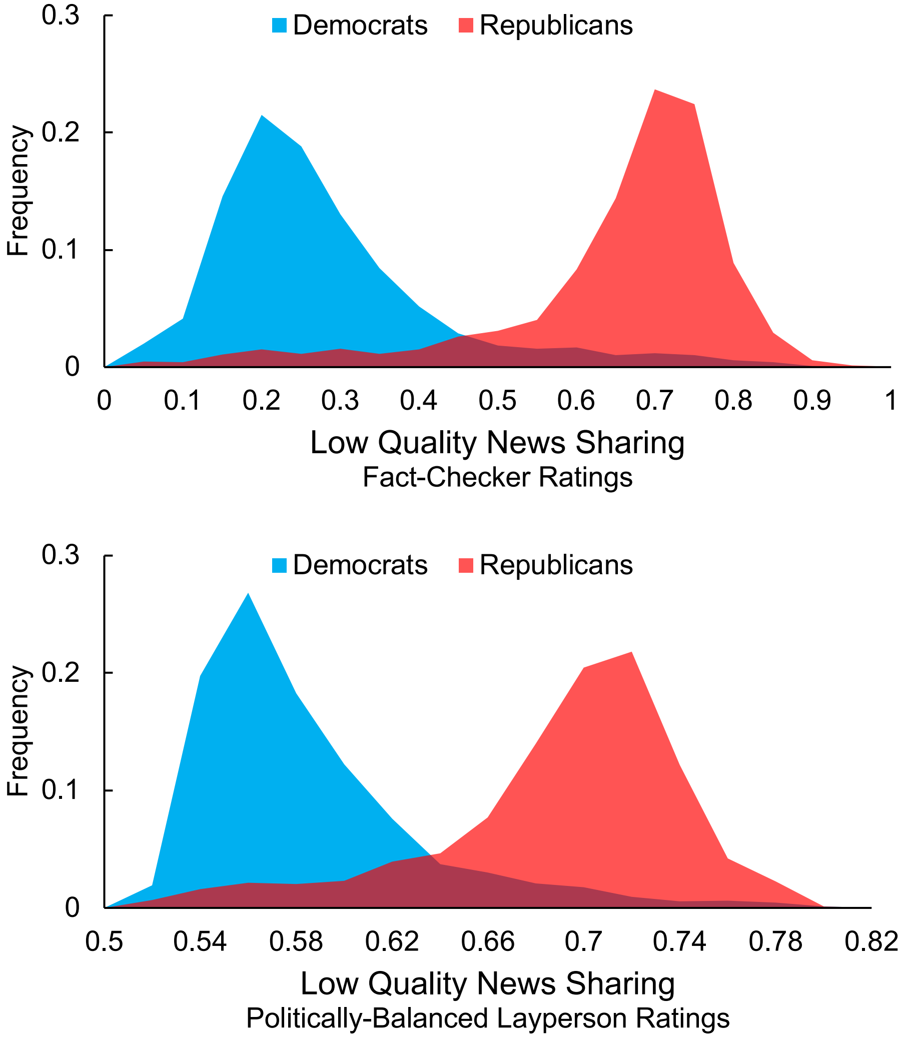
Here, we shed new empirical light on this issue, taking a specific social media policy choice that has drawn intense criticism as a case study: Twitter's suspension of users following the 2020 U.S. Presidential Election. Specifically, in October 2020 we identified 100,000 Twitter users who shared hashtags related to the U.S. Presidential Election, and randomly sampled 4,500 of those users who shared at least one #VoteBidenHarris2020 hashtag and 4,500 who shared at least one #Trump2020 hashtag. We used each user's data from that pre-election time period to quantify their tendency to share low quality news (as well as numerous other potentially relevant characteristics), and then checked seven months later (after the election season) to determine which users got suspended by Twitter (for methodological details, see Supplementary Information; data and code available at https://osf.io/mrbsw). These data allow us to make several contributions to policy discussions around political bias and anti-misinformation efforts.

First, accusations of political bias are based largely on anecdotes or salient unique cases, such as the suspension of former President Donald Trump. Thus, it remains unclear whether conservatives are, in fact, generally more likely to get suspended than liberals. Our data do support this claim: Accounts that had shared the #Trump2020 hashtag during the election were 4.4 times more likely to have been subsequently suspended than those that shared the #VoteBidenHarris2020 hashtag ($\chi 2(1)=486.9$, $p<0.0001$). Specifically, while only 4.5% of the Democratic users had been suspended as of July 2021, 19.6% of the Republican users had been suspended.

Critically, however, this association does *not* necessarily indicate a causal effect of a user's politics on suspension – because of the potential for political orientation to be confounded with the tendency to share misinformation (or to engage in other sanctioned behaviors). Thus, we also examined how the political orientation of the users in our study related to their sharing of links to low quality news sites in October 2020.

As discussed above, prior studies examining the link between political orientation and misinformation sharing have relied exclusively on professional fact-checkers or journalists to determine news quality, which makes this work susceptible to criticisms of liberal bias in the evaluations. We address this issue by complementing four such sets of news site evaluations generated by experts (*14*) with evaluations of 60 news sites (20 mainstream, 20 hyper-partisan, 20 fake news) generated using N=970 demographically representative (quota-sampled) American laypeople (*14*). We gave the ratings of Democrats and Republicans equal weight when constructing our laypeople ratings, and as a result these laypeople ratings cannot reasonably be accused of having liberal bias. See SI for details on the expert and crowd ratings used in our study.

Critically, we find that Republican Twitter users in our dataset shared news from domains that were on average rated as much less trustworthy than Democratic users, based on not only on the expert ratings ($0.71 < r < .78$, $p < .001$ for all), but also based on the politically-balanced layperson ratings ($r(8943) = .73$, $p < .001$); see Figure 1. We also find high correlations between the various measures of low-quality news sharing and measures of ideology based on the Twitter accounts the users follow (*11*) or the news sites that the users share (*10, 15*) (expert ratings, $0.74 < r < .88$, $p < .001$ for all; layperson ratings, $0.74 < r < .82$, $p < .001$ for all). Thus, among the politically active Twitter users in our study, Republicans and conservatives shared information from much lower quality sites than Democrats and liberals – even when quality was judged by a politically-balanced group of U.S. laypeople. This observation provides clear evidence for a political asymmetry in misinformation sharing in our dataset that cannot be attributed to liberal bias in what is considered misinformation or low quality news.



Republicans
(#Trump2020)

News site	***# Shares***
breitbart.com	45437
nypost.com	38624
hannity.com	17699
babylonbee.com	15490
thegatewaypundit.com	15401

Democrats
(#VoteBidenHarris2020)

News site	***# Shares***
nytimes.com	33396
cnn.com	28385
nbcnews.com	14351
theguardian.com	10923
thedailybeast.com	9689

Figure 1. Republican users shared links to much lower quality news sites than Democratic users in October 2020. *Distribution of low quality news sharing scores across Democrats (users who tweeted #VoteBidenHarris2020) and Republicans (users who tweeted #Trump2020) in our sample of 9000 Twitter users, based on links shared as of October 2020. (A) News site quality ratings given by eight professional fact-checkers. (B) News site quality ratings given by American laypeople recruited from Lucid, quota-matched to the national distribution on age, gender, education, and geographic region; ratings of Democratic respondents and Republican respondents were averaged to create politically-balanced layperson ratings. For details of the ratings, see SI and Ref (14). (C) Top 5 most-shared news sites among the Republican and Democrat users in our sample, using list of news sites from Ref (13).*

To what extent, then, can the apparent preferential suspension of right-leaning users actually be explained by differences in sharing of low quality news? To gain a first piece of insight into this question, we calculate the area-under-the-curve (AUC) metric for each measure (AUC captures accuracy while accounting for differences in base rates and is a standard metric of model performance in fields such as machine learning). We find that the various measures of sharing low quality news predict suspension ($0.68 < \text{AUC} < 0.72$) to a similar degree as the various partisanship and ideology measures ($0.67 < \text{AUC} < 0.71$); Figure 2A. Thus, when examined independently, suspension is not predicted any better by political orientation than by sharing low quality news.

Even more importantly, what happens when political orientation and sharing low quality news are used simultaneously to predict suspension? To answer this question, we construct an aggregate measure of political orientation by taking the first component of a principal component analysis (PCA) of our four ideology/partisanship measures, and an aggregate measure of sharing low quality news by taking the first component of a PCA of our four expert news site quality measures. We then use probit regression to predict whether the user was suspended as of the end of July 2021, with independent variables *z*-scored to facilitate the comparison of effect sizes.

Reproducing the AUC results above, we see a strong positive relationship between being more Republican/conservative and likelihood of being suspended ($b = 0.45$, $z = 22.6$, $p < 0.001$) when using political orientation as the sole independent variable in the probit regression. However, once low quality news sharing is added to the model, the association between suspension and political orientation is reduced by 56.2% ($b = 0.20$, $z = 4.6$, $p < 0.001$; see Figure 2b), and sharing low quality news is also strongly associated with suspension ($b = 0.27$, $z = 6.6$, $p < 0.001$). Further adding likelihood of being a bot (calculated via (*16*)) to the model reduces the association between suspension and political orientation even further ($b = 0.12$, $z = 2.52$, $p = 0.012$), such that the coefficient is now 74.1% smaller than in the model with only political orientation; sharing low quality news continues to be strongly associated with suspension ($b = 0.25$, $z = 5.9$, $p < 0.001$), as is likelihood of being a bot ($b = 0.13$, $z = 4.7$, $p < 0.001$). The results are largely unchanged when including a variety of additional metrics calculated based on each user's profile and tweets retrieved in October 2020; see Figure 2c and SI for details. In the full model, the association between suspension and political orientation ($b = 0.11$, $z = 2.1$, $p = 0.035$) is reduced by 75.3% relative to the political orientation-only model; and the p-value drops to a level that would not be considered statistically significant once accounting for multiple testing. Results are qualitatively similar when using ridge (penalized) regression and logistic regression; see SI.

These findings are particularly striking given that we are *prospectively* predicting subsequent suspension based on pre-election tweets. Simply accounting for the tendency the share low quality news sources shared *prior* to the election dramatically reduces the association between political orientation and suspension. It seems quite plausible that the comparatively small association between suspension and political orientation observed in the full model would be entirely eliminated if we had more precise (e.g., post-level, or real-time) measures of misinformation sharing, or were to include additional features (e.g., harmful content that was deleted prior to API retrieval).

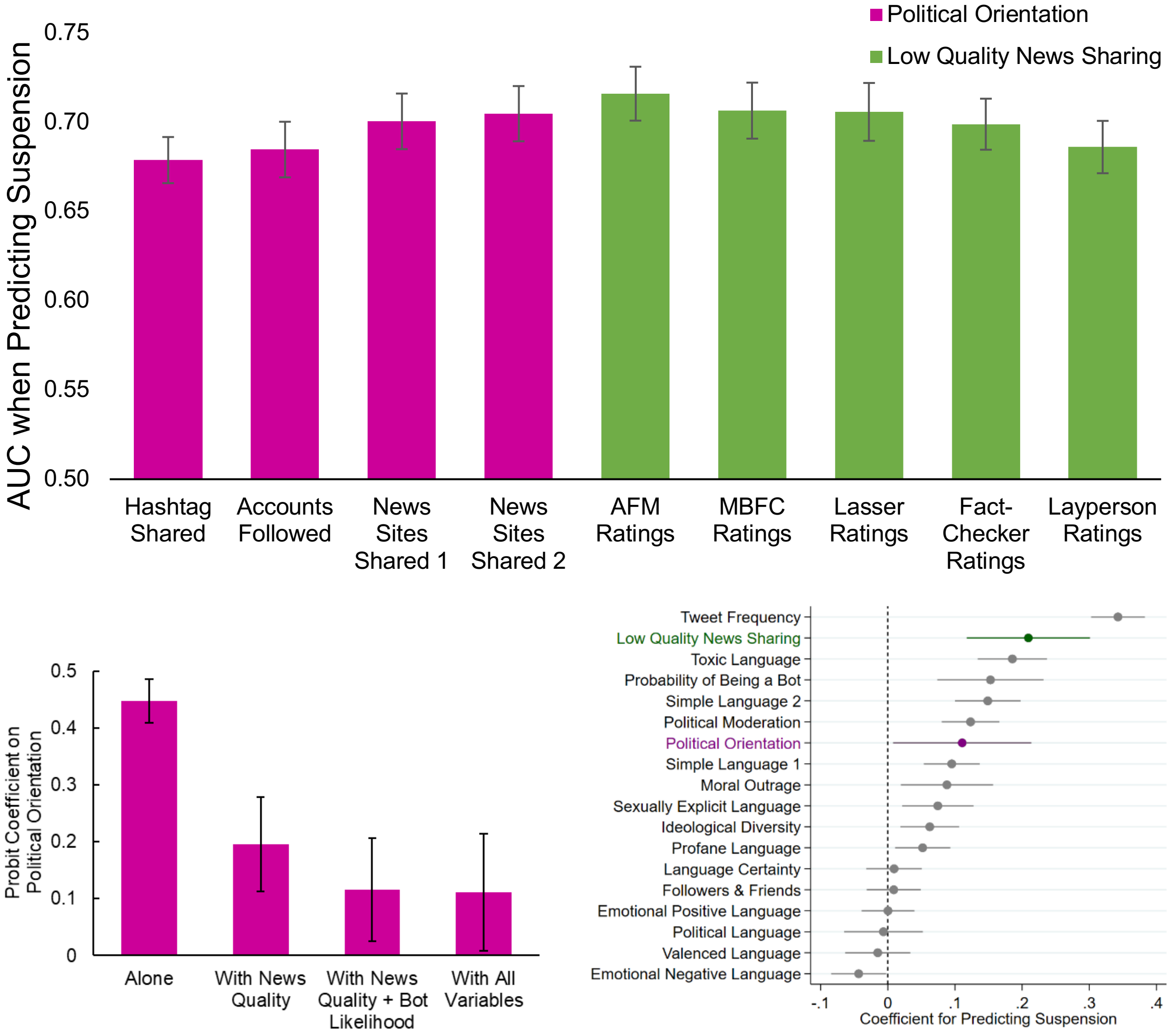


Figure 2. Political orientation is not a key predictor of getting suspended. *(A) When considered separately, political orientation and sharing low quality information are similarly predictive of suspension. Shown is area-under-the-curve (AUC, a standard measure of predictive accuracy) when predicting whether a user was suspended using models that take different features as the independent variable. Green bars indicate AUC for measures of political orientation (partisanship based on sharing of #VoteBidenHarris2020 versus #Trump2020 hashtags; ideology based on accounts followed, estimated using the model of (17); ideology based on news sites shared, estimated using the model of (15) or the model of (10). Purple bars indicate AUC for measures of sharing links to low quality news sites based on ratings from Ad Fontes Media (AFM) (18), Media Bias/Fact Check (MBFC) (19), (13), eight professional fact-checkers (14), and 970 American laypeople weighting Democrats and Republicans equally (14). Error bars indicate bootstrapped 95% confidence intervals. (B) When considered together with other relevant variables, political orientation becomes 75.3% less predictive of suspension than when considered on its own. Shown is the coefficient on an aggregate of the four political orientation variables in a probit model predicting suspension using political orientation alone, controlling for an aggregate of the four expert news site quality ratings of the links shared by the user as of October 2020, controlling for quality of news shared and likelihood of being a bot, and controlling for all available variables. Error bars indicate 95% confidence intervals. (C) Shown are coefficients from the full probit model predicting suspension using all available variables (z-scored to make coefficients comparable). For detailed definitions of the measures in panel B, see SI. Error bars indicate 95% confidence intervals.*

In sum, these results do not support accusations of pervasive anti-conservative bias on the part of Twitter. That is, we do not find evidence of a strong causal effect of political orientation on suspension. Instead, much of the apparent preferential suspension of right-leaning users is better explained by other aspects of their behavior, most notably sharing information from low quality news sites. Beyond the sharing of misinformation or conspiracy theories often promoted by low quality news sites, conservatives may also have been preferentially suspended for engaging in other sanctioned behaviors, such using bots (as suggested by our data) or engaging in calls for violence (e.g., in connection with the insurrection at the US capital on January 6th 2021, which occurred during our study period). Regardless of which prohibited behavior(s) are in operation, the same fundamental point applies – partisan asymmetries in behavior can lead to partisan asymmetries in suspension, even when suspension policies are politically neutral.

The suspension of conservatives in our data thus appears to be largely the result of (unintentional) disparate impact rather than (intentional) disparate treatment. From a legal perspective, political orientation is not a protected class in the United States, and thus neither form of disparate treatment is illegal (although potentially still normatively undesirable). While disparate impact may reasonably be considered to constitute discrimination in some cases (e.g., employment discrimination based on job-irrelevant factors that correlate with race), in the present context reducing the spread of misinformation is a legitimate and necessary goal for social media platforms. This makes a normative case for disparate impact based on political orientation.

Although we focused on Twitter suspensions and U.S. politics during the 2020 election cycle as a case study, the lessons learned from our analysis are relevant whenever there is an association between political orientation and misinformation sharing (or other sanctioned behaviors) – regardless of the specific platform or form of enforcement. Importantly, such an association is not limited to the U.S. (*12, 13*). Thus, there is reason to expect that a variety of politically-neutral actions taken against misinformation on platforms in many different countries will lead to preferential sanctioning of conservatives. This is particularly relevant given the Digital Services Act recently passed by the European Union, which requires platforms to take down content that involves misinformation. Our results suggest that when they comply, platforms are almost certain to face accusations of anti-conservative bias - and that policy-makers in the EU must understand that such patterns will likely arise from the actions they are requiring platforms to take, even if platforms comply in a political-neutral manner. Our results also suggest that using politically-balanced crowds to evaluate content (*20*) may be a way to identify misinformation while ameliorating charges of political bias.

More broadly, the data presented here drive home the point that – at least in the current global political climate and media ecosystem – platforms face a fundamental tradeoff between reducing the spread of misinformation and being politically balanced in their enforcement. In so far as conservatives share more misinformation, it is not possible to be maximally effective in combatting misinformation without preferentially taking action against conservatives. Given the widespread (and bi-partisan) public demand for reducing misinformation online, policy makers must accept that some level of differential enforcement across party lines is necessary if technology companies are to keep misinformation in check. The goal should be neutral policy design, not neutral enforcement.

References

1. D. M. Lazer *et al.*, The science of fake news. *Science* **359**, 1094-1096 (2018).
2. C. Koopman CGO Tech Poll. *The Center for Growth and Oppurtunity at Utah State University*, (2021).
3. Flash Eurobarometer 464: Fake News and Disinformation Online v1.00. *European Commission Directorate-General for Communication*, (2018).
4. A guide to anti-misinformation actions around the world. https://www.poynter.org/ifcn/anti-misinformation-actions/.
5. N. Persily, J. A. Tucker, *Social Media and Democracy: The State of the Field, Prospects for Reform*. (Cambridge University Press, 2020).
6. S. Bond. (NPR, 2020).
7. G. Pennycook, D. G. Rand, The psychology of fake news. *Trends in cognitive sciences*, (2021).
8. J. Bowden, Twitter CEO Jack Dorsey: I 'fully admit' our bias is 'more left-leaning.'. *Hill*, (2018).
9. A. Guess, J. Nagler, J. Tucker, Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science advances* **5**, eaau4586 (2019).
10. N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, Fake news on Twitter during the 2016 US presidential election. *Science* **363**, 374-378 (2019).
11. G. Pennycook, D. G. Rand, Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature communications* **13**, 1-12 (2022).
12. A. A. Arechar *et al.*, Understanding and reducing online misinformation across 16 countries on six continents. (2022).
13. J. Lasser *et al.*, Social media sharing of low-quality news sources by political elites. *PNAS nexus* **1**, pgac186 (2022).
14. G. Pennycook, D. G. Rand, Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* **116**, 2521-2526 (2019).
15. G. Eady, J. Nagler, R. Bonneau, J. Tucker, Political information sharing and ideological polarization. *Midwest Political Science Association, Chicago*, (2019).
16. Social Media Account Classifier. *Botsentinel*.
17. P. Barberá, J. T. Jost, J. Nagler, J. A. Tucker, R. Bonneau, Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science* **26**, 1531-1542 (2015).
18. Ad Fontes Media. http://adfontesmedia.com.
19. Media Bias/Fact Check. http://mediabiasfactcheck.com/.
20. J. Allen, A. A. Arechar, G. Pennycook, D. G. Rand, Scaling up fact-checking using the wisdom of crowds. *Science advances* **7**, eabf4393 (2021).

Supplementary Information

Supplemental Methods

First, we collected a list of Twitter users who tweeted or retweeted either of the election hashtags #Trump2020 and #VoteBidenHarris2020 on October 6, 2020. We also collected the most recent 3,200 tweets sent by each of those accounts. We processed tweets and extracted tweeted domains from 34,920 randomly selected users (15,714 shared #Trump2020 and 19,206 shared #VoteBidenHarris2020), and filtered down to 12,238 users who shared at least 5 links to domains used by the ideology estimator of (*1*). We also excluded 426 'elite' users with more than 15,000 followers who are likely unrepresentative of Twitter users more generally. We then constructed a politically balanced set by randomly selecting 4,500 users each from the remaining 4,756 users who shared #Trump2020 and 7056 users who shared #VoteBidenHarris2020. After nine months, on July 30, 2021, we checked the status of the 9000 users and assessed suspension. We classify an account as having been suspended if the Twitter API returned error code 63 ("User has been suspended") when querying that user.

To measure a user's tendency to share misinformation, we follow most other researchers in this space (*2-5*) and use source quality as a proxy for article accuracy, because it is not feasible to rate the accuracy of individual tweets. Specifically, to quantify the quality of news shared by each user, we leveraged a previously published set of 60 news sites (20 mainstream, 20 hyper-partisan, 20 fake news; see Table S1) whose trustworthiness had been rated by eight professional fact-checkers as well as politically-balanced crowds of laypeople. We also examined Reliability ratings for a set of 283 sites from Ad Fontes Media, Inc. (*6*), Factual Reporting ratings for a set of 3216 sites from Media Bias/Fact Check (*7*), and Accuracy ratings for a set of 4767 sites from a recent academic paper by Lasser and colleagues (*8*). We then used the Twitter API to retrieve the last 3,200 posts (as of October 6, 2020) for each user in our study, and collected all links to any of those sites shared (tweeted or retweeted) by each user. Following the approach used in prior work (*4, 5*), we calculated a news quality score for each user (bounded between 0 and 1) by averaging the ratings of all sites whose links they shared, separately for each set of site ratings. Finally, we transform these ratings into *low* quality news sharing scores by subtracting the news quality ratings from 1. Over 99% of users in our study had shared at least one link to a rated domain; for each set of domain ratings (as well as all other independent variables), we replaced missing values with the sample mean. When combining the four expert-based measures into an aggregate news quality score, PCA indicated that only one component should be retained (87% of variation explained), which had weights of 0.50 on Pennycook & Rand 2019 fact-checker ratings, 0.51 on Ad Fontes Media Reliability ratings, 0.48 on Media Bias/Fact Check Factual Reporting ratings, and 0.51 on Lasser et al. 2022 Accuracy ratings. In all analyses, we use parallel analysis to determine the number of retained components.

Mainstream			Hyper-partisan			Fake news		
Domain	**Politically balanced layperson rating**	**Fact-checker rating**	**Domain**	**Politically balanced layperson rating**	**Fact-checker rating**	**Domain**	**Politically balanced layperson rating**	**Fact-checker rating**
abcnews.go.com	0.45	0.56	activepost.com	0.2	0	americannews.com	0.22	0
aol.com/news	0.35	0.41	antiwar.com	0.18	0	angrypatriotmovement.com	0.18	0
bbc.co.uk	0.38	0.81	blacklistednews.com	0.18	0	bb4sp.com	0.18	0
bostonglobe.com	0.33	0.75	breitbart.com	0.22	0.16	beforeitsnews.com	0.19	0
cbsnews.com	0.48	0.66	commondreams.org	0.18	0.03	channel24news.com	0.25	0.06
chicagotribune.com	0.38	0.53	conservativetribune.com	0.24	0.03	clashdaily.com	0.18	0
cnn.com	0.47	0.84	crooksandliars.com	0.18	0.13	conservativedailypost.com	0.23	0
dailymail.co.uk	0.3	0.44	dailycaller.com	0.21	0.13	dailybuzzlive.com	0.24	0
foxnews.com	0.45	0.44	dailykos.com	0.2	0.16	downtrend.com	0.19	0
huffingtonpost.com	0.41	0.47	dailysignal.com	0.2	0	freedomdaily.com	0.2	0.03
latimes.com	0.33	0.75	dailywire.com	0.25	0.16	newsbreakshere.com	0.19	0
msnbc.com	0.44	0.66	ijr.com	0.19	0.09	notallowedto.com	0.17	0
news.yahoo.com	0.4	0.59	infowars.com	0.21	0.03	now8news.com	0.2	0
nydailynews.com	0.33	0.34	newsmax.com	0.23	0.13	onepoliticalplaza.com	0.19	0
nypost.com	0.38	0.38	patriotpost.us	0.21	0	react365.com	0.17	0
nytimes.com	0.45	0.91	rawstory.com	0.19	0.09	realnewsrightnow.com	0.21	0
sfchronicle.com	0.26	0.59	redstate.com	0.2	0.06	socialeverythings.com	0.18	0
usatoday.com	0.45	0.66	thedailysheeple.com	0.18	0.09	thenewyorkevening.com	0.24	0
washingtonpost.com	0.45	0.91	thepoliticalinsider.com	0.22	0.03	whatdoesitmean.com	0.19	0
wsj.com	0.34	0.72	westernjournal.com	0.22	0.06	yournewswire.com	0.19	0.06

Table S1. Set of 60 news site quality scores generated by 8 professional fact-checker trustworthiness ratings and politically-balanced trustworthiness ratings from 970 laypeople; see (9) for details. These scores are subtracted from 1 to generate the low quality news site sharing scores shown in main text Figure 1.

To measure a user's political orientation, we first classify their partisanship based on whether they shared more #Trump2020 or #VoteBidenHarris2020 hashtags. Additionally, we retrieved all accounts followed by users in our sample and used the statistical model from (*10*) to obtain a continuous measure of users' ideology based on the ideological leaning of the accounts they followed. Similarly, we used the statistical models from (*11*) and (*3*) to estimate users' ideology using the ideological leanings of the news sites that the users shared content from. We also calculated user ideology by averaging political leanings of domains they shared through tweets or retweets based on the method in (*3*). The intuition behind these approaches is that users on social media are more likely to follow accounts (and share news stories from sources) are align with their own ideology than those that are politically distant. Thus, ideology of accounts the user followed (and share news stories shared by the user) provides insight into the user's ideology. When combining these four measures into an aggregate political orientation score, PCA indicated that only one component should be retained (88% of variation explained), which had weights of 0.49 on hashtag-based partisanship, 0.49 on follower-based ideology, 0.51 on sharing-based ideology via (*11*), and 0.51 on sharing-based ideology via (*3*). We also used this aggregate measure to calculate a user's extent of ideological moderation versus extremity by taking the absolute value of the aggregate ideology measure, and used the standard deviation of news site ideology scores from (*3*) across a user's tweets as a measure of the ideological diversity of news shared by the user.

Furthermore, we used each user's most recent 200 tweets (rather than the 3200 available tweets, for tractability) as of October 6, 2020 to calculate the average of the following metrics for language use; all language metrics are winsorized at the 99^{th} percentile. We examined harmful language used in the tweets using Google Jigsaw Perspective API (*12*) (including "toxicity", "severe toxicity", "identity attack", "insult", "profanity", and "threat") and using Rewire API (*13*) (including "abuse", "hate", "profanity", "violent", and "sexually explicit"). PCA of these items indicated that three components should be retained, whose loadings are shown in Table S2 and which describe general toxicity, sexually explicit language, and profane language. We examined language simplicity versus complexity using a variety of metrics including, Kincaid, ARI, Coleman-Liau, Flesch Reading Ease, Gunning Fog Index, LIX, SMOG Index, RIX, and Dale Chall Index (*14*). PCA of these items indicated that two components should be retrained, whose loadings are shown in Table S3 and which describe general language simplicity and language which is simple interactions of number of syllables but not number of characters/words. We used VADER to measure the extent to which the tweets had positive valence, negative valence, and neutral valence (*15*). PCA which indicated that these measures all loaded on a single component (70.8% of variation explained) with weightings of 0.56 on positive valence, 0.46 on negative valence, and -0.69 on neutral valence. We used the approach from (*16*) to quantify certainty in language used in the tweets. We used the approach from (*17*) to quantify the level of emotionality in the language separately for positive and negative words. We used the approach from (*18*) to quantify the expression of moral outrage in the tweets.

We used the API from (*19*) to estimate the likelihood of each user being a social bot based on the text of the user's tweets. Finally, we also collected each user's number of followers and number of friends, and calculated each user's friend/follower ratio, which were highly correlated with each other; we log-transformed each of these variables and performed PCA, which indicated that these measures all loaded on a single component (70.4% of variation explained) with weightings of 0.69 on log(friends+1), 0.57 on log(followers+1), and -0.45 on log((friends+1)/(followers+1)).

In all regression models, all independent variables are z-scored to be coefficients comparable.

	Toxic	Sexually Explicit	Profane
Jigsaw: Toxicity	0.39	-0.14	0.07
Jigsaw: Insult	0.37	-0.14	0.03
Jigsaw: Profanity	0.36	0.20	0.16
Jigsaw: Severe Toxicity	0.34	0.31	-0.24
Jigsaw: Identity Attack	0.31	0.07	-0.46
Jigsaw: Threat	0.28	0.20	-0.51
Rewire: Abuse	0.33	-0.36	0.29
Rewire: Hate	0.24	-0.58	0.01
Rewire: Profanity	0.25	0.21	0.44
Rewire: Violent	0.23	-0.07	0.04
Rewire: Sexually Explicit	0.13	0.52	0.40
Variance explained	*56.1%*	*11.9%*	*10.8%*

Table S2. Loadings on the 3 retained components for toxicity measures.

	Simplicity 1	Simplicity 2
Kincaid	-0.40	-0.12
ARI	-0.38	0.26
Coleman-Liau	-0.28	0.51
Flesch Reading Ease	0.37	-0.04
Gunning Fog Index	-0.33	-0.41
LIX	-0.38	0.26
SMOG Index	-0.31	-0.42
RIX	-0.36	-0.17
Dale Chall Index	-0.05	0.47
Variance explained	*59.9%*	*21.6%*

Table S3. Loadings on the 2 retained components for language simplicity measures.

Supplemental Results

In the main text, Figure 1 shows the distribution of low quality news sharing by party using ratings of fact-checkers and laypeople from Pennycook & Rand 2019. Here, in Figure S1 we show the same distributions using the other three sets of expert news site quality ratings. Figure S2 shows the distribution of Bot Sentinel scores by party.

Turning to our models predicting user suspension, Table S4 shows the full regression models for the probit models reported in the main text. Table S4 also shows the coefficients for the full models estimated using ridge regression with the penalty parameter selected using 5-fold cross validation; the penalty that maximizes out of sample prediction is very small, such that the coefficients in the penalized regression are quite similar to the standard probit model. Finally, Table S5 shows the equivalent models using logistic regression instead of probit regression.

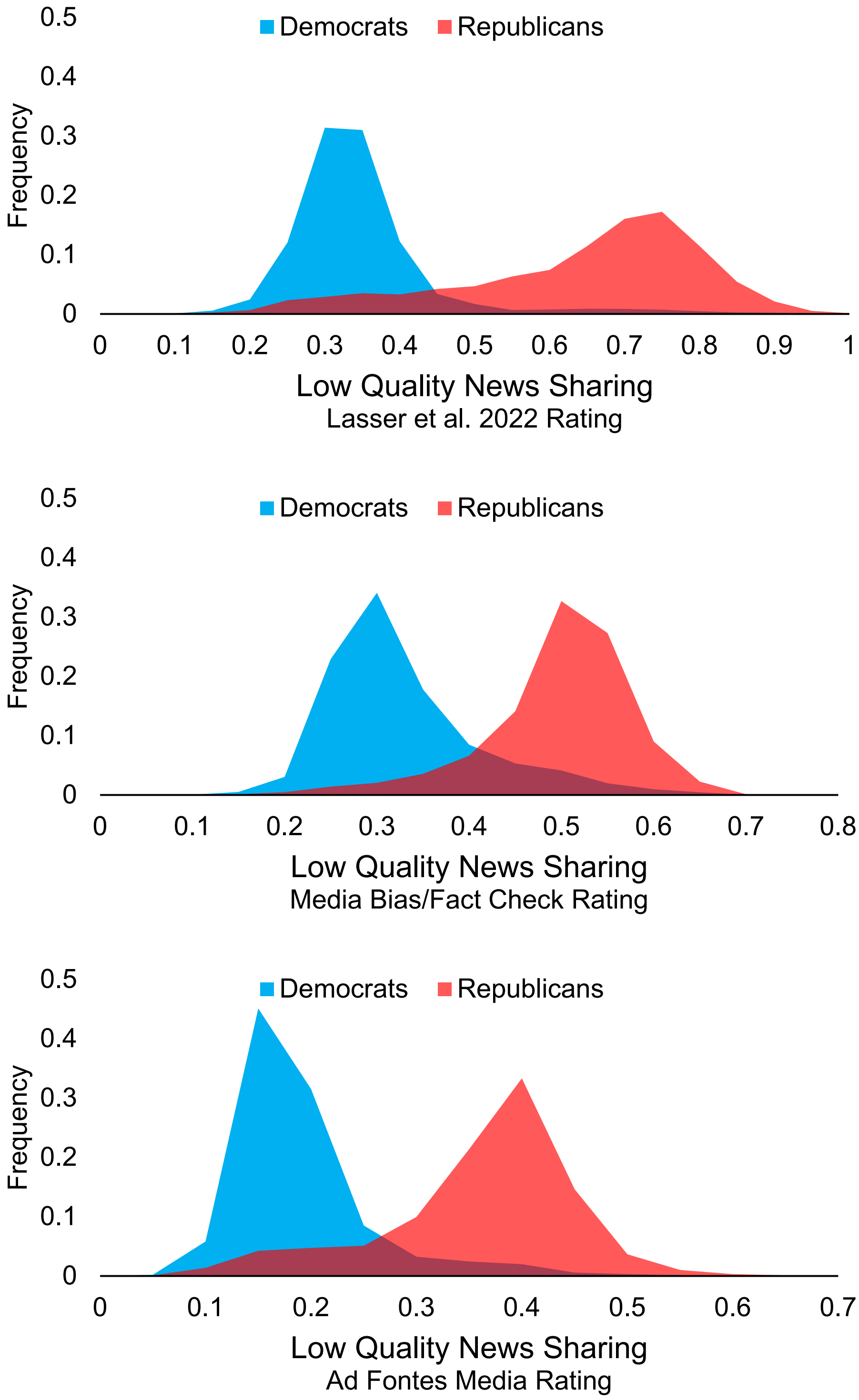


Figure S1. *Low Quality News Site Sharing scores by partisanship using alternative quality rating sets.*

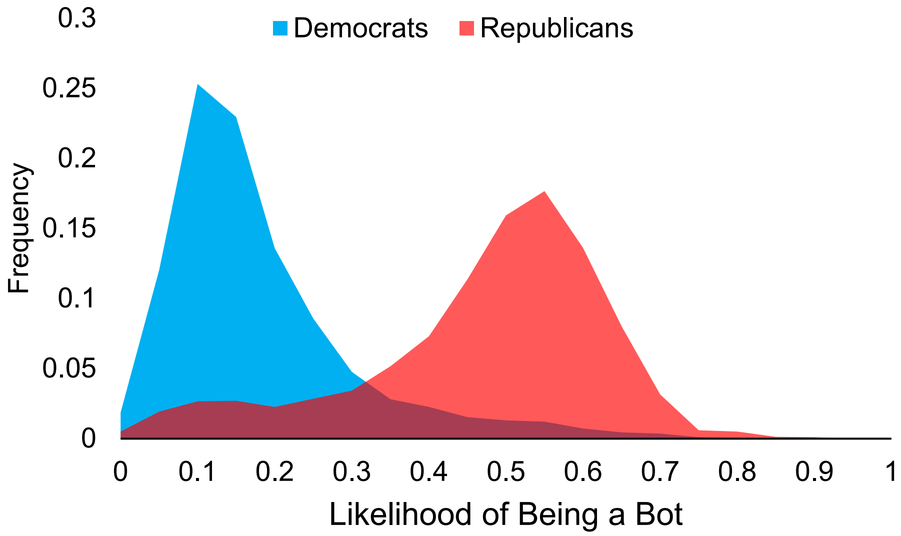


Figure S2. *Bot sentinel scores by partisanship.*

Table S4. Probit regression models. Model 5 shows coefficients from ridge regression.

	(1)	(2)	(3)	(4)	(5)
Political Orientation	0.447***	0.196***	0.116*	0.111*	0.123
	(0.0198)	(0.0426)	(0.0460)	(0.0525)	
Low Quality News Sharing		0.273***	0.247***	0.209***	0.189
		(0.0415)	(0.0421)	(0.0469)	
Bot Score (Botsentinel)			0.135***	0.153***	0.148
			(0.0287)	(0.0403)	
Followers & Friends				0.00880	0.012
				(0.0205)	
Toxic Language				0.185***	0.178
				(0.0263)	
Sexually Explicit Language				0.0744**	0.071
				(0.0271)	
Profane Language				0.0518*	0.050
				(0.0210)	
# Tweets in Past 2 Weeks				0.343***	0.324
				(0.0204)	
Political Moderateness				0.123***	0.110
				(0.0219)	
Simple Language PC1				0.0952***	0.090
				(0.0213)	
Simple Language PC2				0.149***	0.141
				(0.0250)	
Valenced Language				-0.0150	-0.016
				(0.0248)	
Certainty in Language				0.00927	0.011
				(0.0211)	
Emotional Negative Language				-0.0434*	-0.041
				(0.0210)	
Emotional Positive Language				0.000225	0.0009
				(0.0201)	
Moral Outrage				0.0879*	0.084
				(0.0350)	
Political Language				-0.00657	-0.009
				(0.0299)	
Ideology Diversity				0.0624**	0.062
				(0.0223)	
Constant	-1.291***	-1.296***	-1.302***	-1.433***	-1.419
	(0.0198)	(0.0199)	(0.0201)	(0.0235)	
Observations	9,000	9,000	9,000	9,000	9,000

Standard errors in parentheses
*** p<0.001, ** p<0.01, * p<0.05

Table S5. Logistic regression models. Model 5 shows coefficients from ridge regression.

Logistic regression	(1)	(2)	(3)	(4)	(5)
Political Orientation	0.866***	0.382***	0.243**	0.226*	0.244
	(0.0403)	(0.0828)	(0.0895)	(0.0998)	
Low Quality News Sharing		0.518***	0.477***	0.403***	0.365
		(0.0795)	(0.0805)	(0.0873)	
Bot Score (Botsentinel)			0.227***	0.302***	0.291
			(0.0535)	(0.0741)	
Followers & Friends				0.0145	0.0201
				(0.0383)	
Toxic Language				0.322***	0.311
				(0.0483)	
Sexually Explicit Language				0.133**	0.127
				(0.0491)	
Profane Language				0.0863*	0.085
				(0.0379)	
# Tweets in Past 2 Weeks				0.631***	0.598
				(0.0369)	
Political Moderateness				0.250***	0.226
				(0.0409)	
Simple Language PC1				0.183***	0.174
				(0.0394)	
Simple Language PC2				0.267***	0.255
				(0.0453)	
Valenced Language				-0.0276	-0.029
				(0.0453)	
Certainty in Language				0.0123	0.017
				(0.0390)	
Emotional Negative Language				-0.0821*	-0.078
				(0.0388)	
Emotional Positive Language				0.00502	0.006
				(0.0368)	
Moral Outrage				0.173**	0.163
				(0.0642)	
Political Language				-0.0291	-0.031
				(0.0555)	
Ideology Diversity				0.110**	-0.110
				(0.0412)	
Constant	-2.262***	-2.273***	-2.282***	-2.557***	-2.52
	(0.0412)	(0.0414)	(0.0417)	(0.0495)	
Observations	9,000	9,000	9,000	9,000	

Standard errors in parentheses
*** p<0.001, ** p<0.01, * p<0.05

Supplementary References

1. G. Eady, J. Nagler, A. Guess, J. Zilinsky, J. A. Tucker, How many people live in political bubbles on social media? Evidence from linked survey and Twitter data. *Sage Open* **9**, 2158244019832705 (2019).
2. A. Guess, J. Nagler, J. Tucker, Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science advances* **5**, eaau4586 (2019).
3. N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, Fake news on Twitter during the 2016 US presidential election. *Science* **363**, 374-378 (2019).
4. G. Pennycook *et al.*, Shifting attention to accuracy can reduce misinformation online. *Nature* **592**, 590-595 (2021).
5. M. Mosleh, C. Martel, D. Eckles, D. G. Rand, in *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. (2021), pp. 1-13.
6. Ad Fontes Media. *http://adfontesmedia.com*.
7. Media Bias/Fact Check. *http://mediabiasfactcheck.com/*.
8. J. Lasser *et al.*, Social media sharing of low-quality news sources by political elites. *PNAS nexus* **1**, pgac186 (2022).
9. G. Pennycook, D. G. Rand, Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* **116**, 2521-2526 (2019).
10. P. Barberá, J. T. Jost, J. Nagler, J. A. Tucker, R. Bonneau, Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science* **26**, 1531-1542 (2015).
11. G. Eady, J. Nagler, R. Bonneau, J. Tucker, Political information sharing and ideological polarization. *Midwest Political Science Association, Chicago*, (2019).
12. Using machine learning to reduce toxicity online. *Google Prospective API*.
13. Socially responsible AI for online safety. *Rewire Online*.
14. An implementation of traditional readability measures based on simple surface characteristics. *Python readability library*.
15. C. Hutto, E. Gilbert, in *Proceedings of the international AAAI conference on web and social media*. (2014), vol. 8, pp. 216-225.
16. M. D. Rocklage, S. He, D. D. Rucker, L. F. Nordgren, EXPRESS: Beyond Sentiment: The Value and Measurement of Consumer Certainty in Language. *Journal of Marketing Research*, 00222437221134802 (2022).
17. M. D. Rocklage, D. D. Rucker, L. F. Nordgren, The Evaluative Lexicon 2.0: The measurement of emotionality, extremity, and valence in language. *Behavior research methods* **50**, 1327-1344 (2018).
18. W. J. Brady, K. McLoughlin, T. N. Doan, M. J. Crockett, How social learning amplifies moral outrage expression in online social networks. *Science Advances* **7**, eabe5641 (2021).
19. Social Media Account Classifier. *Botsentinel*.