Article

Post-January 6th deplatforming reduced the reach of misinformation on Twitter

https://doi.org/10.1038/s41586-024-07524-8

Received: 27 October 2023

Accepted: 6 May 2024

Published online: 5 June 2024

Check for updates

Stefan D. McCabe[1,7], Diogo Ferrari[2,7], Jon Green[3], David M. J. Lazer[4,5]✉ & Kevin M. Esterling[2,6]

The social media platforms of the twenty-first century have an enormous role in regulating speech in the USA and worldwide[1]. However, there has been little research on platform-wide interventions on speech[2,3]. Here we evaluate the effect of the decision by Twitter to suddenly deplatform 70,000 misinformation traffickers in response to the violence at the US Capitol on 6 January 2021 (a series of events commonly known as and referred to here as 'January 6th'). Using a panel of more than 500,000 active Twitter users[4,5] and natural experimental designs[6,7], we evaluate the effects of this intervention on the circulation of misinformation on Twitter. We show that the intervention reduced circulation of misinformation by the deplatformed users as well as by those who followed the deplatformed users, though we cannot identify the magnitude of the causal estimates owing to the co-occurrence of the deplatforming intervention with the events surrounding January 6th. We also find that many of the misinformation traffickers who were not deplatformed left Twitter following the intervention. The results inform the historical record surrounding the insurrection, a momentous event in US history, and indicate the capacity of social media platforms to control the circulation of misinformation, and more generally to regulate public discourse.

Social media platforms such as X (previously known as Twitter), Facebook, YouTube and Instagram are key mediators of civic and political speech in the modern world. Tech companies' policies regarding permissible speech, including what they allow to be posted, what they allow to be seen and who is allowed to post, loom large as the de facto laws governing political communication through much of the world[1]. Each platform actively determines the content that users see using two mechanisms: algorithmic sorting of user-posted content, and enforcing terms of use and community conduct policies that specify acceptable content and posting behaviour.

Given the importance of social media platforms in regulating our contemporary public sphere, surprisingly little scholarly attention has focused on the effect of the interventions these platforms undertake to curate content. Here we evaluate the effect of a terms-of-use intervention undertaken by Twitter following the events of 6 January 2021. That day witnessed a violent storming of the US Capitol by supporters of then-President Donald Trump in an effort to prevent the certification of the election of Joe Biden as President of the USA. On 8 to 12 January 2021, in response to the violence, Twitter deplatformed a large number of right-wing misinformation traffickers, including Trump. Twitter had a central role in the events of January 6th, and this intervention to regulate speech on the platform was widely reported.

Here we use natural experimental designs to evaluate the effects of this intervention on the amount of misinformation that circulated on Twitter, using a panel of more than 500,000 Twitter users[4,5,8] who were active in the election cycle between June 2020 and February 2021. We specifically evaluate the following questions: (1) Did Twitter's intervention have an effect on the volume of misinformation circulating on the platform? (2) Was this change due only to the deplatforming of a set of users who had heavily trafficked in misinformation but could no longer share it, or was there also a spillover effect of the deplatforming event on the sharing behaviour of those who remained, and in particular among those who followed the deplatformed accounts[9]? and (3) Among misinformation traffickers who were not deplatformed, to what extent did the intervention lead some to voluntarily exit the platform?

Our core analytical challenge is that Twitter's deplatforming intervention coincided with the insurrection and election certification, along with massive media reporting on each of these events, and thus there are inherent confounds in each of our comparisons. Since we use observational data, the effects that we identify can be interpreted as causal only under the assumptions that we state and, to the extent possible, warrant below.

The January 6th insurrection was a momentous event in US history, and understanding engagement on Twitter is of strong historical interest given the centrality of its role in those events on one of the darkest days for US democracy. More generally, the results are informative regarding the current role of social media companies in the regulation of speech, and how terms-of-use interventions may be applied in other settings.

[1]Institute for Data, Democracy & Politics, George Washington University, Washington, DC, USA. [2]Department of Political Science, University of California, Riverside, Riverside, CA, USA. [3]Department of Political Science, Duke University, Durham, NC, USA. [4]Network Science Institute, Northeastern University, Boston, MA, USA. [5]Institute for Quantitative Social Science, Harvard University, Cambridge, MA, USA. [6]School of Public Policy, University of California, Riverside, Riverside, CA, USA. [7]These authors contributed equally: Stefan D. McCabe, Diogo Ferrari. ✉e-mail: d.lazer@northeastern.edu

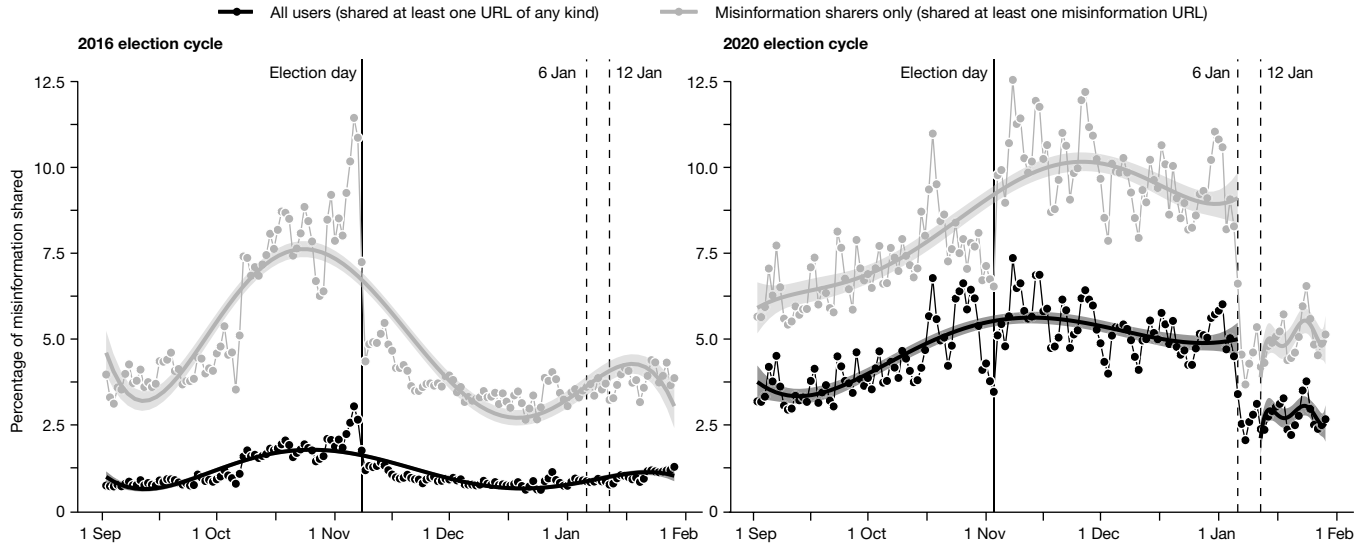


Fig. 1 | Misinformation sharing on Twitter during the 2016 and 2020 US election cycles. Daily percentage of posts shared (tweeted or retweeted) containing a misinformation URL among all tweets and retweets containing a URL for all users and for users classified as misinformation users during the 2016 (left) and 2020 (right) US election cycles. 2016 election cycle: 149 days (2 September 2016 to 29 January 2017); average daily total number of shared posts (tweeted or retweeted), 174,429; 79,393 daily active users on average. 2020 election cycle: 149 days (2 September 2020 to 29 January 2021); average daily total number of shared posts, 217,464; 63,411 active users on average. Fitted lines are 5-degree polynomials and shaded areas show bootstrap 95% confidence intervals. The graph for the 2020 election cycle uses two polynomials, fitted separately before 6 January 2021 and after 12 January 2021.

Regulating speech on big tech platforms

Social media companies filter content on their platforms for two reasons. First, with millions or billions of users contributing content, the limits of human attention require that users can only consume a tiny fraction of the content of the platform at any one time. Given these limits, the platforms have a business interest in developing algorithms to promote content that users find most engaging[10]. Second, users can choose to circulate material that is misinforming, harmful or even illegal, and the platforms tailor their terms of use to walk a fine line between suppressing objectionable content and permitting content that many users find attractive.

In the USA, Section 230 of the Communications Decency Act codifies tech platforms' right to moderate content[11], but does not affirmatively require content moderation in most cases[12], and indeed Section 230 shields the platforms from legal responsibility for the spread of misinformation[13]. In traditional media, content editors belong to a profession that is largely guided by journalistic ethics and judgment. By contrast, content moderation on social media is guided by terms of use and algorithmic design that is intended to protect the platforms' commercial interests rather than any larger public interest[14–16].

Individual platforms articulate content moderation guidelines and the penalties for violating them, in their terms of use and community standards policies, such as Twitter's Civic Integrity Policy[17,18] and Facebook's Community Standards policy[19,20]. However, content suppression results in difficult business decisions for tech companies. It is well-established that social media users post a considerable volume of misinformation, and that political content is more engaging[21] and extreme content perhaps especially so[8,22]. Because of these engagement patterns, curation policies that optimize engagement metrics may promote content that fosters polarization, divisiveness and extremism[23], and tech companies often find themselves under intense political pressure to limit the extent to which their platforms undermine democracy and the public interest[24].

The actions that platforms undertake to regulate speech are of increasing importance. Yet academic research on the causal effects of the platforms' attempts to limit misinformation and harmful content is relatively limited. A small number of papers do show compelling results. For example, Chandrasekharan et al.[25] used matching and difference-in-differences (DID) to identify the causal effects of banning Reddit threads that promote hate speech on reducing hate content. Matias[26] used a randomized design to evaluate the beneficial effects of messages regarding community norms. Jhaver et al.[2] examined the effect of Twitter's deplatforming of prominent malicious influencers on the spread of toxic content. Yildirim et al.[27] used suspensions for hate speech on Twitter to assess the deterrent effect of warning users to avoid terms-of-use violations. Several recent papers have reported on experiments with changes to Facebook's curation algorithm[10,28,29]. Broniatowski et al.[3] investigated the effect of the sudden removal of vaccine misinformation on Facebook; their paper is the most similar to this Article, and we shall revisit it in the conclusion.

The number of studies examining platform regulation, however, is not commensurate with the importance of this form of regulation for modern social discourse. One reason for this dearth of studies is the limited access that researchers have to data from social media companies[30]. Another reason is the design problem that platforms typically enforce terms of use in an incremental and piecemeal manner, stepping up consequences only over multiple violations, and in an arbitrary case-by-case approach that is not observable systematically or at scale[31]. Twitter's massive and sudden deplatforming following the January 6th insurrection creates a natural experiment to show the effect of terms-of-use enforcement. Similar deplatforming events include an event in June 2020 when Twitter removed 23,750 Chinese fake accounts that were spreading misinformation on COVID-19[32], and when Facebook banned accounts associated with the Myanmar military in light of human rights abuses[33].

January 6th and the crisis at Twitter

On 6 January 2021, Congress met to exercise its constitutional duty to certify the electoral victory of President-elect Joe Biden, and witnessed the storming of the Capitol by supporters of Donald Trump in a violent insurrection to prevent that certification. Twitter had a central role in that event[34]. It was the primary medium used by then-President Trump to call people to Washington. For example, on December 19 Trump tweeted "Big protest in DC on January 6th. Be there, will be wild!".

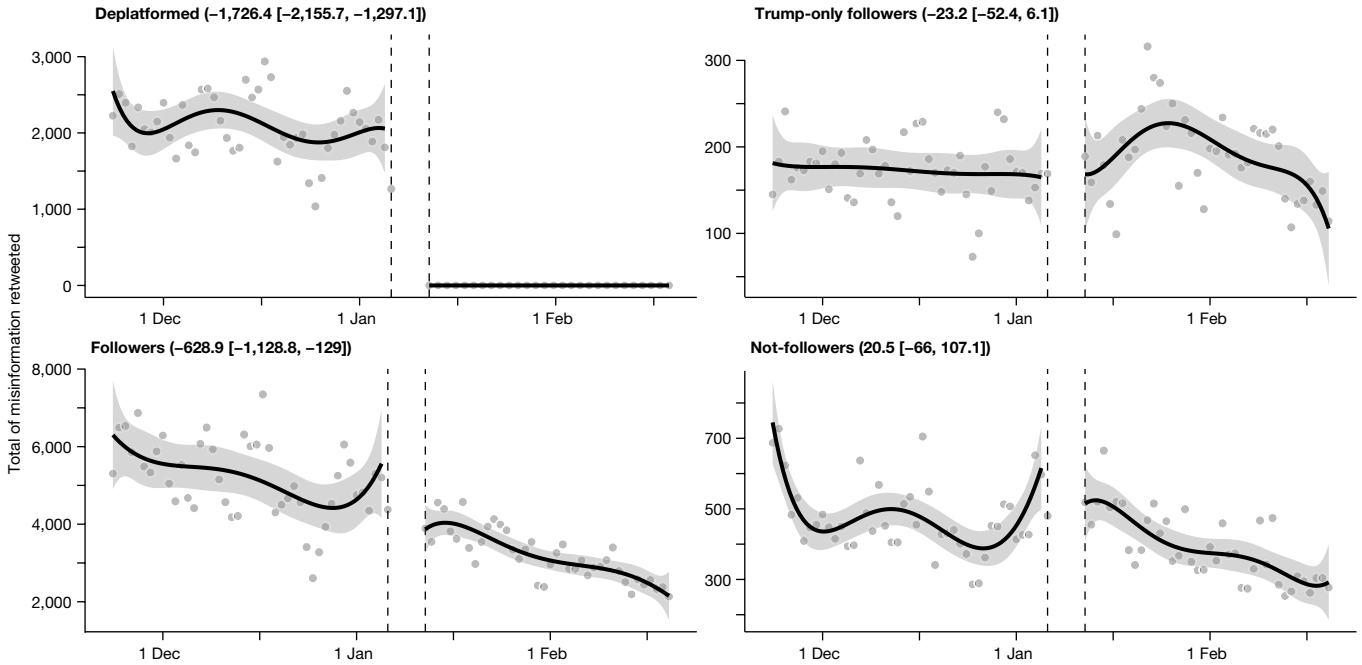


Fig. 2 | Reduction in misinformation retweets among misinformation users on Twitter following the deplatforming after January 6th. SRD estimates of the reduction in total misinformation retweets among users who shared at least one misinformation URL for deplatformed and not-deplatformed Twitter users. The SRD estimates are presented in parentheses above each graph, with 95% bias-corrected confidence intervals in square brackets. For the point estimates, the sample includes 546 observations (days) on average across groups, 404 before and 137 after 6 January 2021. The effective number of observations is 64.31 days before and after 6 January 2021 on average. The estimation excludes data between 6 January (cutoff point) and 12 January (included) 2021. The score on 6 January is set at 0, and the score on 12 January is 1. Optimal bandwidth of 32.7 days with triangular kernel and first-order polynomial. Fitted lines are five-degree polynomials, and shaded areas indicate 95% confidence intervals of the fitted lines.

During the Trump presidency, Twitter benefitted enormously from Trump's volatile tweets, which kept the platform at the centre of public discourse[35]. However, the insurrection created a moment of crisis for the company. From news accounts, as the violence in Washington, DC escalated, there was a series of urgent communications between the corporate centre and Jack Dorsey, the chief executive officer of Twitter at the time, who was on holiday in the South Pacific[36,37].

The unfolding crisis at Twitter is illustrated in Fig. 1, right, which shows the daily percentage of posts that include a 'misinformation URL' among all tweets and retweets containing a URL that users in our sample shared around time of the 2016 and 2020 elections (the dataset and classification of 'misinformation URLs' are described in Methods). Each panel shows two time series. The dark line indicates the percentage of misinformation posts among all users who shared any URL during the election cycle (the June before the election to the following February), and the light line indicates the percentage of misinformation posts among those users who shared at least one misinformation URL during each election cycle; the latter is a subset of the sample that we call 'misinformation sharers' and that our analyses below will focus on. Note that the misinformation URL sharing rate among all users was five times higher in 2020 than in 2016, and about twice as high among misinformation sharers, which suggests a marked change in the misinformation ecosystem on Twitter over time.

Note also that the rates of misinformation tweets and retweets remained relatively high after the November 2020 election until 6 January 2021, immediately after which the rate declined by about 50%. Some of this reduction is owing to a drop in misinformation sharing starting on that date, which we document below, and some of it is owing to a large increase in the amount of non-misinformation sharing that also occurred after that date. This pattern contrasts with the 2016 election cycle, during which the rate of misinformation sharing increased leading up to the November election, but declined immediately following Trump's victory.

Twitter's enforcement action was reported widely in the mainstream media[38,39] following an official announcement from Twitter[17], indicating that the deplatforming intervention targeted prominent conservative activists as well as accounts associated with QAnon that were influential in spreading election misinformation[40]. The immediate drop in misinformation following January 6th was also widely reported in the news[35,41,42]. For our purposes, January 6th is a moment that highlights the relationship between real-world violence and misinformation[43], as well as the attempt at purposeful control over political speech on Twitter by its leadership.

Research design

We aimed to explore whether social media companies have control over the spread of misinformation on their platforms. We use observational data to evaluate this question in two steps, using the subset of Twitter users in our panel who shared at least one misinformation URL during the 2020 election cycle.

First, we exploit Twitter's deplatforming intervention to evaluate whether a reduction in the circulation of misinformation occurred when Twitter suspended some of the most prolific spreaders of misinformation. We use a sharp regression discontinuity (SRD) design to identify the effect of the deplatforming intervention on the reduction of misinformation circulated by the deplatformed users via either tweets or retweets. In a press release dated 12 January 2021[17], Twitter announced that it conducted the deplatforming intervention on 8 to 12 January, although the press release indicated that the company began to suppress misinformation content in the days prior, and had suspended Donald Trump for 12 h on 6 January 2021[36]. We include outcomes starting on 12 January to measure post-intervention misinformation sharing.

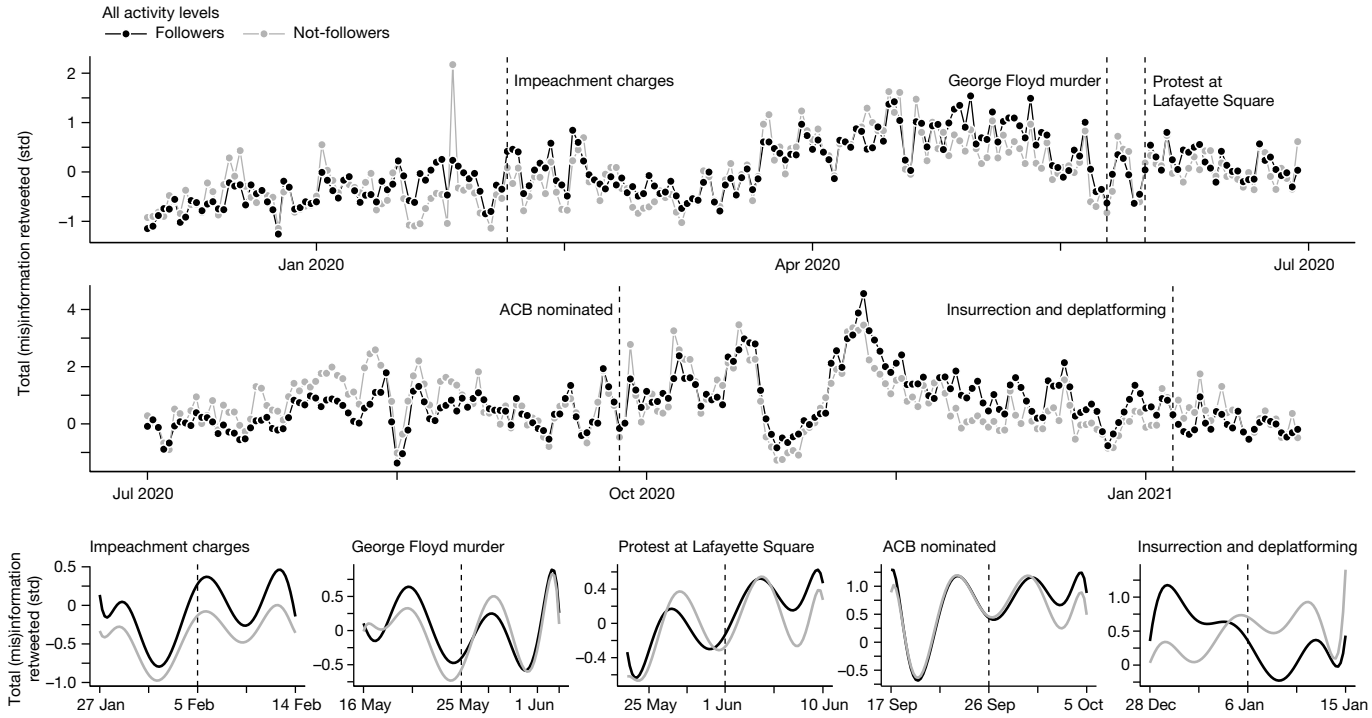


Fig. 3 | Time series of misinformation retweeting, for followers and not-followers of deplatformed users, across all activity levels. Top and middle, standardized (std) misinformation retweeting activity from 1 December 2019 to 29 January 2021. Vertical lines indicate times of highly salient events for comparison against the insurrection and deplatforming. Bottom, expanded view of activity nine days before and after each event, using an eight-degree polynomial to fit the trends. 'ACB nominated' refers to the nomination of Amy Coney Barrett to the position of Associate Justice of the Supreme Court of the USA.

Interpreting the magnitude of the SRD estimate as the causal effect of Twitter's intervention would depend on a strong 'continuity' assumption that these deplatformed users would have continued to share a similar amount of misinformation immediately after the intervention as they did immediately before, had the intervention not occurred[6]. This assumption is strong because the deplatforming intervention did not occur on a typical day; the deplatforming coincided with the insurrection and the election certification, which were widely covered in the media, and there is no way to separately identify the causal effects of these real-world political events from the intervention. The SRD bundles these events, and so this test does not disentangle the specific effect of the deplatforming unless one were to assume that the insurrection and certification events would have had no effect on deplatformed users' misinformation sharing behaviour, which would be a strong and probably implausible assumption.

This deplatforming intervention was necessarily causal in an obvious way, however, in that whatever misinformation those deplatformed users would have contributed was, by design, set to zero. If one were to assume that the deplatformed users would have continued to circulate some amount of misinformation after the insurrection, had the deplatforming not occurred, then our results can show that this causal effect is not zero. This is a reasonable assumption in that the deplatformed users were those who trafficked heavily in misinformation.

Second, we use a DID design to evaluate whether removing these deplatformed users as a source of misinformation also reduced misinformation circulation by other misinformation sharers who were not themselves deplatformed. Our DID design takes advantage of two of Twitter's affordances that are common to other social media platforms: users' ability to 'follow' other users and to recirculate or 'retweet' other users' posts. To set up our DID design, we created three mutually exclusive Twitter user types. First we divided all misinformation sharers in our sample into those who were deplatformed during the terms-of-use intervention and those who were not ('deplatformed' and 'not-deplatformed' groups). Then we further divide the not-deplatformed users into two types: those who followed at least one of the deplatformed users, who we refer to as 'followers' of the deplatformed accounts, and those who did not follow any deplatformed users, who we refer to as 'not-followers' of the deplatformed accounts.

For the followers, the deplatforming intervention directly reduced the sources of misinformation available in their feeds that they could retweet, compared to not-followers, and thus followers were more directly exposed to the intervention. At the same time, both followers and not-followers were similarly exposed to any confounding real-world events that coincided with the intervention. We thus can use weaker assumptions to identify causal effects of the deplatforming intervention in the DID compared with the SRD, given that all remaining users were similarly exposed to the confounding real world political events surrounding January 6th.

Identification of causal effects in the DID requires a 'parallel path' assumption, under which we can identify the amount of misinformation that followers would have posted had they stayed on the same misinformation sharing path as not-followers, had the intervention not occurred. We warrant the parallel path assumption by tracing the levels of misinformation circulation among followers and not-followers through prior political events between December 2019 and February 2021. The parallel path assumption holds that users responded to the insurrection the same way they did to previous events such as the George Floyd murder or when the Trump administration cleared protesters at Lafayette Square. This assumption is not testable, which means that caution is needed to interpret the results causally. But as these groups of users responded similarly to previous real-world events, it is reasonable to believe they responded similarly to the insurrection.

Coinciding with any behavioural changes among users, on 12 January 2021, Twitter announced that it had modified its algorithm to limit the visibility of misinformation and the ability of users to retweet misinformation[17]. Thus, this algorithm change is part of the intervention that

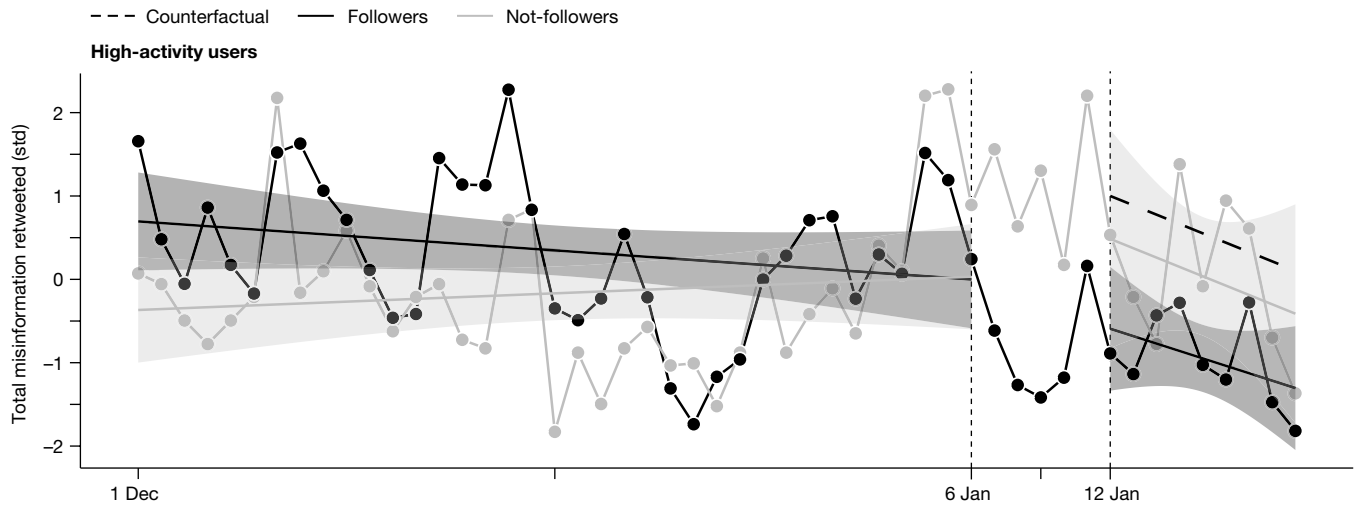


Fig. 4 | Time series of misinformation retweeting for followers and not-followers. Sample size includes 51 observations (days) from 1 December 2020 to 20 January 2021. The counterfactual identified under the parallel path assumption is shown as a dashed line after 12 January 2021. Fitted straight lines are ordinary least squares regressions of standardized daily total retweeted misinformation, fitted separately before 6 January 2021 and after 12 January 2021 and by group. The shaded areas around the fitted lines are 95% confidence intervals.

we evaluate in both the SRD and DID. Further, mainstream reporting of the intervention probably amplified the effect of the deplatforming intervention and is also bundled in.

Data

The details of the panel data collection, variable measurements and analyses are described in Methods. Our primary outcome is the daily counts of tweets and retweets, separately, containing URLs from a domain that appears on the lists of misinformation and disinformation domains described in the Methods. As a result, our measurement identifies URLs that point to unreliable sources of information or fake news, but does not evaluate the truth value of the contents. For simplicity, we refer to these URLs as 'misinformation' but the measurement does not distinguish misinformation from disinformation[44]. We examine both the average and the total count of misinformation URLs within each subgroup for both tweets and retweets.

This paper focuses on the behaviours of real residents of the USA on Twitter by using only accounts that had previously been matched to US voter file records[4,8]. The strength of this approach is that it enables us to evaluate the effect of the intervention on the speech of actual people, who from the perspective of democratic theory are the most important actors. The limitation is that it does not offer an understanding of the effect of the intervention on non-human accounts such as organizations or bots, as well as human users who choose to not post under their own name.

Within our panel of over 550,000 Twitter users who were active during the 2020 election cycle, 1,361 panellists (approximately 0.23%) were deplatformed. Extended Data Tables 1 and 2 show that deplatformed users were substantially older, more Republican, and (perhaps surprisingly) more likely to be women compared with the overall composition of the panel. These deplatformed users accounted for 4.35% of all URL sharing and 24.13% of all misinformation URLs shared among our sample during December 2020. Among the users who were not deplatformed, 26.4% were followers of at least one of the deplatformed accounts (including those who followed Trump). A total of 44,743 users (7.46% of the full panel) were misinformation sharers (users who shared at least one misinformation URL during the 2020 election cycle), which is the subset of interest for our SRD and DID analyses.

We use different definitions of follower in order to make comparisons. Our primary definition classifies users as followers of deplatformed accounts if they follow at least one account that was identified by Abilov et al.[45] as deplatformed. In Extended Data Fig. 1, we explore whether the 'dose' of misinformation matters by classifying users as followers if they follow k deplatformed accounts, where k is between 1 and 10. Finally, we separate those who exclusively followed Trump (who was a deplatformed user) from followers of the other deplatformed accounts, since users might have a variety of motivations for following Trump as a newsworthy public figure.

SRD results

Figure 1 shows the overall patterns of misinformation sharing for our full sample, as well as for the subset of users who circulate misinformation URLs. Figure 2 shows the SRD analysis of Twitter's intervention (which bundles the intervention with any effect of the real-world political events on 6 January 2021) on the total count of misinformation retweets specifically for each of our subgroups of interest, including point estimates of the SRD analysis with 95% robust bias-corrected confidence intervals[46]. In Extended Data Fig. 2, we show the effect for the means, as well as the corresponding results for tweeting misinformation.

Since Twitter targeted the worst offenders of its Civic Integrity Policy for deplatforming, it is no surprise the intervention coincided with a significant reduction in misinformation sharing among deplatformed users, for both tweets and retweets, as shown in Fig. 2, top left and Extended Data Fig. 2. In particular, deplatforming this group reduced the total amount of misinformation by 95 daily misinformation tweets and 1,726 misinformation retweets after the intervention within the Twitter panel. Because this contrast is confounded by the real-world events that occurred surrounding January 6th, the SRD cannot identify the magnitude of the causal effect; however, under the assumption that the deplatformed users would have circulated some amount of misinformation after 6 January 2021, we can assert that the causal effect for this subgroup is greater than zero.

The remaining panels of Fig. 2 report the SRD effects for the misinformation sharers who were not deplatformed, subsetting into whether or not they followed one of the deplatformed users. A reduction is apparent among the subset of users who followed the deplatformed users, but not among those who did not follow the deplatformed users. The data do not reject the null hypothesis that those who followed Trump but none of the other deplatformed users were not affected in the total

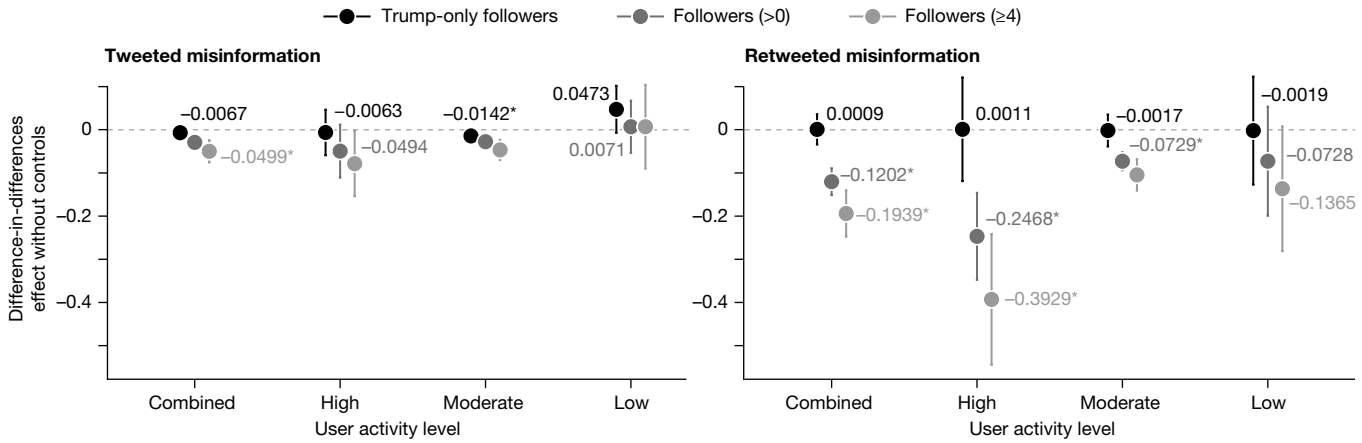


Fig. 5 | DID estimates of effect of deplatforming on followers of deplatformed Twitter users. DID two-way fixed-effect point estimates (dots) and 95% confidence intervals (bars) of the DID for high-, moderate- and low-activity users, as well as all activity levels combined. Standard errors are clustered at the user level. Estimates compare followers (treated group) and not-followers (reference group) of deplatformed users after 12 January 2021 (post-treatment period) and before 6 January 2021 (pre-treatment period). Supplementary Tables 1–3 show details with all activity level users combined. Total sample sizes of not-followers (reference) and Trump followers (treatment group): combined, 318,074; high, 56,273; moderate, 225,799; low, 35,208. Not-followers and followers (>0): combined, 688,174; high, 163,037; moderate, 461,644; low, 60,901. Not-followers and followers (≥4): combined, 481,684; high, 120,005; moderate, 310, 971; low, 48,718. *$P < 0.05$.

number of misinformation they shared, which is not a surprise given that many users followed Trump for reasons other than consumption of misinformation (these users actually show a slight increase in misinformation circulation after 12 January 2021 for unknown reasons, although the amount of the increase is comparatively small).

Extended Data Fig. 2 shows that we do not reject the null hypothesis that deplatforming had no effect on tweeting misinformation among the not-deplatformed users, suggesting that the deplatforming only matters for recirculating misinformation through retweets among these users, but not for introducing new misinformation through tweets. The effect on deplatformed users' tweets is significant but miniscule. We also conducted a series of placebo tests in which we do not expect to see significant effects. Extended Data Fig. 3 reports the SRD polynomials related to our placebo tests, showing that the intervention is unrelated to sharing of non-fake news, including conservative-leaning non-fake news. Extended Data Fig. 6 reports a placebo test on the circulation of information about sports and shopping, Extended Data Figs. 7 and 8 report placebo tests that use 20 December 2020 and 18 January 2021 as the cutoff dates, respectively.

DID results

In addition to the direct effect of the intervention on deplatformed users, we also expect to see a reduction in the retweeting of misinformation among those misinformation sharers who followed the deplatformed users (that is, the group that we are calling the followers), since these followers will have experienced a reduction in the availability of misinformation tweets in their own feeds. That we observe a reduction in retweets among the followers but not among the not-followers in the SRD might lead one to interpret the difference to demonstrate such a spillover effect, but the SRD design cannot support this conclusion even if one were to permit the strong SRD identification assumptions. The reason is that the SRD only estimates a local effect of the intervention—that is, the effect of the intervention at that moment in time. The SRD cannot rule out that the sizes of the causal effects could vary over time, and in particular, it is possible that the effect could vary between followers and not-followers if they respond to exogenous shocks in different ways. It is possible that the small reduction among not-followers is merely coincidental to that point in time.

To address this limitation, we must make and warrant a 'parallel path' assumption that the follower and not-follower retweet behaviours respond to exogenous shocks in a similar manner over time, among users who share misinformation. Theoretically, we have no reason to expect any such difference. In Fig. 3, we present a time series plot of retweeting behaviour for both groups over the whole of 2020, indicating with vertical lines the substantial political shocks that occurred throughout that year. The data collection for the panel measured high- and moderate-activity users more frequently than low-activity users. Figure 3 shows parallel trends for all activity-level groups combined, and similar patterns emerged when we ran the analysis separately by activity levels (Extended Data Fig. 4). Although there is some random variability, in general followers and not-followers maintain parallel paths throughout.

Under the parallel path assumption, the DID identifies the causal effect of the deplatforming intervention on the not-deplatformed misinformation sharers. Formally, the parallel path assumption holds that the followers would have remained on a parallel path to the not-followers after the intervention had the intervention not happened. In effect, the DID takes the not-followers as a control group that was similarly exposed to the real-world political events. Figure 4 shows the post-intervention counterfactual path of the followers under this assumption using a dashed line. Note that after the deplatforming intervention, the not-followers increased their levels of misinformation retweeting compared to before the intervention. The followers decreased their level of misinformation retweets, but had the followers remained on a parallel path as the not-followers, they would have instead increased the amount of misinformation sharing even more.

The difference between the followers' observed and counterfactual misinformation sharing shown in Fig. 4 is the causal effect of deplatforming under the DID assumptions. We estimate the magnitude of this causal effect using two-way fixed effects. Figure 5 summarizes the results. Estimates using adjustment covariates are presented in Extended Data Fig. 5. Misinformation retweeting activity of followers and not-followers were affected in different ways. There was an average reduction after the intervention of between 0.14 (adjusted) and 0.4 (unadjusted) misinformation retweets among high-activity followers of 4 or more deplatformed users relative to the change among not-followers. Specifically, high-activity misinformation sharers who were not followers of deplatformed users retweeted an average total of 154 misinformation posts per day in the 36 days between 1 December 2020 and 6 January 2021. This represents an average of 1,000 active users per day in our panel, who retweeted misinformation an average

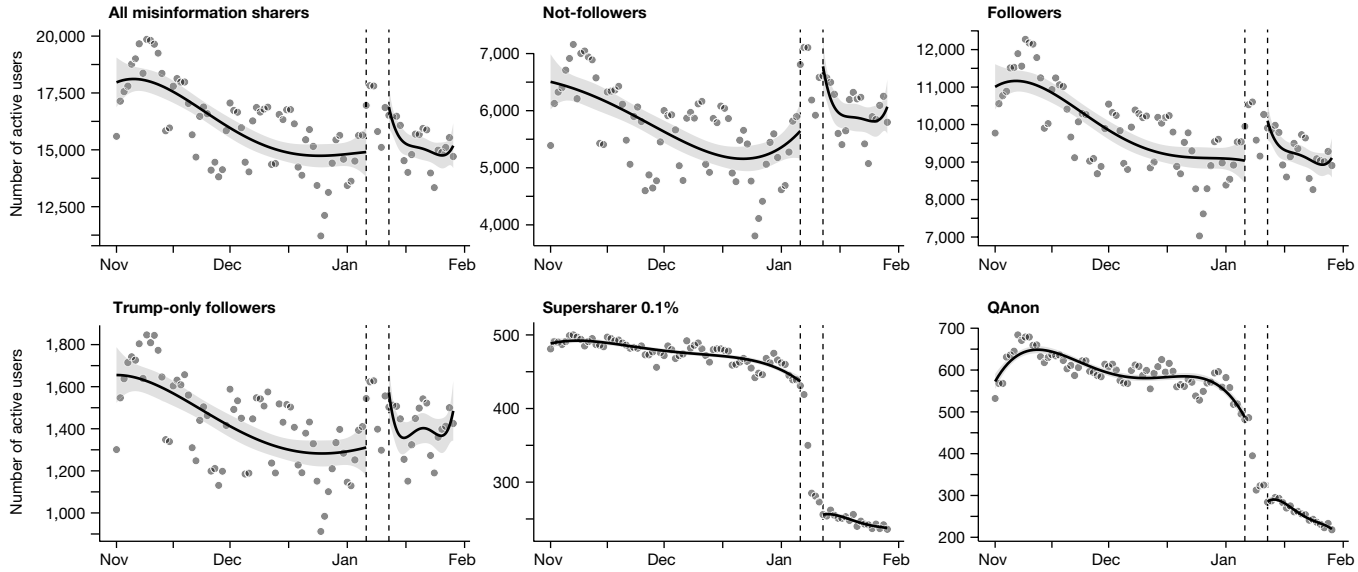


Fig. 6 | Time series of the number of not-deplatformed users within each subgroup. Four-degree polynomial regression (fitted line) before and after the deplatforming intervention, separated by subgroup. The subgroups are not mutually exclusive (Supplementary Fig. 1). Shaded area around the fitted line is the 95% confidence interval of the fitted values.

total of 162 times after the intervention, which is a 5% increase. Conversely, high-activity misinformation sharers who were followers of four or more deplatformed (on average, around 1,450 active users in that group in the period in our panel) retweeted 2,287 misinformation posts in that same before-intervention period. After the intervention, that number dropped to 1,607, representing a 30% reduction in misinformation sharing among the followers.

In Extended Data Fig. 1, we explore the DID effect of increasing the dose of the number of deplatformed users the follower follows, showing an increase in the size of the causal effect as the number increases from one to ten. The Methods section discusses how this test is a relaxation of the strong stable unit treatment value assumptions[7] (SUTVAs) used for the SRD. Notably, our test is a conservative estimate of the spillover effect in that it does not include second and higher order degree spillover, since one might expect not-followers to occasionally retweet followers' retweets from deplatformed accounts.

It is possible that part of the decline in misinformation in the not-deplatformed subgroups was due to the worst remaining offenders leaving Twitter in favour of alt-right platforms, even if they were not themselves deplatformed, in response to the then increasingly aggressive stance of Twitter towards election misinformation. To assess the extent of exit, we specify indicator variables for the subgroups of users who we identified as sharing similarities to those who were deplatformed. These subgroups include: those who followed deplatformed users, those who share misinformation, and those who share QAnon content. We define the misinformation sharer subgroup in two ways, corresponding to different levels of the intensity of misinformation sharing. First, as above, we classify whether a user has shared any misinformation URL during the 2020 election cycle. Second, we classify a supersharer as an individual who is in the top 0.1% of the distribution of misinformation sharing[8]. We measure QAnon users as the subset of users who used five distinct QAnon hashtags during the 2020 cycle[47]. QAnon users comprise 0.16% of the panel. As we show in Supplementary Fig. 1, these additional subgroups are not mutually exclusive.

Figure 6 evaluates exit among these groups of non-deplatformed users by examining the number of active users within each subgroup before and after the deplatforming intervention. Note that even though they themselves were not deplatformed, QAnon sharers and misinformation supersharers became distinctly less active after the intervention, and particularly so after 8 January 2021, but the other groups, including the followers, maintained a similar activity level as before the intervention. That is, individuals from targeted groups that remained on the platform seemed to have been especially likely to go quiescent after Twitter's intervention.

Discussion and conclusion

By 2020 Twitter had emerged as a virtual national town square in the USA, allowing the free flow of a wide range of speech. Despite the fact that its share of users, then at about 25% of the US adult population, was smaller than those of Facebook, Instagram, Google and YouTube[48], Twitter was the platform of choice among elected officials and journalists, and so had an outsized role in public discourse[49]. It is illustrative to note the lack of traction of Donald Trump's subsequent social media posts on other platforms, including Truth Social and his short-lived blog, which have received less attention despite their similar content. The content was accessible but lacked the machinery of dissemination and perceived social significance of a post on Twitter[35,42].

Our evaluation indicates that Twitter's deplatforming of the worst offenders of its Civic Integrity Policy affected not only the users who themselves were deplatformed, but also users who followed those who were deplatformed. We interpret these findings as causal effects in two specific ways.

First, if one were to assume the deplatformed users would have continued to circulate some amount of misinformation after the insurrection had the deplatforming not occurred, then we can say that the causal effect on this subgroup is greater than zero. We cannot precisely identify the magnitude of this causal effect, however, without adding a strong assumption that the events of January 6th themselves had no effect on these users' tendency to circulate misinformation. This latter assumption is probably implausible given the reality of the insurrection's violence and given that the certification of the election was likely to have changed the underlying reason that Trump supporters were circulating misinformation during this period.

Second, if one were to assume that the users who followed the deplatformed accounts responded to real world political events in the same way as do the users who did not follow the deplatformed accounts, among all users who circulated misinformation, we can identify the spillover effect of deplatforming on the users who followed the deplatformed accounts using DID. We presented evidence suggesting

that the parallel path assumption holds. In this case, we can take the not-followers as a control group that were exposed to the events of January 6th but not the misinformation posted by deplatformed users.

If one grants these assumptions, the findings demonstrate that Twitter exercised a degree of control over the circulation of misinformation on its platform, which propagated beyond the deplatformed users through their direct followers. These results are in contrast to those of Broniatowski et al.[3], who found that the sudden removal of anti-vaccine content on Facebook did not decrease overall engagement with anti-vaccine content. These authors identify Facebook's 'layered hierarchy' user interface for the failure of that intervention, in that Facebook's architecture provides redundant channels for misinformation to spread. The apparent effectiveness of Twitter's intervention implies either that such redundant channels do not exist to the same degree in Twitter's platform architecture, or that Twitter's strategy to remove users rather than content was more effective. Redundant channels of misinformation spread on a platform presumably do not matter if the users spreading misinformation are removed.

We emphasize that our design does not prove that the intervention is causal, in that there is no way for us to fully separate the effect of the insurrection and other real-world political events from Twitter's enforcement. Further, even if the deplatforming was effective in this way, it is likely to be partly owing to amplification from the widespread reporting that may not generalize to future deplatforming events. We do show, however, that many misinformation supersharers and traffickers in QAnon content chose to exit the platform after Twitter enhanced its enforcement stance, a compositional effect that probably accounts for some of the reduction in harmful content. The SRD analysis indicating the immediate drop in misinformation that had been circulated by the deplatformed users, and the DID establishing the followers of deplatformed users reduced their recirculating of misinformation via retweets, support the idea that Twitter's terms-of-use intervention probably resulted in a substantial reduction in misinformation trafficking overall, and indicates that the company had some degree of control over the quality of discourse on its own platform.

Twitter (now X) is now owned by X Corp. and ultimately by Elon Musk, under whose ownership content moderation has been vastly scaled back[50]. Irrespective of the future of X as a platform, we can be sure that social media platforms will continue to have a key role in civic discourse. Our data enable us to evaluate the counterfactual of the effectiveness of deplatforming under natural experimental designs on an occasion when a social media platform made the effort to do so. The results are informative for social media companies, regulators, elected officials and the general public to understand the capacities of social media companies as we consider regulations and alternatives to the market for public discourse as a public good.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-024-07524-8.

1. Lazer, D. The rise of the social algorithm. *Science* **348**, 1090–1091 (2015).
2. Jhaver, S., Boylston, C., Yang, D. & Bruckman, A. Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proc. ACM Hum.-Comput. Interact.* **5**, 381 (2021).
3. Broniatowski, D. A., Simons, J. R., Gu, J., Jamison, A. M. & Abroms, L. C. The efficacy of Facebook's vaccine misinformation policies and architecture during the COVID-19 pandemic. *Sci. Adv.* **9**, eadh2132 (2023).
4. Hughes, A. G. et al. Using administrative records and survey data to construct samples of tweeters and tweets. *Public Opin. Q.* **85**, 323–346 (2021).
5. Shugars, S. et al. Pandemics, protests, and publics: demographic activity and engagement on Twitter in 2020. *J. Quant. Descr. Digit. Media* https://doi.org/10.51685/jqd.2021.002 (2021).
6. Imbens, G. W., & Lemieux, T. Regression discontinuity designs: a guide to practice. *J. Econom.* **142**, 615–635 (2008).
7. Gerber, A. S. & Green, D. P. *Field Experiments: Design, Analysis, and Interpretation* (W.W. Norton, 2012).
8. Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. & Lazer, D. Fake news on Twitter during the 2016 U.S. presidential election. *Science* **363**, 374–378 (2019).
9. Munger, K. & Phillips, J. Right-wing YouTube: a supply and demand perspective. *Int. J. Press Polit.* **27**, 186–219 (2022).
10. Guess, et al. How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science* **381**, 398–404 (2023).
11. Persily, N. in *New Technologies of Communication and the First Amendment: The Internet, Social Media and Censorship* (ed. Bollinger L. C. & Stone, G. R.) (Oxford Univ. Press, 2022).
12. Sevanian, A. M. Section 230 of the Communications Decency Act: a 'good Samaritan' law without the requirement of acting as a 'good Samaritan'. *UCLA Ent. L. Rev.* https://doi.org/10.5070/LR8211027178 (2014).
13. Lazer, D. M. J. et al. The science of fake news. *Science* **359**, 1094–1096 (2018).
14. Suzor, N. Digital constitutionalism: using the rule of law to evaluate the legitimacy of governance by platforms. *Soc. Media Soc.* **4**, 2056305118787812 (2018).
15. Napoli, P. M. *Social Media and the Public Interest* (Columbia Univ. Press, 2019).
16. DeNardis, L. & Hackl, A. M. Internet governance by social media platforms. *Telecomm. Policy* **39**, 761–770 (2015).
17. TwitterSafety. An update following the riots in Washington, DC. *Twitter* https://blog.x.com/en_us/topics/company/2021/protecting--the-conversation-following-the-riots-in-washington-- (2021).
18. Twitter. Civic Integrity Policy. *Twitter* https://help.twitter.com/en/rules-and-policies/election-integrity-policy (2021).
19. Promoting safety and expression. *Facebook* https://about.facebook.com/actions/promoting-safety-and-expression/ (2021).
20. Dwoskin, E. Trump is suspended from Facebook for 2 years and can't return until 'risk to public safety is receded'. *The Washington Post* https://www.washingtonpost.com/technology/2021/06/03/trump-facebook-oversight-board/ (4 June 2021).
21. Huszár, F. et al. Algorithmic amplification of politics on Twitter. *Proc. Natl Acad. Sci. USA* **119**, e2025334119 (2021).
22. Guess, A. M., Nyhan, B. & Reifler, J. Exposure to untrustworthy websites in the 2016 US election. *Nat. Hum. Behav.* **4**, 472–480 (2020).
23. Sunstein, C. R. *#Republic: Divided Democracy in the Age of Social Media* (Princeton Univ. Press, 2017).
24. Timberg, C., Dwoskin, E. & Albergotti, R. Inside Facebook, Jan. 6 violence fueled anger, regret over missed warning signs. *The Washington Post* https://www.washingtonpost.com/technology/2021/10/22/jan-6-capitol-riot-facebook/ (22 October 2021).
25. Chandrasekharan, E. et al. You can't stay here: the efficacy of Reddit's 2015 ban examined through hate speech. *Proc. ACM Hum. Comput. Interact.* **1**, 31 (2017).
26. Matias, J. N. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proc. Natl Acad. Sci. USA* **116**, 9785–9789 (2019).
27. Yildirim, M. M., Nagler, J., Bonneau, R. & Tucker, J. A. Short of suspension: how suspension warnings can reduce hate speech on Twitter. *Perspect. Politics* **21**, 651–663 (2023).
28. Guess, A. M. et al. Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science* **381**, 404–408 (2023).
29. Nyhan, B. et al. Like-minded sources on Facebook are prevalent but not polarizing. *Nature* **620**, 137–144 (2023).
30. Dang, S. Elon Musk's X restructuring curtails disinformation research, spurs legal fears. *Reuters* https://www.reuters.com/technology/elon-musks-x-restructuring-curtails-disinformation-research-spurs-legal-fears-2023-11-06/ (6 November 2023).
31. Duffy, C. For misinformation peddlers on social media, it's three strikes and you're out. Or five. Maybe more. *CNN Business* https://edition.cnn.com/2021/09/01/tech/social-media-misinformation-strike-policies/index.html (1 September 2021).
32. Conger, K. Twitter removes Chinese disinformation campaign. *The New York Times* https://www.nytimes.com/2020/06/11/technology/twitter-chinese-misinformation.html (11 June 2020).
33. Timberg, C. & Mahtani, S. Facebook bans Myanmar's military, citing threat of new violence after Feb. 1 coup. *The Washington Post* https://www.washingtonpost.com/technology/2021/02/24/facebook-myanmar-coup-genocide/ (24 February 2021).
34. Barry, D. & Frenkel, S. 'Be there. Will be wild!': Trump all but circled the date. *The New York Times* https://www.nytimes.com/2021/01/06/us/politics/capitol-mob-trump-supporters.html (6 January 2021).
35. Timberg, C. Twitter ban reveals that tech companies held keys to Trump's power all along. *The Washington Post* https://www.washingtonpost.com/technology/2021/01/14/trump-twitter-megaphone/ (14 January 2021).
36. Dwoskin, E. & Tiku, N. How Twitter, on the front lines of history, finally decided to ban Trump. *The Washington Post* https://www.washingtonpost.com/technology/2021/01/16/how-twitter-banned-trump/ (16 January 2021).
37. Harwell, D. New video undercuts claim Twitter censored pro-Trump views before Jan. 6. *The Washington Post* https://www.washingtonpost.com/technology/2023/06/23/new-twitter-video-jan6/ (23 June 2023).
38. Romm, T. & Dwoskin, E. Twitter purged more than 70,000 accounts affiliated with QAnon following Capitol riot. *The Washington Post* https://www.washingtonpost.com/technology/2021/01/11/trump-twitter-ban/ (11 January 2021).
39. Denham, H. These are the platforms that have banned Trump and his allies. *The Washington Post* https://www.washingtonpost.com/technology/2021/01/11/trump-banned-social-media/ (13 January 2021).
40. Graphika Team. DisQualified: network impact of Twitter's latest QAnon enforcement. *Graphika Blog* https://graphika.com/posts/disqualified-network-impact-of-twitters-latest-qanon-enforcement/ (2021).

41. Dwoskin, E. & Timberg, C. Misinformation dropped dramatically the week after Twitter banned Trump and some allies. *The Washington Post* https://www.washingtonpost.com/technology/2021/01/16/misinformation-trump-twitter/ (16 January 2021).
42. Harwell, D. & Dawsey, J. Trump is sliding toward online irrelevance. His new blog isn't helping. *The Washington Post* https://www.washingtonpost.com/technology/2021/05/21/trump-online-traffic-plunge/ (21 May 2021).
43. Olteanu, A., Castillo, C., Boy, J. & Varshney, K. The effect of extremist violence on hateful speech online. In *Proc. 12th International AAAI Conference on Web and Social Media* https://doi.org/10.1609/icwsm.v12i1.15040 (ICWSM, 2018).
44. Lin, H. et al. High level of correspondence across different news domain quality rating sets. *PNAS Nexus* **2**, gad286 (2023).
45. Abilov, A., Hua, Y., Matatov, H., Amir, O., & Naaman, M. VoterFraud2020: a multi-modal dataset of election fraud claims on Twitter." *Proc. Int. AAAI Conf. Weblogs Soc. Media* **15**, 901–912 (2021).
46. Calonico, S., Cattaneo, M. D. & Titiunik, R. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica* **82**, 2295–2326 (2014).
47. Jackson, S., Gorman, B. & Nakatsuka, M. *QAnon on Twitter: An Overview* (Institute for Data, Democracy and Politics, George Washington Univ. 2021).
48. Shearer, E. & Mitchell, A. News use across social media platforms in 2020. *Pew Research Center* https://www.pewresearch.org/journalism/2021/01/12/news-use-across-social-media-platforms-in-2020/ (2021).
49. McGregor, S. C. Social media as public opinion: How journalists use social media to represent public opinion. *Journalism* **20**, 1070–1086 (2019).
50. Hammond-Errey, M. Elon Musk's Twitter is becoming a sewer of disinformation. *Foreign Policy* https://foreignpolicy.com/2023/07/15/elon-musk-twitter-blue-checks-verification-disinformation-propaganda-russia-china-trust-safety/ (15 July 2023).

Methods

The data were collected under Northeastern IRB protocol no. 17-12-13. To protect user privacy, the Northeastern-based research team has produced user-level aggregate data from the raw tweets, obscuring the actual user identifiers using salted one-way hashes, and this anonymized user-level data is what is analysed here.

Twitter panel data

This study uses a dataset of tweets and retweets that was previously collected for Grinberg et al.[8], Hughes et al.[4] and Shugars et al[5]. The full dataset (the 'Twitter panel') is a dataset of 1.5 million Twitter users whose identities have been linked to a commercial US voter file. For each of these users, we have regularly collected all panellists' tweets and retweets since late 2017. The data collection began in late 2017, so we can only be sure of full coverage at that point—but we have historical data (the most recent 3,200 tweets from each user around autumn 2017) that stretches back further in time; for most users this includes all of their Tweets from 2016. This has produced a sample of 3 billion tweets and retweets dating from the beginning of Twitter to 12 July 2021. All time stamps are standardized to Eastern Standard Time.

We limit the analysis to active users matched to the US voter file. We define an active user as anyone who shared (via tweet or retweet) one URL during the 2020 election cycle between 30 June 2020 and 31 January 2021. There are 599,697 active users in our panel under this definition, who tweeted more than 8 million URLs during the study period. We exclude 11 users who were deplatformed around the time of the January 2021 terms-of-use intervention but had their accounts restored. Beyond that, we do not exclude any data.

The SRD and DID analyses focus on the subset of users who shared at least one misinformation URL; there are 44,734 such users. We use the full Twitter panel to create Fig. 1; it is only our estimation sample for the SRD and DID that excludes panellists who are known to never share misinformation. In the analysis, we subsetted the estimation sample in this way to ensure that the DID comparison groups—the followers and not-followers—are reasonably comparable to each other. A large proportion of the follower group engages in misinformation sharing. If we were to use the full panel in the DID, that would make the not-follower group largely composed of users who never share misinformation, and thus would make the not-follower group possibly different from the follower group. Selecting on the panel members who have circulated misinformation to define both followers and not-followers makes the two groups reasonably comparable, given any latent causes for this selection, and thus weakens the parallel path assumption required in that design. For consistency, we used the same estimation sample for the SRD, although the selection has no material effect on the SRD results since the excluded users contribute zero misinformation URLs both before and after the cutoff date. This decision reduces the estimation sample for both analyses since only about 10% of the panellists share misinformation URLs and so is conservative regarding power.

In practice, this decision on the estimation sample has no effect on the results. When re-estimating the DID results using all URL sharers rather than the subset of misinformation sharers, we get almost identical results for both the parallel path analysis and the DID treatment effect estimates, although the DID estimates using the full sample are a little larger and the parallel path a little less warranted. We prefer to report the results using the subset of misinformation sharers because they are more conservative and conceptually apparently more defensible.

Measured variables

Our main outcome variables are each defined as the number of misinformation URLs shared each day. For each tweet or retweet, we perform the following three-step procedure:

(1) We extract all URLs contained in each tweet or retweet. (If a tweet or retweet contains no URLs, it is excluded). We examine user-authored tweets and retweets separately. User-authored tweets that contain a misinformation URL introduce new misinformation into the system, whereas retweets circulate already existing URLs more widely within the system. We omit user-authored quote tweets; although quote tweets help to amplify controversial content or figures, it can be difficult to determine a user's stance towards the quoted material[51].
(2) For each URL, we follow the link and any associated redirects and record the final domain it points to (for example, this method resolves links from the URL shortener https://nyti.ms to https://nytimes.com).
(3) For each domain, we determine whether or not it is a fake-news domain by referring to a predefined list of fake-news domains. Our list of fake-news domains is defined by combining data from Grinberg et al.[8] and the proprietary list from the journalism company NewsGuard (https://www.newsguardtech.com/newsguard-for-researchers), using the 29 March 2021 snapshot. NewsGuard gives websites a score between 0 and 100, with a score below 60 indicating a low-credibility domain. As robustness checks, we perform the analyses with the Grinberg list alone, the NewsGuard list alone, and Indiana University's Iffy list as alternative definitions and report these in Supplementary Fig. 5.

Our domain list includes the list of 'fake news' sites from Grinberg et al.[8], which operationalizes "sites that lack editorial norms and processes to ensure the credibility of information." Newsguard's index uses a combination of attributes—editorial processes, fact checks and financial disclosure—to capture the 'reliability' or 'quality' of a domain. These criteria capture both misinformation and disinformation domains but our measurement does not evaluate the truth value of a URL's content. Although the specific criteria can vary from list to list, Lin et al.[44] find high levels of agreement between domain lists.

Our threshold of 60 points for the Newsguard list is based on the recommendations of Newsguard. To give an idea as to what domains are on either side of that threshold, the two most shared domains with a Newsguard score above 50 but below 60 are https://www.judicialwatch.org/ (which is also red in the Grinberg list) and https://www.dailywire.com/ (which is orange in Grinberg). The three most shared domains with a Newsguard score above 60 but below 70 are https://www.foxnews.com/ (not in Grinberg), https://www.msnbc.com/ (not in Grinberg), and https://dailycaller.com/ (orange in Grinberg). Notably, Newsguard later revised their scores of Fox News and MSNBC to place them below the 60-point threshold; our snapshot predates this decision, unlike the scores presented in Lin et al.[44] where https://www.foxnews.com/ is shown as having a score of 57.

Our supplementary outcome variables are defined as the number of non-fake conservative news domains shared in each day and the number of non-fake liberal news domains shared in each day. Among non-fake URLs, we use a similar URL-resolving procedure to determine whether the URL represented non-fake conservative news, or non-fake liberal news, defined as all domains with audience-based slant scores in the top and bottom quintile of the scores defined in Robertson et al.[52].

Subgroups

We construct group indicator variables for the groups we have identified as potentially affected by the terms-of-use intervention: the users who were themselves deplatformed, and then among those not deplatformed, we subsetted those who followed deplatformed users or those who did not follow deplatformed users. We identify deplatformed users in our panel by collecting a snapshot of all user profiles for active members of the Twitter panel; this snapshot was collected between 14 January 2021 and 27 January 2021. If we failed to collect a user profile, the Twitter API returned an error code indicating that the user has either deleted their account or has been suspended. We define a user as deplatformed if they were active prior to the intervention and suspended at the time we snapshotted their user profile on or after 14 January 2021.

We identify followers of deplatformed users using a snapshot of following behaviour collected between 19 September 2020 and 7 January 2021. We identify all accounts in our panel who follow any users described in the definition of accounts engaged in spreading voter-fraud-related misinformation (from Abilov et al.[45]) who were suspended.

In the analysis, we define a 'follower' as a user that followed any deplatformed user, including then-President Trump. For comparison, we construct alternate measures that require the user to follow at least k deplatformed users including Trump, with $0 < k < 11$. Finally, we use a third definition that includes those who followed Trump but none of the other deplatformed users; because Trump was the elected President of the USA at the time, users may have had different motivations to follow him given his tweets were seen as newsworthy.

For alternative supplementary comparisons, we also identified those who share misinformation and those who share QAnon content. We define the misinformation sharer group in three ways. First, we classify whether a user has shared even a single misinformation URL during the study period. Second, we define a supersharer, following Grinberg et al.[8] as an individual who is in the top 0.1% of the distribution of misinformation sharing among the panel (excluding deplatformed users). Finally, we measure QAnon users as those who use at least five of the list of hashtags in Jackson et al.[47] during the study period.

Within our panel of 0.5 million active Twitter users, 1,361 panellists were deplatformed, which accounts for 0.23% of our panellists. These deplatformed users account for 4.35% of content on Twitter in December 2020 and 24.13% of misinformation URLs. Among our subgroups, 1% supersharers account for 0.1% of the sample; QAnon traffickers are 0.16%; and the followers of the deplatformed users (including those following Trump) are 26.4%. As we show in Supplementary Fig. 1, these subgroups have considerable overlap and are not mutually exclusive.

The data for analysis (merging subgroups and demographics with the user URL sharing activity) and the files with daily summaries (percentage, average and total) were prepared using Python 3.11.6.

Missing data

Some tweets and retweets are missing. Our data collection progress did not occur in real time, therefore it is possible that we are missing tweets or retweets that were deleted shortly after they were sent. In our data collection process, we partition users into 3 levels based on activity level—the 10,000 most active users (high activity) are collected daily, the next 500,000 most active users (moderate activity) are collected approximately weekly, and then everyone else (low activity) is collected approximately biweekly. Collection dates within each of these user types are randomized. Because Twitter suspensions were concentrated among the most active users, we expect potential missingness in the less active users to have minimal effect. In the SRD what matters are the expected counts on each side of the cutoff. Because of the collection procedures, the composition of users will be different on each side of the cutoff, and the difference in estimated expectations is the treatment effect. SRD does not require the same units on both sides of the cutoff, and randomization of the collection date within user types makes this aspect of continuity plausible. In the DID analysis we report all results for the combined users as well as by user type.

SRD design

We use a SRD design to test for the direct effect of the deplatforming intervention. The running variable is the date, or the days before and after the intervention. Here we focus on the effect of the terms-of-use intervention on those users who were deplatformed, that is, those who had their account removed by the intervention on 8 to 12 January 2021 (with aspects of the intervention started on 6 and 7 January 2021), and when Twitter imposed a separate restriction on retweeting of harmful content on 12 January 2021. Because these users' sharing is deterministically zero post-intervention, this test only depends on the quantity of misinformation sharing among these users prior to the intervention; it evaluates the causal effect of the intervention on those deplatformed under a continuity assumption that the misinformation sharing would have occurred at the same rate had the intervention not occurred.

The treatment is the Twitter administration intervention between 8 and 12 January (inclusive), although the company banned Trump on 6 January 2021 and began algorithmic suppression of content on that day[17]. The results in Fig. 2 show the SRD results comparing misinformation sharing up to and including 6 January 2021 as the cutoff point to misinformation sharing after 12 January 2021, including this date.

We adopt all standard SRD analysis procedures, as described[6], including robustness checks to evaluate any influence of model specifications, such as selecting the optimal bandwidth and testing the robustness of the kernel selection. We include as robustness and model specification checks the evaluation of the continuity of the running variable at the cutoff point[53], and several placebo tests. We test the robustness of the treatment effect across different cutoff dates using standard methods and show the results in Supplementary Figs. 3 and 4. Supplementary Fig. 5 shows the results under different measures of misinformation to show the robustness of the results to different misinformation URL lists. The placebo tests consider sharing of shopping-related or sports-related URLs in Extended Data Fig. 6, and evaluating cutoffs on non-intervention dates (20 December 2020 and 18 January 2021) shown in Extended Data Figs. 7 and 8. In the placebo test, if our expectations are correct, the behaviour of those users should not change due to the intervention. This will indicate that the intervention did not affect the behaviour it was not supposed to, increasing our confidence on the effect of the intervention on misinformation sharing.

To test the statistical parameters, we evaluate the statistical significance of differences of average values of the within-group outcome near the cutoff point before and after the intervention for each group of users described above. Our SRD model does not include any covariates. The parametric specification for the SRD refers to the regression model used to estimate the conditional expectations before and after the intervention. We use a local linear regression with a rectangular kernel, and evaluate the dependence of the results on those choices (robustness checks) by varying the kernel choice (such as gaussian kernel, triangular kernel, and others), sensitivity to the choice of the cutoff, and we always choose the optimal bandwidth. We find that our results are robust to using the uniform kernel and the Epanechnikov kernel (results not shown). We use bias-corrected 95% confidence intervals first derived in Calonicao et al.[46].

DID design

Our DID design uses two-way fixed effects with ordinary least squares estimators and clustered standard errors at the user level. The outcomes are total misinformation tweets and retweets for high-, moderate- and low-activity users, among all users who shared at least one misinformation URL during the previous year. We selected cases between 1 December 2020 and 6 January 2021 as the pre-treatment period, and 12 to 29 January 2021 as the post-treatment periods, and used binary indicators for pre- and post-treatment periods and followers and not-followers of deplatformed users. As usual, the interaction between these indicators give the causal effect of interest, in this case the effect of deplatforming intervention on misinformation sharing activity among followers of deplatformed users, among users who had previously circulated misinformation.

Figure 3 shows the time series plot to warrant the parallel path assumption comparing deplatformed follower activity versus deplatformed not-follower activity for all users who shared at least one misinformation URL during the study period. The replication materials contain the script to recreate these figures for each of the three definitions of follower (follower, Trump-only follower, 4+ follower) for each activity level (combined, high, moderate and low). The plots showing each of these comparisons are similar so we omit them here.

Supplementary Tables 1–3 show the results for the DID analysis for each of three definitions of follower (deplatformed follower, Trump-only follower, 4+ follower) that combine across the activity levels. The replication package contains the corresponding tables that show the results separately for high-, moderate- and low-activity users.

DID is a well-established identification strategy used to estimate causal effects with observational data in various fields[54,55], and it is well-suited for our analysis. Most of the concerns regarding the application of the DID approach revolve around the appropriateness of the untestable identification assumptions that are specific to the DID design. These assumptions include the parallel path assumption and the no anticipation (or no prescience) assumption.

No anticipation refers to the assumption that individuals in both the treated and control groups were unaware of the upcoming treatment and therefore did not self-select into either group. In the context of our research, this means that followers and not-followers were not aware that a massive deplatforming was going to take place until it actually occurred. Therefore these users did not consciously select whom to follow or change their level of misinformation sharing activity in response to that anticipation. In other words, followers did not intentionally stop following users who were later deplatformed because they anticipated their removal in advance. Similarly, not-followers did not suddenly start following deplatformed users and retweeting these users' misinformation posts prior to the intervention because not-followers anticipated the removal of the deplatformed accounts. This assumption appears reasonable in our application.

Parallel paths refer to the assumption that the difference in the average outcome between treated and untreated groups would have stayed the same as it was prior to the treatment, if the treated group had not received the treatment. In classical DID designs, there are typically two time periods (before and after the treatment) and two groups (treated and untreated). In this scenario, since there is only one observation of the average outcome per group before the treatment, there is limited evidence to suggest a parallel path existed prior to the treatment. Additionally, because there is only one observation of the average for each group after the treatment, there is little evidence to suggest that the paths would continue to be parallel after the treatment. However, in our case, we have the advantage of having multiple observations both before and after the treatment. Our analysis, particularly Figs. 3 and 4, provides robust evidence that followers and not-followers exhibited parallel behaviour leading up to the events surrounding January 6th. Specifically, Fig. 3 demonstrates that parallel paths remained largely intact across various significant political occurrences, such as the impeachment charges against Trump and the protest at Lafayette Square. The only time this trend shifted was during the deplatforming period.

One problem that arises in multi-period DID analysis is the estimation of treatment effects using two-way fixed effects when there are staggered treatments—that is, treatments occurring at different times for different groups with varying effect sizes[56]. In such cases, two-way fixed effects can lead to biased estimates. Although there are solutions available for this issue[57], we are not faced with this issue in our analysis. First, we are comparing only two groups: followers and not-followers. Second, we do not have staggered treatments. Third, our parallel paths analysis indicates that parallel paths hold both before and after the treatment. As shown in Fig. 3, the paths of followers and not-followers appear to be parallel before and after the deplatforming period, with a shift in their average relative levels. Figure 4 further confirms that the parallel paths assumption is reasonable and that the paths in the post-treatment period remain parallel, albeit with average level shifts among followers and not-followers. This is evident from the counterfactual dotted line after the treatment period and the observed fitted line for both followers and not-followers during the pre- and post-treatment periods.

Although these observations cannot prove that the parallel paths assumption holds—this assumption, as any other identification assumption, is untestable—they give us confidence in the reliability and appropriateness of that assumption in our application, especially when compared to classical applications with two periods and two groups. The major advantage of the DID approach in our analysis is that it can be argued that both followers and not-followers were equally exposed to the insurrection and the media coverage of the events, but the deplatforming affected these groups differently for obvious reasons. Combined, our results suggest that the deplatforming had significant consequences that extended beyond its direct effect on those who were deplatformed.

Estimation

All estimation was conducted in the statistical software R[58]. The estimation of the SRD models uses the R package rdrobust[59]. The estimation uses local polynomial regression discontinuity point estimators with robust bias-corrected confidence intervals. Details of the estimation and inference procedures can be found in refs. 46,60,61. The estimation of the DiD parameters used the function lm of the R base package stats, which implements linear regression models. The point estimates were computed using ordinary least squares. The standard errors were clustered at user level using the function coeftest of the package lmtest[62] combined with the clustered sandwich estimator[63], which is implemented using the package sandwich and the function vcovCL[64].

We report the results of all statistical tests. In the reported results we do not correct the implied *P* values for multiple comparisons. Instead, since we are reporting all results, readers can understand the number of significant results compared to a known expected false positive rate. Supplementary Table 5 shows the *P* values for the DID analyses under different multiple comparisons corrections, each of which offers a different trade-off between type I and type II errors. In most cases, the *P* values that are significant in the raw results are also significant under the corrections. The project initially filed a pre-analysis plan available at the project repository (https://doi.org/10.17605/OSF.IO/KU8Z4) but we abandoned the pre-analysis plan because it did not anticipate the DID analysis which is the core of this paper. The pre-analysis plan instead had a less defensible set of subgroup analyses.

Network interference and our SUTVA assumptions

The SRD and DID use different interference assumptions[7]. First, our SRD analysis assumes that the deplatformed users only have potential outcomes defined by their own exposure to the intervention z_i—that is, two potential outcomes only, with the treated outcome deterministically zero, $Y(z_i = 1) = 0$, the control outcome determined by the continuity assumption $Y(z_i = 0)$, and the strong SUTVA assumption implies that the vector of assignments of everyone else in the network is ignorable for this analysis. This assumption is strong but supportable since the deplatforming intervention had a direct effect on deplatformed users that was not dependent on the deplatforming status of any other users. Second, in the SRD and DID we also assume that the not-deplatformed users have potential outcomes defined both by their own assignment (which is always $z_i = 0$) and by the number of deplatformed accounts they follow, rather than the specific vector of accounts that were deplatformed. This is a simplifying assumption because assuming the full vector of assignments matters would lead to an astronomical number of potential outcomes that could not be modeled.

To state this assumption for not-deplatformed users formally, define an assignment vector $\mathbf{z}_{-i}$ that tracks the assignments of units one degree connected in the network only and z_i indicates the unit's own assignment. Then for $K \geq 1$, the potential outcomes are $Y(z_i = 0, \mathbf{1}'\mathbf{z}_{-i} \geq K)$ for followers and $Y(z_i = 0, \mathbf{1}'\mathbf{z}_{-i} = 0)$ for not-followers. This in effect is a weakened SUTVA assumption that the deplatformed users are exchangeable but not ignorable, and the vector of assignments of the remaining units in the network is ignorable, and this simplifying assumption enables us to model spillover in the network. This is similar to the assumptions in ref. 65, which assumes independence

at some network degree. Finally, in the DID, the treatment group is defined as whether the user follows K deplatformed accounts ($K \geq 1$), and the control group is not-followers, and the potential outcome for the treated group is defined by the parallel path assumption. For the robustness tests, we set $K = \{1,2,3,4,5,\ldots,10\}$.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Aggregate data used in the analysis are publicly available at the OSF project website (https://doi.org/10.17605/OSF.IO/KU8Z4) to any researcher for purposes of reproducing or extending the analysis. The tweet-level data and specific user demographics cannot be publicly shared owing to privacy concerns arising from matching data to administrative records, data use agreements and platforms' terms of service. Our replication materials include the code used to produce the aggregate data from the tweet-level data, and the tweet-level data can be accessed after signing a data-use agreement. For access requests, please contact D.M.J.L.

Code availability

All code necessary for reproduction of the results is available at the OSF project site https://doi.org/10.17605/OSF.IO/KU8Z4.

51. Joseph, K. et al. (Mis)alignment between stance expressed in social media data and public opinion surveys. *Proc. 2021 Conference on Empirical Methods in Natural Language Processing* 312–324 (Association for Computational Linguistics, 2021).
52. Robertson, R. E. et al. Auditing partisan audience bias within Google search. *Proc. ACM Hum. Comput. Interact.* **2**, 148 (2018).
53. McCrary, J. Manipulation of the running variable in the regression discontinuity design: a density. *Test* **142**, 698–714 (2008).
54. Roth, J., Sant'Anna, P. H. C., Bilinski, A. & Poe, J. What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *J. Econom.* **235**, 2218–2244 (2023).
55. Wing, C., Simon, K. & Bello-Gomez, R. A. Designing difference in difference studies: best practices for public health policy research. *Annu. Rev. Public Health* **39**, 453–469 (2018).
56. Baker, A. C., Larcker, D. F. & Wang, C. C. Y. How much should we trust staggered difference-in-differences estimates? *J. Financ. Econ.* **144**, 370–395 (2022).
57. Callaway, B. & Sant'Anna, P. H. C. Difference-in-differences with multiple time periods. *J. Econom.* **225**, 200–230 (2021).
58. R Core Team. R: A Language and Environment for Statistical Computing, v.4.3.1. https://www.R-project.org/ (2023).
59. rdrobust: Robust data-driven statistical inference in regression-discontinuity designs. https://cran.r-project.org/package=rdrobust (2023).
60. Calonico, S., Cattaneo, M. D. & Titiunik, R. Optimal data-driven regression discontinuity plots. *J. Am. Stat. Assoc.* **110**, 1753–1769 (2015).
61. Calonico, S., Cattaneo, M. D. & Farrell, M. H. On the effect of bias estimation on coverage accuracy in nonparametric inference. *J. Am. Stat. Assoc.* **113**, 767–779 (2018).
62. Zeileis, A. & Hothorn, T. Diagnostic checking in regression relationships. *R News* **2**, 7–10 (2002).
63. Cameron, A. C., Gelbach, J. B. & Miller, D. L. Robust inference with multiway clustering. *J. Bus. Econ. Stat.* **29**, 238–249 (2011).
64. Zeileis, A. Econometric computing with HC and HAC covariance matrix estimators. *J. Stat. Softw.* https://doi.org/10.18637/jss.v011.i10 (2004).
65. Eckles, D., Karrer, B. & Johan, U. Design and analysis of experiments in networks: reducing bias from interference. *J. Causal Inference* https://doi.org/10.1515/jci-2015-0021 (2016).

Acknowledgements The authors thank N. Grinberg, L. Friedland and K. Joseph for earlier technical work on the development of the Twitter dataset. Earlier versions of this paper were presented at the Social Media Analysis Workshop, UC Riverside, 26 August 2022; at the Annual Meeting of the American Political Science Association, 17 September 2022; and at the Center for Social Media and Politics, NYU, 23 April 2021. Special thanks go to A. Guess for suggesting the DID analysis. D.M.J.L. acknowledges support from the William & Flora Hewlett Foundation and the Volkswagen Foundation. S.D.M. was supported by the John S. and James L. Knight Foundation through a grant to the Institute for Data, Democracy & Politics at the George Washington University.

Author contributions The order of author listed here does not indicate level of contribution. Conceptualization of theory and research design: S.D.M., D.M.J.L., D.F., K.M.E. and J.G. Data curation: S.D.M. and J.G. Methodology: D.F. Visualization: D.F. Funding acquisition: D.M.J.L. Project administration: K.M.E., S.D.M. and D.M.J.L. Writing, original draft: K.M.E. and D.M.J.L. Writing, review and editing: K.M.E., D.F., S.D.M., D.M.J.L. and J.G.

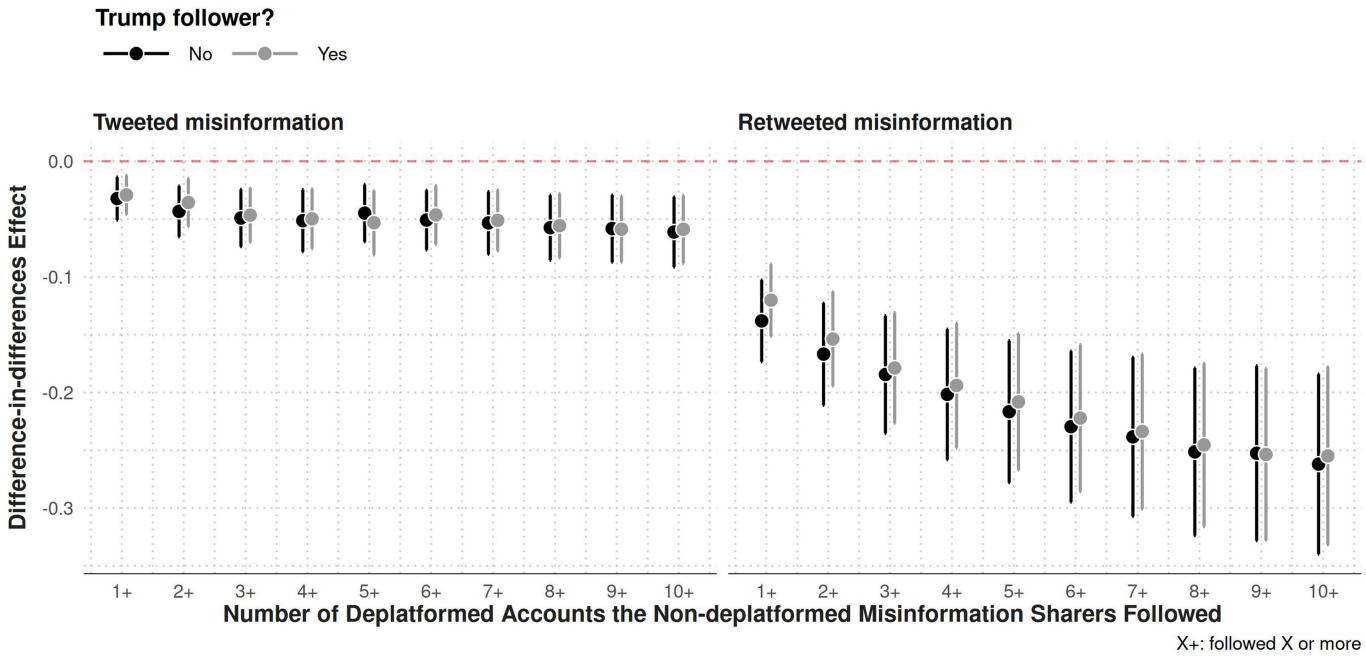
Competing interests The authors declare no competing interests.

Additional information

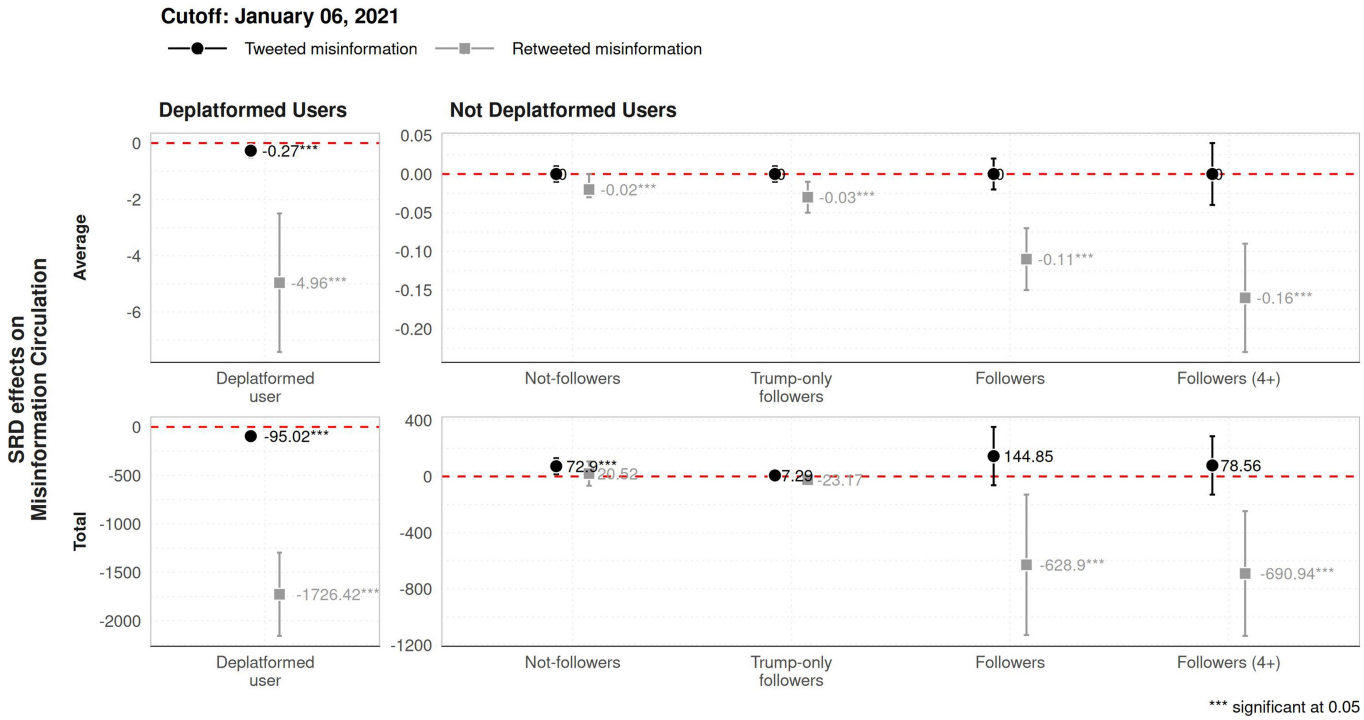
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-07524-8.

Correspondence and requests for materials should be addressed to David M. J. Lazer.

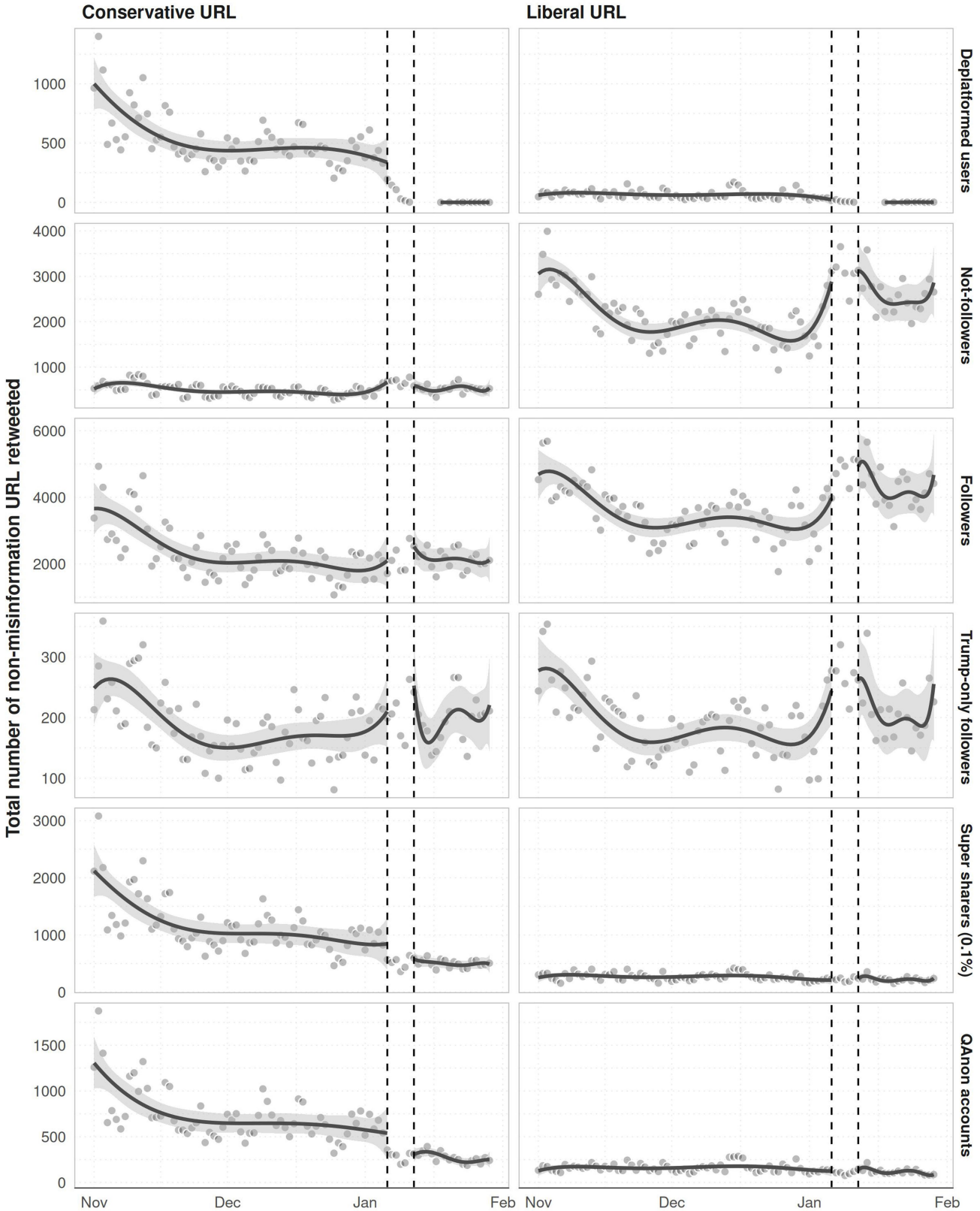
Peer review information *Nature* thanks Jason Reifler and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.



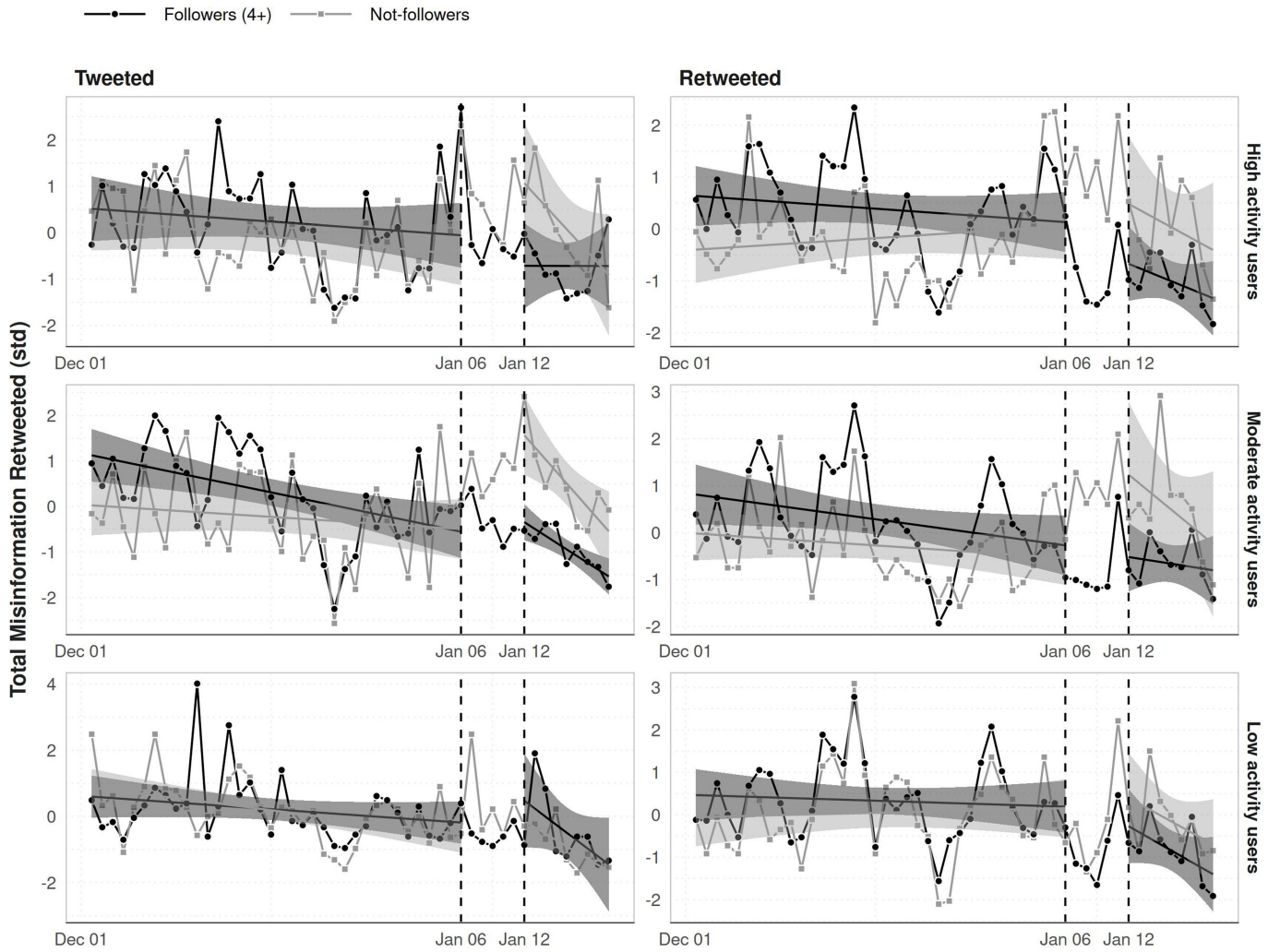
Extended Data Fig. 1 | Replication of the DID results varying the number of deplatformed accounts. DID estimates where the intervention depends on the number of deplatformed users that were followed by the not-deplatformed misinformation sharers. Results are two-way fixed effect point estimates (dots) and 95% confidence intervals (bars) of the difference-in-differences for all activity levels combined. Estimates use ordinary least squares with clustered standard errors at user-level. The Figure shows results including and excluding Trump followers (color code). The x-axis shows the minimum number of deplatformed accounts the user followed from at least one (1+) to at least ten (10+). Total sample sizes for each dosage level: Follow Trump (No): 1: 625,865; 2: 538,460; 3: 495,723; 4: 470,380; 5: 451,468; 6: 437,574; 7: 426,772; 8: 417,200; 9: 408,672; 10: 401,467; Follow Trump (Yes): 1: 688,174; 2: 570,637; 3: 514,352; 4: 481,684; 5: 460,676; 6: 444,656; 7: 432,659; 8: 421,924; 9: 413,241; 10: 405,766.



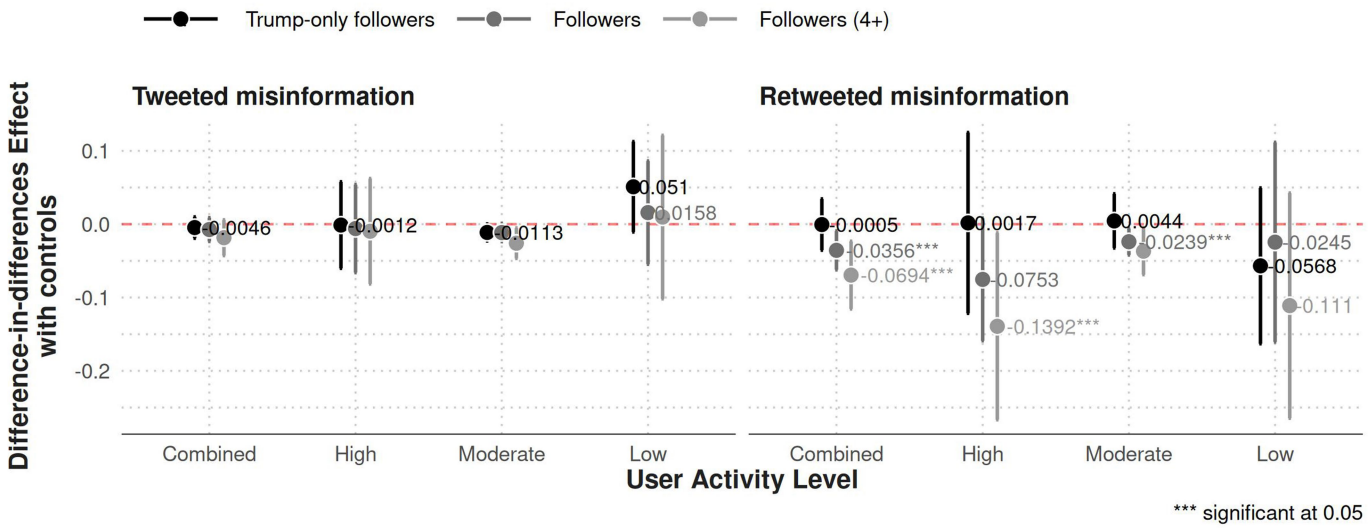
Extended Data Fig. 2 | SRD results for total (bottom row) and average (top row) misinformation tweets and retweets, for deplatformed and not-deplatformed users. Sample size includes 546 observations (days) on average across groups (x-axis), 404 before and 136 after. The effective number of observations is 64.31 days before and after on average. The estimation excludes data between Jan 6 (cutoff point) and 12 (included). January 6th is the score value 0, and January 12th the score value 1. Optimal bandwidth of 32.6 days with triangular kernel and order-one polynomial. Bars indicate 95% robust bias-corrected confidence intervals.



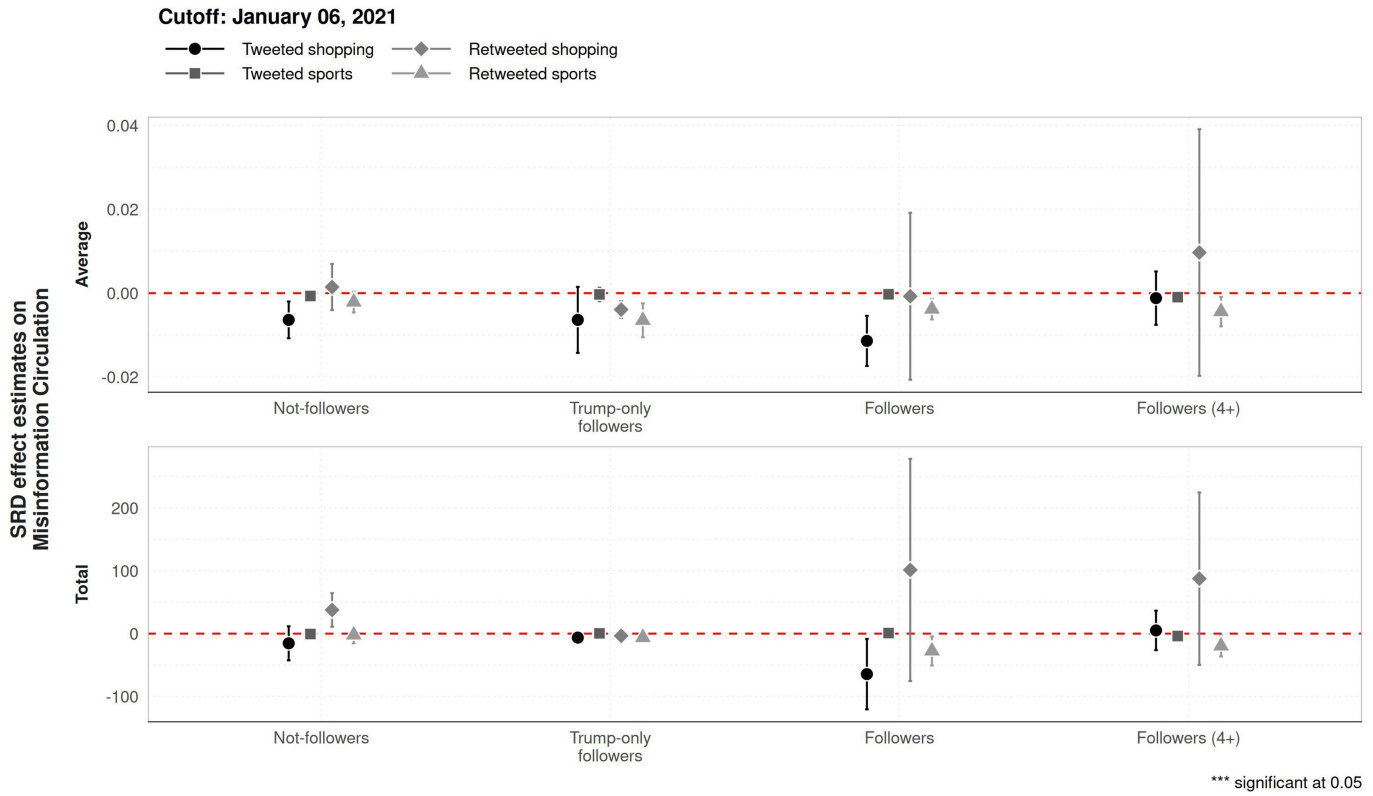
Extended Data Fig. 3 | Time series of the daily mean of non-misinformation URL sharing. Degree five polynomial regression (fitted line) before and after the deplatforming intervention, separated by subgroup (panel rows), for liberal-slant news (right column), and conservative-slant news (left column) sharing activity. Shaded area around the fitted line is the 95% confidence interval of the fitted values. As a placebo test we evaluate the effect of the intervention on sharing non-fake news for each of our subgroups. Since sharing non-misinformation does not violate Twitter's Civic Integrity policy – irrespective of the ideological slant of the news – we do not expect the intervention to have an impact on this form of Twitter engagement; see SI for how we identify liberal and conservative slant of these domains from ref. 52. Among the subgroups, users typically did not change their sharing of liberal or conservative non-fake news. Taking these results alongside those in Fig. 2 implies that these subgroups of users did not substitute non-misinformation conservative news sharing during and after the insurrection in place of misinformation.



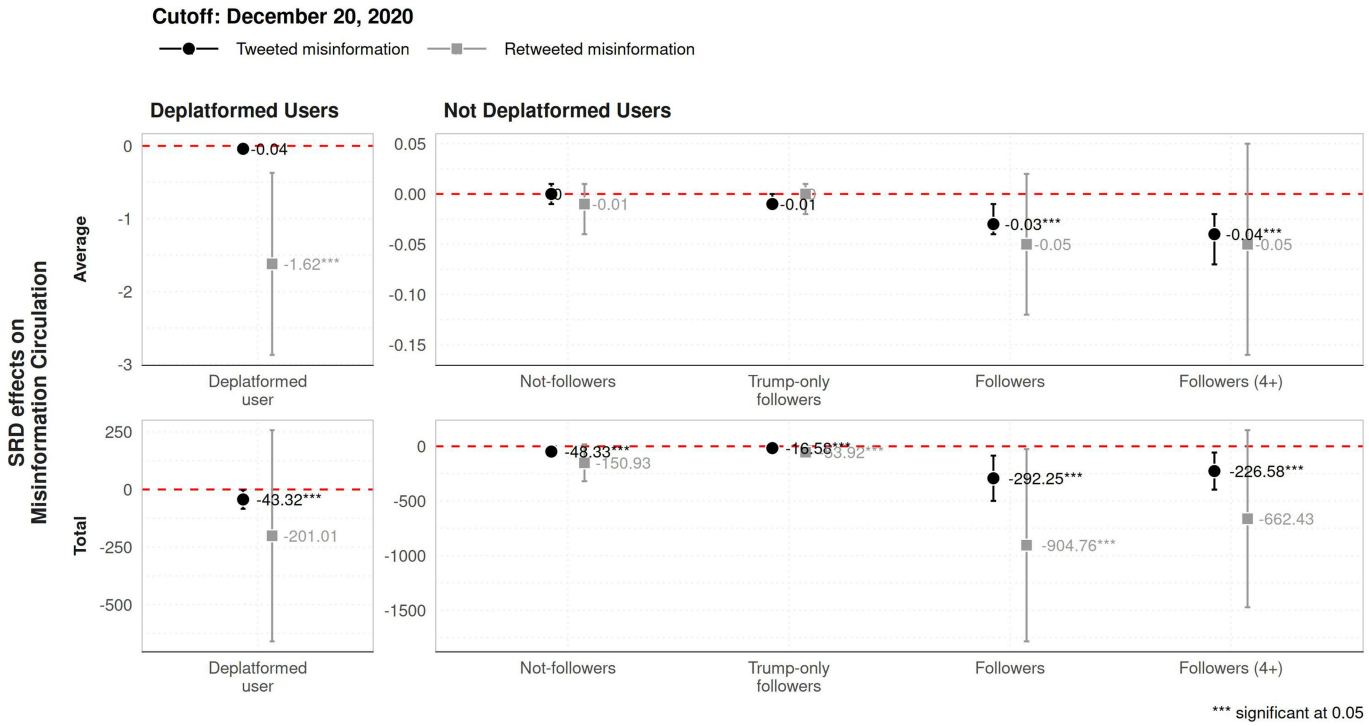
Extended Data Fig. 4 | Time series of misinformation tweets and retweets (panel columns), separately for high, medium and low activity users (panel rows). Fitted straight lines describe a linear regression fitted using ordinary least squares of daily total misinformation retweeted standardized (y-axis) on days (x-axis) before January 6th and after January 12th. Shaded areas around the fitted line are 95% confidence intervals.



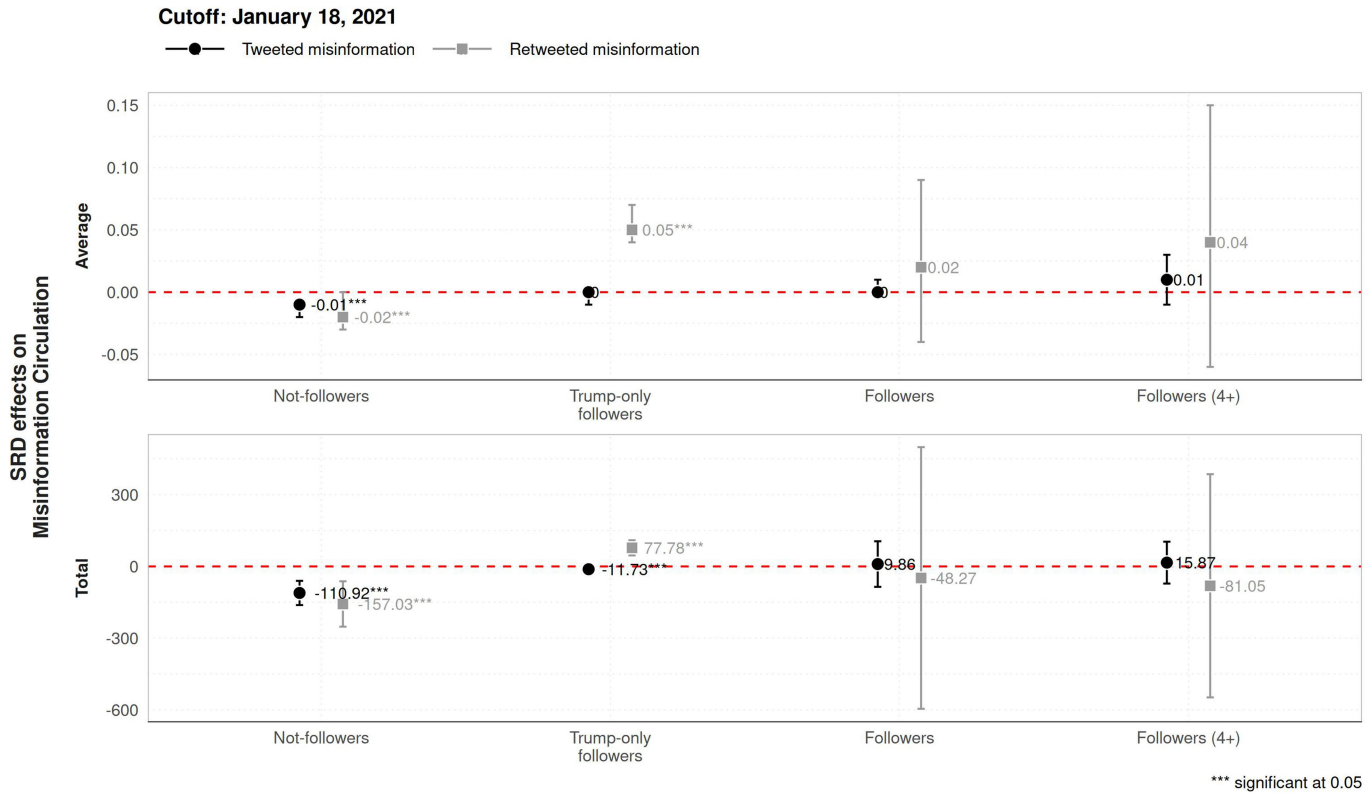
Extended Data Fig. 5 | Replicates Fig. 5 but with adjustment covariates. Corresponding regression tables are Supplementary Information Tables 1 to 3. Two-way fixed effect point estimates (dots) and 95% confidence intervals (bars) of the difference-in-differences for high, moderate, and low activity users, as well as all these levels combined (x-axis). P-values (stars) are from two-sided t-tests based on ordinary least squares estimates with clustered standard errors at user-level. Estimates compare followers (treated group) and not-followers (reference group) of deplatformed users after January 12th (post-treatment period) and before January 6th (pre-treatment period). No multiple test correction was used. See Supplementary Information Tables 1–3 for exact values with all activity level users combined. Total sample sizes of not-followers (reference) and Trump-only followers: combined: 306,089, high: 53,962, moderate: 219,375, low: 32,003; Followers: combined: 662,216, high: 156,941, moderate: 449,560, low: 53,442; Followers (4+): combined: 463,176, high: 115,264, moderate: 302,907, low: 43,218.



Extended Data Fig. 6 | Placebo test of SRD results for total (bottom row) and average (top row) shopping and sports tweets and retweets at the deplatforming intervention, among those not deplatformed. Sample size includes 545 observations (days), 404 before the intervention and 141 after. Optimal bandwidth of 843.6 days with triangular kernel and order-one polynomial. Cutoff points on January 6th (score 0) and January 12th (score 1). Bars indicate 95% robust bias-corrected confidence intervals. These are placebo tests since tweets about sports and shoppings should not be affected by the insurrection or deplatforming.



Extended Data Fig. 7 | Placebo test of SRD results for total (bottom row) and average (top row) misinformation tweets and retweets using December 20th as an arbitrary cutoff point. Sample size includes 551 observations (days), 387 before the intervention and 164 after. Optimal bandwidth of 37.2 days with triangular kernel and order-one polynomial. Bars indicate 95% robust bias-corrected confidence intervals about the SRD coefficients. This is a placebo test of the intervention period.



Extended Data Fig. 8 | Placebo test of SRD results for total (bottom row) and average (top row) misinformation tweets and retweets using January 18th as a cutoff point. The parameters are very similar to Extended Data Fig. 7.

Extended Data Table 1 | Demographics of Twitter Panel and Associated Subgroups

Group	Mean Age	% White	% Republican	% Female
All Users	39.2	85.2	30.3	50.6
Non-suspended	39.2	85.1	30.2	50.6
Suspended	48.1	92.0	55.1	56.3
QAnon	51.7	93.1	58.4	66.8
Misinfo sharers	46.8	89.1	42.3	48.6
Supersharers	57.5	93.7	69.5	61.5
Supersharers (alt)	55.2	93.8	63.4	59.9
Followers	50.3	93.0	53.4	50.5
Trump followers	42.2	90.1	47.5	45.3

The panel of active users has previously been linked to a proprietary voter file. This allows for comparison of the demographic attributes of users in our sample. The panel's demographics are broadly representative of the population of Twitter users, although it skews somewhat white and female[5]. Partisanship is derived from a modeled propensity to identify as one party or the other, not party registration; validation of this measure is presented in Hughes et al[4]. Other authors have noted that misinformation sharers tend to be older and more Republican[8]. Extended Data Table 1 presents the demographic composition of our subgroups. We note that this descriptive pattern holds here, and we expand this further to note that this tendency is present across our subgroups of interest, ranging from relatively anodyne behaviors such as Trump following up to suspension (and QAnon support among non-suspended users). Relative to non-suspended users, suspended users are significantly older (suspended mean 48.1, non-suspended mean 39.2, $t=22.8$, $p<1e-16$), significantly more white (suspended 92.0%, non-suspended 85.1%, $t=7.0$, $p<1e-11$), significantly more female (suspended 56.3%, non-suspended 50.6%, $t=29.6$, $p<1e-16$), and significantly more Republican (suspended 55.1%, non-suspended 30%, $t=20.0$, $p<1e-16$).

Extended Data Table 2 | Overrepresentation of Demographic Cells in Subgroups

Group	Republican	White Republican	White Female Republican	White Female Republican 65+
Suspended	1.82	1.83	2.46	5.06
QAnon	1.93	1.94	2.95	8.08
Misinfo sharers	1.39	1.40	1.40	3.26
Supersharers	2.29	2.30	3.19	13.94
Supersharers (alt)	2.09	2.10	2.89	11.14
Followers	1.76	1.78	1.94	4.88
Trump followers	1.57	1.58	1.52	2.09

Numbers are odds ratios for members of the select demographic being in the behavioral subgroup relative to the population as whole. Notably, these behaviors are especially concentrated among the intersections of these demographic groups. As shown in Extended Data Table 2, white Republican women over 65 were five times more likely to be deplatformed than Twitter users as a whole. The rate for this group is lower than the rate for supersharers and QAnon supporters, however.

nature portfolio

Corresponding author(s): David M.J. Lazer

Last updated by author(s): Apr 24, 2024

Reporting Summary

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

Statistics

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

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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We used the (then) Twitter API for data collection.
Data analysis	All of the analyses of the manuscript and SI (Figures, Tables, SRD, and DiD) used R version 4.3.1. I The data for analysis (merging subgroups and demographics with the user URL sharing activity) and the files with daily summaries (pct, average, and total) were prepared using Python 3.11.6. The code repository is publicly available at https://osf.io/ku8z4/

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

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

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

The analyses presented in this paper were performed on user-level aggregates, with user IDs anonymized to preserve privacy. We make this data available as part of the replication package on OSF: https://osf.io/ku8z4/. The tweet-level data, and specific user demographics, cannot be publicly shared due to privacy concerns

arising from matching data to administrative records, data use agreements, and platforms' terms of service. Our replication materials include the code used to produce the aggregate data from the tweet-level data, and the tweet-level data can be accessed after signing a data-use agreement; contact author D.L. with access requests.

Research involving human participants, their data, or biological material

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

Reporting on sex and gender	We use administrative data from voting records for reporting on sex in the Extended Data.
Reporting on race, ethnicity, or other socially relevant groupings	We use inferred race from the voter data vendor (TargetSmart) in the Extended Data.
Population characteristics	The population is US registered voters that have active accounts on Twitter (now X) whose accounts tweeted between 2014 and 2016, who are registered voters, and whose self-reported names and locations could be linked to a fall 2017 snapshot of a commercial voter file. Because of these inclusion characteristics, the sample of users in this paper is not representative of the US registered voting population.
Recruitment	We developed the sample by matching voter data to Twitter handles. The sample size is based on the number of matches we were able to make successfully. This sample has been used in various prior papers, with a careful examination of the biases that might exist in the sample as a result. We include references in the paper. There are 599,697 active users in our panel under this definition, who tweeted over 8 million URLs during the study period. The sample size was not determined by a prospective power analysis. However, the sample size is sufficient to make appropriate conclusions because the study is about the effects of deplatforming on misinformation on twitter, and generally for social science applications such as this the typical sample sizes for making inferences are on the order of 1,000 participants while we have over a half million.
Ethics oversight	Northeastern University IRB protocol #17-12-13.

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

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

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

Study description	This is a quantitative study, involving times series analysis of posting on Twitter, with an evaluation of whether patterns changed after January 6, 2021.
Research sample	The population is US registered voters that have active accounts on Twitter (now X) whose accounts tweeted between 2014 and 2016, who are registered voters, and whose self-reported names and locations could be linked to a fall 2017 snapshot of a commercial voter file. Because of these inclusion characteristics, the sample of users in this paper is not representative of the US registered voting population.
Sampling strategy	We developed the sample by matching voter data to Twitter handles. The sample size is based on the number of matches we were able to make successfully. This sample has been used in various prior papers, with a careful examination of the biases that might exist in the sample as a result. We include references in the paper. There are 599,697 active users in our panel under this definition, who tweeted over 8 million URLs during the study period. The sample size was not determined by a prospective power analysis. However, the sample size is sufficient to make appropriate conclusions because the study is about the effects of deplatforming on misinformation on twitter, and generally for social science applications such as this the typical sample sizes for making inferences are on the order of 1,000 participants while we have over a half million.
Data collection	We used the Twitter API for data collection of the Twitter data; and used the voter administrative data for demographic information (mostly contained in analyses in the SM).
Timing	This was a rolling data collection that started in 2017, and ceased when Twitter shut down their APIs in the summer of 2023.
Data exclusions	We exclude 11 users that were deplatformed during the January 2021 terms of use intervention but had their accounts restored. Beyond that, we do not exclude data.
Non-participation	NA

Randomization | NA

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Plants

Seed stocks | *Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*

Novel plant genotypes | *Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*

Authentication | *Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*