Peer Review File

Manuscript Title: Twitter Deplatforming: Limiting Misinformation after the Jan.6 Insurrection

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

This is a really important paper and topic. I think the paper gives an interesting analysis of some observational data about tweet and retweet behavior before and after the Jan 6 response by Twitter. By making some assumptions to do causal inference, the paper investigates whether the deplatforming led to a drop in misinformation links in tweets and retweets.

While none of the methodology is particularly novel or controversial, broadly similar analyses have been done before (according to the authors of the paper) on Facebook data, this particular analysis has important value to the larger community as a study of the impacts of account removal from social media companies. In my view, the motivating reason for this paper is the particular subject material and what can be gleaned from it. To this end, I think the rhetorical tack of the paper is a little uneven- early sections are too brazen about what conclusions we can draw from this analysis, while the technical sections and parts of the conclusion are more careful and appropriate.
I think this paper would benefit from a revision that more carefully considers the likely violations for the assumptions needed for causal inference and tones down the rhetoric accordingly, but that the underlying analysis can merit publication.

A.
The goal of the paper is trying to estimate the impact of Twitter's suspension of accounts on misinformation tweeting and retweeting. Two standard approaches are taken- SRD and diff-in-diff. The analysis is done by looking at accounts linked to voter registrations. In my view, this is an important and interesting analysis to do, so I think the paper has a high degree of relevance and interest. A key challenge for researchers external to a social media company is the quality of the data, but matching on voter registrations makes me believe the data is likely high quality. I thought the data exploration and presentation of the analyses was generally good, users were split in interesting ways (e.g. by number of followers in the suspended group), and the SI did a good job of exploring the data.

B.
I see no issues with the validity of the methodology given the constraints- observational data collected externally about user behavior around a national incident. However, I think the writing in the abstract and earlier sections downplays the strength of assumptions needed in this particular analysis. Of course, these types of objections can always be raised when trying to do causal inference with observational data, but I think in this case the likelihood of external causes getting tangled in is quite high.

For example, the continuity assumption seems strong to me because the Jan 6 incident and extensive news coverage around it could have caused a discontinuity in misinformation tweet/retweet behavior even in the absence of intervention on the platform. Jan 6 was a monumental news event, and all the subsequent reporting on the incident seem likely to impact user behaviors that pertain to election misinformation. Furthermore, the election was certified and the power transition was made shortly after (including in the measurement window of the paper), so misinformation surrounding the election might have naturally lost relevance around this time. In my view, the paper should be deliberate about our inability to separate these external factors from the impact of Twitter's intervention, even in the abstract and introductory sections.
In addition to the suspensions impacting user behavior directly, the paper points out that reporting on the suspensions was extensive:
"Twitter's enforcement action was reported widely in the mainstream media (34, 35) following an official announcement from Twitter (15), indicating that the deplatforming intervention targeted prominent conservative activists as well as QAnon accounts that were influential in spreading election misinformation (36)."
It's possible that this reporting also impacted user behavior, and one might argue that the reporting spread the news of the suspension in a way that helped the deterrent effect. It's possible that a suspension without so much press wouldn't have helped as much if users found out about the interventions through the press. This is of course not really an external effect, and one might be happy to roll it in with the suspension, but in the sense of the one sentence summary "Social media companies have the capacity to limit misinformation in discourse on their platforms, if they choose to do so," it's not a priori clear to me that a deplatforming event will always get this sort of news coverage/amplification in the absence of a national event like Jan 6.
I also think that the impact of algorithmic changes outside of the suspensions are underplayed. The authors mention:
"Any change in the algorithm will not matter for the DID analysis as it affects both followers and not-followers, and is accommodated by the parallel path assumption."
But the algorithmic change effects could be quite heterogeneous across the user base, and indeed were probably intended to be so. If Twitter was trying to reduce the spread of misinformation via an algorithmic change, it likely stands the users who have engaged with misinformation in the past are going to see larger effects from such a change.

The paper is more of a case study than an original approach, though to my knowledge the data and topic area are underexplored. I think the results are of immediate interest to the community and have broad interdisciplinary appeal.

C.
The paper says (around line 622)
"Aggregate data data, code, and materials used in the analysis will be publicly available upon publication "
Where this is a violation of "A condition of publication in a Nature Portfolio journal is that authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications." written here, I will leave to the editors.

While data was not released, I was able to review the aggregate information in the SI, which I found satisfying. Results cannot be reproduced without direct data access. Assuming the matching process to U.S. voter registration data was good, the resulting data should be pretty high quality.

D.
All of the error bars seem appropriate under the varying assumptions of those tests. I didn't see any statistical tests that I had a particular issue with, up to the previously stated caveats about the strength of necessary assumptions.

One area I think was underexplored was network based interference between e.g. the subgroups of users analyzed (followers of suspended accounts, Trump, etc.) Although the authors mention that the suspension can lead to a contagion effect, I didn't see any discussion about how the mechanism can interfere across these subgroups- there is certainly going to be some overlap in tweet/retweeted content between the different groups we're splitting on, causing interference. See e.g. https://arxiv.org/abs/1404.7530

E.
I broadly find the conclusions reasonable once the caveats that this analysis cannot separate effects of Twitter's actions from the effects of the reporting, coverage, etc. of both the Jan 6 incident and coverage of the deplatforming event itself.

F.

For me, the most important area of improvement is aligning the strength of claims in the abstract/intro with what we can actually know from the data. I think this sentence from the "Data and methods" section:
"Since the insurrection event essentially coincided with the terms of use intervention, the SRD bundles these two events, the insurrection itself with the deplatforming intervention, and so this test does not disentangle the separate effect of the deplatforming on the deplatformed users unless one were to assume that the insurrection would have had no effect on deplatformed users' misinformation sharing behavior."

Needs to be featured sooner in the paper. Also appropriate would be
"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 impact of the insurrection from Twitter's enforcement."
especially with
"we evaluate the direct causal effect of this intervention."
On a smaller note, a few typos:
who retweeted misinformation an average total of 162 after the intervention, which is a 5% increase. => 162 times?
Data availability: Aggregate data data, => Aggregate data

G.

This sentence: “Tech companies’ policies regarding permissible speech, ranging from what they allow to be posted, to what they allow to be seen, to who is allowed to post, loom large as the de facto laws governing political communication through much of the world (1).”
Appears to be citing a Washington Post article about the impacts of algorithmic ranking on users and some discussion of changing section 230 or how, e.g., China/Europe/the U.S. etc. regulate social media differently. I think the claim that a tech company’s policy entails setting “de facto laws governing political communication through much of the world” is well beyond the citation.
Some discussion and references about the impact of network interference on these methods would improve the paper.

H.
The abstract is clear and accessible, but the abstract and introduction overstates what the data shows because it does not include disclaimers about the external impacts of the insurrection and reporting on it as potential confounders. In my view, the discussion/conclusion section is more measured, though I still think sentences like “We demonstrate that Twitter’s deplatforming of the worst offenders of its Civic Integrity policy substantially directly reduced the amount of misinformation circulating on its platform.” are incompatible with the later “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 impact of the insurrection from Twitter’s enforcement.” While making this emphasis late in the conclusion is helpful, the earlier sentence I think needs to be written in a more passive way- the amount of misinformation was reduced after the deplatforming but we can’t say without including language about the causal assumptions that the reduction was directly from the suspensions.

Referee #2 (Remarks to the Author):

This is an exceptionally high quality manuscript that addresses critically important questions central to the interplay between democracy and information technology systems as well as further advancing the nascent subfield of computational social science. I overwhelmingly support publication in a high profile outlet like Nature.

Below are a number of suggestions for the authors (and editor) to consider to potentially improve an already outstanding manuscript. I know a lot of these are pedantic points about communicating subtle points -- please read the nature of these comments are complete buy-in about the the importance and quality of the research, and good faith suggestions how to more effectively communicate core findings.

Page (P) 4/Line (L)150. Can these two panels be combined into a single figure? (At an absolute minimum, I'd like the y-axis to be scaled the same.) Also, I find the language here very opaque -- "percentage of misinformation shared." What exactly is the numerator and denominator in this percentage? Is it the percentage of tweets that contain a misinformation link? (I think this is it -- the universe of tweets at issue are those with links.) As the authors well know, "sharing" can be two different things -- including content of any type (like a URL) or specifically reposting another person's post/tweet. I encourage consistent language throughout that separates these two different actions, and clearly signal that to readers.

Returning to the quantity in Fig 1, I think the quantity is the percentage of tweets with a URL where that link ultimately points to a low-credibility website (based on domain definitions from Grinberg et al and/or NewsGuard). I see this figure as a stage setting figure driving to help drive home the normative importance (I was convinced already). But, as someone who tries to teach these topics to undergraduates with limited technical skills, I have found the years that anything I struggle to put into words exactly what the quantity is will be even more difficult for students to understand. I think combining into a single figure might allow discussion -- the quantity is still significantly higher in early 2021 (versus 2017) even after Twitter's action. I think evaluating Twitter's policy decisions requires considering not only the magnitude of the *change* in key quantities on either side of Jan 6, but where the broader information ecosystem is post Jan 7, 2021 compared to all of 2016/17.

P7/L256. " Among the users who were not deplatformed, the followers of
the deplatformed users (including Trump followers) are 26.63 percent." I'm having trouble parsing this sentence. I think this means that approximately one quarter of the sample followed one or more deplatformed accounts.

P7/L264. Thinking again about communicating this in my teaching, would a a line graph (or series of line graphs) be more intuitive/effective here, going from pre-deplatformed levels to post-deplatfomring levels? (The post level for deplatformed accounts would, unless I'm very confused, be zero, so the change quantity for deplatformed uses is really telling us about their levels of posting content with links to misinformation domains.) I can think of a number of potential way this information.

Also, I think "deplatformed followers" is potentially confusing in the second panel. I know it is not as snappy, but is "followers of deplatformed accounts" more precise? I would encourage avoid using the term "deplatformed" in labels that sometimes refers to people who were actually deplatformed, and sometimes to people who were not deplatformed but were following deplatformed accounts (which is how I think the term is used across the two panels in Fig 2). Trying to explain this word means slightly different things in the same Figure to undergrads feels like 5 minutes of extra class time. (This issue comes up across a number of figures, for instance in Fig 4.)

P9/L303. I am primarily a survey and survey experiments researcher, and am always playing catch up on the latest advances (and weaknesses) in DiD approaches. I would greatly appreciate a section (SM is fine) explaining where the DiD approaches used here fit within the broader concerns about the reliability and appropriateness of DiD approaches.

My basic conclusion is that this is important and meaningful research that should be published in a high level outlet. But, I think there are some minor issues around how to most effectively communicate this important points.

Referee #3 (Remarks to the Author):

Summary of the key results

The paper analyzes the impact of platform moderation policies in the aftermath of the US Capitol building's attack. It specifically analyses the effects of Twitter's moderation policy on the spread of misinformation on the platform. It also assesses the effects of the platform's policy decisions on the behavior of the followers of deplatformed users in further engaging with misinformation (versus non-followers of deplatformed). The authors found that platform moderation policies and suspension of users reduced misinformation circulating on Twitter and limited the amplification of the disinformation by followers of suspended users as well. Overall, the paper shows how platform moderation policies directly and indirectly impacts misinformation circulating on social media – directly by targeting users spreading false information and indirectly by limiting the spreading of this content by followers of suspended users.

Originality and significance

Original and significant. I specially value the effort of the authors in providing novel evidence to understand the magnitude of the effects of platform intervention on public discourse on social media. In doing that, the authors provide a background against which to assess other types of interventions aimed at ensuring the quality of public discourse e.g., EU policy regulation.

The question tackled in this study is of extreme importance not only because of Twitter's relevance as one of the main mediators of public discourse during the Capitol's attack, but also because of the relevance of existent and emerging platforms in future political mobilization and coordination actions. That said, my comments below are aimed to clarify the article and help the authors to improve an already relevant work.

Appropriate use of statistics

Statistical analyses seem adequate, and the authors have been very careful in reporting all the statistical tests details. The multiple robustness checks show the rigor with which the study was conducted.

Data & methodology

For their analysis the authors rely on NewsGuard's index, and they combine this data with data from Grinberg et al. 2019. Based on this approach they identify misinformation providers and compile the list of domains to be detected on the extracted URLs. Additionally, as robustness checks, they perform separate analyses for each of these lists and an additional analysis including the Indiana University's Iffy list. I applaud the effort of the authors in combining these different sources of information for the detection of misinformation providers. My question here though is do the authors concentrate exclusively on misinformation providers (sites providing inaccurate, misleading or false information)? Did I misunderstand that the study excludes disinformation providers (sites mimicking news outlets and purposely providing false information). Do the authors use the term misinformation to include both mis- and dis- information? The SI does not help to clarify the conceptual boundaries of the concept analyzed. The SI describes the identification of fake-news sites more generally. The use of the fake news label there, contrasts with the concept of misinformation used throughout the manuscript. It is true the authors refer to Grinberg et al. 2019 for further information on their approach. In turn, Grinberg et al. 2019 follow Lazer et al., which argue that fake news overlaps with other information disorders, such us misinformation (false or misleading

information) and disinformation (false information that is purposely spread to deceive). However, there is room for clarification in this regard. The authors use the threshold of 60 points for the selection of the NewsGuard sites for their analysis. One wonders which type of information providers are included after applying that threshold? Are information providers included in the study all providing content fabricated with the purpose to deceive? Are some of them falling short on standard journalistic routines and therefore, their content lacks the quality of professional journalistic sites? This differentiation is much clearer on Grinberg et al. 2019 when they distinguish between black, red and orange websites (or providers of fabricated stories and questionable claims respectively). The distinction is important and necessary if we aim to fully understand the quality of information available for public discussion and the impact of platform moderation policies in ensuring it.

My main concerns is related though to the causal continuity assumption that the misinformation sharing would have occurred at the same rate had the intervention not occurred. I found the authors fall short to convince us this would have been the case and to discuss evidence to support their assumption. Several events concurred on that day and the following ones that could have influenced the behavior of misinformation providers and sharers on Twitter. One can think about the multiple reminders to abide by the law send by various sources from the judiciary, the executive, or the fact that the eldest daughter of the President deleted a tweet herself. Consistently, the authors should provide a lengthier discussion in the main manuscript to support their argument. Only towards the end and the discussion (lines #274 and #431), they acknowledge and emphasize that “their design does not prove the intervention is causal, in that there is no way for us to fully separate the impact of the insurrection from Twitter’s enforcement”. However, the authors are indeed interested in causality and the wording throughout the manuscript proves it: “we evaluate the direct causal effect”; “the contagion causal effect of this intervention”; “we also show a deterrent effect”. If the current research design i.e., the SRD design, does not allow to disentangle the effect of the intervention from the effect of the insurrections itself as they eventually acknowledge, a careful consideration of the wording should be in place.

The authors aim to understand if the “dose” of misinformation received has an impact on followers of deplatformed users. Towards this end, they define a follower as someone following at least four deplatformed users. How the results change if this threshold varies? The threshold seems somehow arbitrary, and one wonders if different thresholds would yield different and more relevant results?

In the discussion section, there is room for improvement when it comes to define what kind of studies are needed to overcome limitations of the study as well as what kind of data should be necessary for future replications of it. Even if those replications are currently not possible due restrictions to proprietary social media data, new regulations to grant data access elsewhere i.e., EU, might overcome those limitations in the near future. The authors insights on this regard would be highly valuable and make the discussion section stronger.

Perhaps irrelevant for some readers, but still worthy to ask in my opinion is that in the introduction, the authors stated (line #59) “Twitter deplatformed a large number of right-wing extremists, including Trump himself”, should we assume from this that the US president is classified as a right-

wing extremist here?

To provide a broader context to the enforcement of Twitter's moderation policy, the authors might want to acknowledge previous deplatforming processes, at various scales, which have taken place before, both in western and non-western countries (see for instance the case of Myanmar, after legacy media raised attention of the use of Facebook to coordinate attacks against Rohingya minorities; or the suspension of multiple accounts following the outbreak of the Covid-19 pandemic, when social media platforms argued to have put in place unprecedented measures to tackle incivility online:

https://blog.twitter.com/en_us/topics/company/2020/information-operations-june-2020.html
https://transparency.twitter.com/
https://www.nytimes.com/2020/06/11/technology/twitter-chinese-misinformation.html
https://about.fb.com/news/2020/12/coronavirus/

The design problem is alluded as one of the two potential reasons for the relatively small number of studies examining platform regulation. But the limited access to comprehensive datasets should be pointed out as well.

The authors currently argue that their results demonstrate the capacity of social media platforms to regulate public discourse. The assumption here is that their study shows the effects of Twitter's decision over public opinion. Yet, the results cannot inform how or when the ideas spread by right-wing extremists shaped people's opinion in or outside Twitter. We do get useful insights limited though to the extent to which misinformation was reduced and the effects of suspension of users over followers sharing behavior.

The authors argue (line #96) that "U.S. tech companies have a First Amendment right to determine what content their platforms amplify and suppress, and who is permitted to post content […]". Here, I question if this is strictly the case. Laws regulating content platforms can carry and who can access the platforms are frequently held to be violating the First Amendment, but isn't this because they see these laws as violating the rights of the social media users not that of the tech companies themselves?

Conclusions
As mentioned, I found the discussion section to fall short in highlighting the limitations of the research design and the analytical approach. If space is a constrain for the authors, I believe they can trim details on the decision-making process surrounding the suspension of the accounts. A good example of this being "the then-CEO of Twitter, Jack Dorsey, who was vacationing at the time in the South Pacific". The need for further elaboration on the above mentioned points should prevail in front of the perhaps too generous details in the introduction of the current manuscript.

Minor
Edit for consistency (line #63) June 30th, 2020 and January 31, 2021 to June 30, 2020

Author Rebuttals to Initial Comments:

Referee #1

I think the rhetorical tack of the paper is a little uneven- early sections are too brazen about what conclusions we can draw from this analysis, while the technical sections and parts of the conclusion are more careful and appropriate. I think this paper would benefit from a revision that more carefully considers the likely violations for the assumptions needed for causal inference and tones down the rhetoric accordingly, but that the underlying analysis can merit publication.

Thank you very much for these comments. As we explain in our response to the editor above, we have extensively revised the paper to make it clear that our SRD analysis is potentially confounded by the real-world events of January 6, and we soften claims about identifying causal effects and to underscore the strength of the assumptions. We have made extensive revisions in the abstract, introduction, the data and methods section, and the discussion and conclusion. This more careful wording to match the nature of the observational data and the natural experimental designs greatly improves the paper and we are very thankful for this feedback on our writing.

At the same time, as we explain above in our note to the editors, while we agree that the continuity assumption in the SRD is strong, we also believe that other assumptions we make are more reasonable, such as the parallel path assumption for the DID analysis that we can use data to warrant. Relatively weak assumptions support a causal claim for the DID because the followers and not-followers were both exposed to the events of January 6, and so the not-followers serve as a control for that analysis. And while the strong continuity assumption is needed to quantify the magnitude in the reduction of misinformation circulation among deplatformed users in the SRD, a causal claim that the reduction was greater than zero only depends on a weak assumption that the deplatformed users would have continued sharing some amount of misinformation had the deplatforming not occurred, which is a very reasonable assumption. We hope our revisions reflect the nuance that our observational data can reasonably underwrite some claims about causality, but not all, and that any causal claims remain tentative.

To be specific, here are some examples of the edits that we have made. We have completely rewritten the abstract and we now write "Our results indicate that Twitter's intervention reduced misinformation circulation among those users who followed the deplatformed users, in addition to among the deplatformed users themselves, though we cannot identify the magnitude of the causal estimates due to the co-occurrence of the deplatforming intervention with the events surrounding January 6th." In the introduction, we added "Our core analytical challenge is that Twitter's deplatforming intervention coincided with the insurrection and certification along with massive media reporting on each of these events, and hence there are inherent confounds in each of our comparisons. Since we use observational data, the effects we identify can be interpreted as causal under assumptions that we state and, to the extent possible, warrant below." In the research design section, we now write, "This assumption [continuity] is strong because Twitter's 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 separate 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 behavior, which would be a strong and likely implausible assumption."

We have included a marked up version of the revised manuscript and SI that shows each of our revisions, which are extensive. We feel the manuscript is much improved from your feedback on our writing and we hope that we have struck a correct balance in our interpretations.

I think the writing in the abstract and earlier sections downplays the strength of assumptions needed in this particular analysis. Of course, these types of objections can always be raised when trying to do causal inference with observational data, but I think in this case the likelihood of external causes getting tangled in is quite high….

For example, the continuity assumption seems strong to me because the Jan 6 incident and extensive news coverage around it could have caused a discontinuity in misinformation tweet/retweet behavior even in the absence of intervention on the platform. Jan 6 was a monumental news event, and all the subsequent reporting on the incident seems likely to impact user behaviors that pertain to election misinformation. Furthermore, the election was certified and the power transition was made shortly after (including in the measurement window of the paper), so misinformation surrounding the election might have naturally lost relevance around this time. In my view, the paper should be deliberate about our inability to separate these external factors from the impact of Twitter's intervention, even in the abstract and introductory sections.

As we mention above, we have made extensive revisions to highlight the strength of the continuity assumption within the SRD and the likelihood that the deplatforming intervention was confounded by the real-world events that occurred at the same time. We also have added the certification of the election as another potential confounder in addition to the insurrection.

In addition to the suspensions impacting user behavior directly, the paper points out that reporting on the suspensions was extensive…. It's possible that this reporting also impacted user behavior, and one might argue that the reporting spread the news of the suspension in a way that helped the deterrent effect. It's possible that a suspension without so much press wouldn't have helped as much if users found out about the interventions through the press. This is of course not really an external effect, and one might be happy to roll it in with the suspension, but in the sense of the one sentence summary "Social media companies have the capacity to limit misinformation in discourse on their platforms, if they choose to do so," it's not a priori clear to me that a deplatforming event will always get this sort of news coverage/amplification in the absence of a national event like Jan 6.

This is a very helpful point that had not occurred to us when writing the original version. We have added sentences in the introduction and at the end of the research design section that the news reporting likely amplified any effect of the deplatforming, and that future interventions might not be as effective if not bundled with widespread reporting. We also have added a sentence in the conclusion to this effect. Finally, we have deleted the one-sentence summary that we had included on the previous version as that sentence was overstated and potentially misleading.

I also think that the impact of algorithmic changes outside of the suspensions are underplayed. The authors mention: "Any change in the algorithm will not matter for the DID analysis as it affects both followers and not-followers, and is accommodated by the parallel path assumption." But the algorithmic change effects could be quite heterogeneous across the user base, and indeed were probably intended to be so. If Twitter was trying to reduce the spread of misinformation via an algorithmic change, it likely stands that the users who have engaged with misinformation in the past are going to see larger effects from such a change.

This is another very helpful point. This is correct – we do not know that the algorithm affected the followers and not-followers equally. We have deleted the sentence where we had said the algorithm affects both groups equally in the DID, and instead on p. 6 we now write "Hence this algorithm change is part of the intervention that we evaluate in both the SRD and DID."

The paper says (around line 622) "Aggregate data, code, and materials used in the analysis will be publicly available upon publication." Whether this is a violation of "A condition of publication in a Nature Portfolio journal is that authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications." written here, I will leave to the editors. While data was not released, I was able to review the aggregate information in the SI, which I found satisfying. Results cannot be reproduced without direct data access. Assuming the matching process to U.S. voter registration data was good, the resulting data should be pretty high quality.

Our research team strongly shares these commitments to reproducible science. If this paper is accepted, we will work closely with the editors and the *Nature* staff to meet our data and code sharing requirements, given the limitations on what we can share under the Twitter and TargetSmart terms of use that govern the data. Here we can point to a few examples to illustrate what we have previously made available. Grinberg et al. (Science, 2019) made code and aggregate data available, with individual-level data available under a data-use agreement. Green et al. (PNAS, 2022) made all code, as well as high-level aggregate data for reproducing the models available, with the input data available under a data-use agreement. Hughes et al. (Public Opinion Quarterly, 2021) only released replication code; however, that included the code underpinning the record linkage. What we currently write in the data availability statement for this paper most closely matches the Grinberg example.

One area I think was underexplored was network based interference between e.g. the subgroups of users analyzed (followers of suspended accounts, Trump, etc.) Although the authors mention that the suspension can lead to a contagion effect, I didn't see any discussion about how the mechanism can interfere across these subgroups. There is

certainly going to be some overlap in tweet/retweeted content between the different groups we're splitting on, causing interference. See e.g. https://arxiv.org/abs/1404.7530

We find this also to be an extremely helpful suggestion. We are indeed (partially) modeling dependence in the social network by studying whether deplatforming impacted the followers of the deplatformed users, and we make use of weakened SUTVA assumptions very much like those in the Eckles et. al (2014) paper that you mention here (in the arXiv URL). We have added a discussion of SUTVA in a new section in the SI entitled "Network interference and our SUTVA assumptions," where we include a discussion of SUTVA in the SRD and DID designs. That section provides our technical details, but in essence, in the DID we model dependence in misinformation circulation for those users that had followed the deplatformed accounts. In the DID, we make use of a weakened SUTVA assumption that the deplatformed users are exchangeable but not ignorable, and the vector of treatment assignments of the remaining units in the network is ignorable (citing the Eckles et al paper), and this simplifying assumption enables us to model spillover in the network in a way that is tractable. The new section in the SI explains our logic in more technical detail. We also have placed a sentence pointing to the SI SUTVA section at the place where we discuss the DID results on p. 11.

For me, the most important area of improvement is aligning the strength of claims in the abstract/intro with what we can actually know from the data. I think this sentence from the "Data and methods" section: "Since the insurrection event essentially coincided with the terms of use intervention, the SRD bundles these two events, the insurrection itself with the deplatforming intervention, and so this test does not disentangle the separate effect of the deplatforming on the deplatformed users unless one were to assume that the insurrection would have had no effect on deplatformed users' misinformation sharing behavior." – needs to be featured sooner in the paper. Also appropriate would be "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 impact of the insurrection from Twitter's enforcement." especially with "we evaluate the direct causal effect of this intervention."

As we mention in our response to your first comment, we have extensively edited the abstract and introduction to soften the wording of having identified causal effects from our natural experimental designs. We did not intend to overclaim the findings of our natural experimental designs, and it is very helpful feedback for us to temper the language. For example, in the manuscript we no longer make reference to "direct causal" effects.

On a smaller note, a few typos: "who retweeted misinformation an average total of 162 after the intervention, which is a 5% increase." => 162 times?

Yes, that is correct, and we have made this correction.

Data availability: "Aggregate data data," => Aggregate data

We have made this correction as well.

This sentence: "Tech companies' policies regarding permissible speech, ranging from what they allow to be posted, to what they allow to be seen, to who is allowed to post, loom large as the de facto laws governing political communication through much of the world (1)." – appears to be citing a Washington Post article about the impacts of algorithmic ranking on users and some discussion of changing section 230 or how, e.g., China/Europe/the U.S. etc. regulate social media differently. I think the claim that a tech company's policy entails setting "de facto laws governing political communication through much of the world" is well beyond the citation.

We deleted this cite to the Washington Post article and replaced it with a cite to D. Lazer, The rise of the social algorithm. Science. 348, 1090–1091 (2015), which is more squarely on point.

Some discussion and references about the impact of network interference on these methods would improve the paper.

As we mention above, we have added a new section to the SI that discusses network interference and the SUTVA assumptions, and we have added cites to the Eckles et al (2014) paper you mention and also to Gerber and Green which has a textbook treatment of SUTVA.

The abstract is clear and accessible, but the abstract and introduction overstates what the data shows because it does not include disclaimers about the external impacts of the insurrection and reporting on it as potential confounders. In my view, the discussion/conclusion section is more measured, though I still think sentences like "We demonstrate that Twitter's deplatforming of the worst offenders of its Civic Integrity policy substantially directly reduced the amount of misinformation circulating on its platform." are incompatible with the later "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 impact of the insurrection from Twitter's enforcement." While making this emphasis late in the conclusion is helpful, the earlier sentence I think needs to be written in a more passive way- the amount of misinformation was reduced after the deplatforming but we can't say without including language about the causal assumptions that the reduction was directly from the suspensions.

We agree – the first sentence you highlighted, that says "We demonstrate…," was stated too forcefully. We deleted it since it was not necessary for the point we were making in that

paragraph and as we mention, we have made extensive revisions throughout the manuscript to address this concern. Thank you again for leading us to improve the paper in this way.

Referee #2

Page (P) 4/Line (L)150. Can these two panels be combined into a single figure? (At an absolute minimum, I'd like the y-axis to be scaled the same.)

This is a very helpful suggestion. In the original version of Figure 1, we used different scales for the Y-axis in each panel, which made the pattern in each time series more apparent and visible, but you are correct that in effect this masked the comparisons in the level of misinformation circulation across the two election cycles. We experimented with combining the two time series into one figure instead of across two panels, but we found the combined figure is hard to interpret since the vertical line for election day is on a different date in each cycle, and so adding in multiple vertical lines made the figure busy and hard to interpret. Instead, we have kept the time series for each election cycle in a different panel, but made the y-axes the same, and this much better highlights the larger volume of misinformation in circulation in 2020 compared to 2016. We hope that this change communicates the differences across the years adequately.

Also, I find the language here very opaque -- "percentage of misinformation shared." What exactly is the numerator and denominator in this percentage? Is it the percentage of tweets that contain a misinformation link? (I think this is it -- the universe of tweets at issue are those with links.)

Thank you for pointing out this ambiguity. We have revised the caption of this figure to be clear about the numerator and denominator for the points in that time series. As we explain above in a response to the editor, we also have modified Figure 1 to include two time series plots in each panel, one where the denominator is the set of all panel members who share any URL, and one where the denominator is the set of panel members who shared a misinformation URL. In the caption for Figure 1 we now write, “All users series: daily percentage of posts shared (i.e., tweeted or retweeted) containing a misinformation URL among all tweets and retweets containing a URL. Misinformation users series: same as the all users series except limited to the set of users who shared at least one misinformation URL during the 2020 election cycle.”

As the authors well know, "sharing" can be two different things -- including content of any type (like a URL) or specifically reposting another person's post/tweet. I encourage consistent language throughout that separates these two different actions, and clearly signal that to readers.

This also was a problem in our caption for Figure 1. As we show in our previous response, we edited the caption to make it clearer now that we use "sharing" to capture both tweeting and retweeting, which we combine to do the time series for this figure. In the analyses to follow we separate tweets from retweets and so do not use the term "sharing" elsewhere.

Returning to the quantity in Fig 1, I think the quantity is the percentage of tweets with a URL where that link ultimately points to a low-credibility website (based on domain definitions from Grinberg et al and/or NewsGuard). I see this figure as a stage setting figure to help drive home the normative importance (I was convinced already). But, as someone who tries to teach these topics to undergraduates with limited technical skills, I have found over the years that anything I struggle to put into words exactly what the quantity is will be even more difficult for students to understand. I think combining the panels into a single figure might allow discussion -- the quantity is still significantly higher in early 2021 (versus 2017) even after Twitter's action. I think evaluating Twitter's policy decisions requires considering not only the magnitude of the *change* in key quantities on either side of Jan 6, but where the broader information ecosystem is post Jan 7, 2021 compared to all of 2016/17.

We agree very much that this is a powerful way to think about science communication, that our language and visualizations should readily make sense to undergraduates or even to the well-informed public. As we mention above, we have fixed this figure so that the y-axes on both panels are the same scale, and as a result, the differences in the volume of misinformation circulating during each of the two election cycles is much more apparent. Further, we have added a second time series to each panel, where in addition to the rate of misinformation sharing among our whole sample, we also show the rate of misinformation sharing among the users who have a history of sharing misinformation. While adding this second time series to each panel adds some complexity, we think the comparisons are interesting and could spark conversations among students – and possibly an opportunity for students to think more deeply about numerators and denominators in quantitative reasoning. As we explain above in the note to the editor, the added time series also helps to make it more transparent what our estimation sample is for the subsequent analyses. That said, explaining why the information ecosystem in 2020 is so much worse than 2016 is outside of the scope of this paper, and so we mention the comparison between the two years and that the misinformation ecosystem seems to have changed over this time, but otherwise leave the comparisons implicit – although this comparison too might spark discussion among students and other readers regarding the underlying causes.

P7/L256. "Among the users who were not deplatformed, the followers of the deplatformed users (including Trump followers) are 26.63 percent." I'm having trouble parsing this sentence. I think this means that approximately one quarter of the sample followed one or more deplatformed accounts.

This is helpful – thank you for catching this typo. We have revised this sentence to: "Among the users who were not deplatformed, 26.4 percent were followers of at least one of the deplatformed accounts (including those who followed Trump)."

P7/L264. Thinking again about communicating this in my teaching, would a line graph (or series of line graphs) be more intuitive/effective here, going from pre-deplatformed levels to post-deplatfomring levels? (The post level for deplatformed accounts would, unless I'm very confused, be zero, so the change quantity for deplatformed users is really telling us about their levels of posting content with links to misinformation domains.)

We very much agree that the series of line graphs is a much more intuitive way to visualize the SRD analysis than the visualization of the parameter estimates we previously had for Figure 2. We moved the original figure with the detailed parameter plots to the SI (now Figure 3a), and we have replaced it with a new Figure 2 that has the line graphs.

Also, I think "deplatformed followers" is potentially confusing in the second panel. I know it is not as snappy, but is "followers of deplatformed accounts" more precise? I would encourage avoiding use of the term "deplatformed" in labels that sometimes refers to people who were actually deplatformed, and sometimes to people who were not deplatformed but were following deplatformed accounts (which is how I think the term is used across the two panels in Fig 2). Trying to explain this word means slightly different things in the same Figure to undergrads feels like 5 minutes of extra class time. (This issue comes up across a number of figures, for instance in Fig 4.)

This is helpful feedback. In the text and all of the figures (in both the main text and the SI) we now only use "followers" to indicate followers of deplatformed accounts, and no longer use "deplatformed followers." We define "followers" and "not-followers" on p. 6 when we describe the setup for the DID analysis, and the idea that "followers" are those that were not deplatformed but followed the deplatformed accounts should now be apparent.

P9/L303. I am primarily a survey and survey experiments researcher, and am always playing catch up on the latest advances (and weaknesses) in DiD approaches. I would greatly appreciate a section (SM is fine) explaining where the DiD approaches used here fit within the broader concerns about the reliability and appropriateness of DiD approaches.

This is a helpful suggestion. We have added a discussion of the reliability and appropriateness of DID for our study in the SI section where we describe the DID methods.

My basic conclusion is that this is important and meaningful research that should be published in a high level outlet. But, I think there are some minor issues around how to most effectively communicate these important points.

Again, we strongly agree that clear communication is of core importance for science, and we hope that our revisions, especially the new figures, are an improvement.

Referee #3

For their analysis the authors rely on NewsGuard's index, and they combine this data with data from Grinberg et al. 2019. Based on this approach they identify misinformation providers and compile the list of domains to be detected on the extracted URLs. Additionally, as robustness checks, they perform separate analyses for each of these lists and an additional analysis including the Indiana University's Iffy list. I applaud the effort of the authors in combining these different sources of information for the detection of misinformation providers. My question here though is do the authors concentrate exclusively on misinformation providers (sites providing inaccurate, misleading or false information)? Did I misunderstand that the study excludes disinformation providers (sites mimicking news outlets and purposely providing false information). Do the authors use the term misinformation to include both mis- and dis- information? The SI does not help to clarify the conceptual boundaries of the concept analyzed. The SI describes the identification of fake-news sites more generally. The use of the fake news label there, contrasts with the concept of misinformation used throughout the manuscript. It is true the authors refer to Grinberg et al. 2019 for further information on their approach. In turn, Grinberg et al. 2019 follow Lazer et al., which argue that fake news overlaps with other information disorders, such as misinformation (false or misleading information) and disinformation (false information that is purposely spread to deceive). However, there is room for clarification in this regard.

Thank you for pointing out this ambiguity. We have revised the main text in the first paragraph of the results section, on p. 7, to state that our measurement does not distinguish between misinformation and disinformation domains, where we write "Our primary outcome is the daily counts of tweets and retweets, separately, containing URLs from a domain that appears on the list of misinformation and disinformation domains maintained by Newsguard combined with the list in (4). 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 URL contents. For simplicity we refer to these URLS as 'misinformation' but the measurement does not distinguish misinformation from disinformation (see 43)." And in the SI where we describe how we measure

our outcome variables, we now write, “Our domain list includes Grinberg et al’s list of ‘fake news’ sites 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, 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. While the specific criteria can vary from list to list, Lin et al. (43) finds high levels of agreement between domain lists.”

Lin, Hause, Jana Lasser, Stephan Lewandowsky, Rocky Cole, Andrew Gully, David G Rand, and Gordon Pennycook. 2023. “High Level of Correspondence across Different News Domain Quality Rating Sets.” Edited by Noshir Contractor. *PNAS Nexus* 2 (9): pgad286. https://doi.org/10.1093/pnasnexus/pgad286.

The authors use the threshold of 60 points for the selection of the NewsGuard sites for their analysis. One wonders which type of information providers are included after applying that threshold? Are information providers included in the study all providing content fabricated with the purpose to deceive? Are some of them falling short on standard journalistic routines and therefore, their content lacks the quality of professional journalistic sites? This differentiation is much clearer on Grinberg et al. 2019 when they distinguish between black, red and orange websites (or providers of fabricated stories and questionable claims respectively). The distinction is important and necessary if we aim to fully understand the quality of information available for public discussion and the impact of platform moderation policies in ensuring it.

Unfortunately, we are unable to provide full details on the Newsguard data due to our contract with them; as the Lin et al. cite above also notes, we are only able to identify five domains in their list at a time. In the SI section on measured variables we now write, “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 judicialwatch.org (which is also red in the Grinberg list) and dailywire.com (which is orange in Grinberg). The three most shared domains with a Newsguard score above 60 but below 70 are foxnews.com (not in Grinberg), msnbc.com (not in Grinberg), and 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) where Fox is shown as having a score of 57.”

My main concern is related though to the causal continuity assumption that the misinformation sharing would have occurred at the same rate had the intervention not occurred. I found the authors fall short to convince us this would have been the case and to

discuss evidence to support their assumption. Several events concurred on that day and the following ones that could have influenced the behavior of misinformation providers and sharers on Twitter. One can think about the multiple reminders to abide by the law sent by various sources from the judiciary, the executive, or the fact that the eldest daughter of the President deleted a tweet herself. Consistently, the authors should provide a lengthier discussion in the main manuscript to support their argument. Only towards the end and the discussion (lines #274 and #431), they acknowledge and emphasize that "their design does not prove the intervention is causal, in that there is no way for us to fully separate the impact of the insurrection from Twitter's enforcement." However, the authors are indeed interested in causality and the wording throughout the manuscript proves it: "we evaluate the direct causal effect"; "the contagion causal effect of this intervention"; "we also show a deterrent effect." If the current research design i.e., the SRD design, does not allow to disentangle the effect of the intervention from the effect of the insurrections itself as they eventually acknowledge, a careful consideration of the wording should be in place.

Thank you for this feedback; Reviewer 1 has similar concerns. As we describe above in responses to the editor and Reviewer 1, we have made extensive revisions to soften the language and to emphasize the nature and strength of the assumptions required by our natural experimental designs to take our results to be causal effects. We have revised each of the phrases you mention, and we also no longer rely on the labels "direct," "contagion" or "deterrent" causal effects in the text. We would like to mention that a substantial and core part of our analysis is based on the DID, which does not require the continuity assumption. The manuscript and SI provide various pieces of evidence for the assumptions required for the DID, along with many robustness checks for both the DID and SRD, which covers what is feasible to do with observational data to evaluate the validity of the design. In any case, we agree with your comment and addressed this concern by softening the causal language in the manuscript and discussing in more detail the respective limitations of each design and their necessary identification assumptions for causal interpretation. We give examples of these revisions in a response to Reviewer 1 above, and we have included a marked up version of the manuscript that details our changes.

The authors aim to understand if the "dose" of misinformation received has an impact on followers of deplatformed users. Towards this end, they define a follower as someone following at least four deplatformed users. How do the results change if this threshold varies? The threshold seems somehow arbitrary, and one wonders if different thresholds would yield different and more relevant results?

You are right that our original choice of 4+ followers to evaluate a higher "dose" was arbitrary. We have followed your advice and re-estimated the model with different versions of the measure – each time increasing the required number of deplatformed accounts that the user follows by one – and there is a new figure in the SI (Figure S9) with the results that show an increasing but

diminishing effect as the dose is increased. And we describe these results in the main text on p. 11. As this new added analysis shows (see SI Figure S9), our results are robust to the choice of dosage, and the effects follow the expected direction.

In the discussion section, there is room for improvement when it comes to defining what kind of studies are needed to overcome limitations of the study as well as what kind of data should be necessary for future replications of it. Even if those replications are currently not possible due restrictions to proprietary social media data, new regulations to grant data access elsewhere i.e., the EU, might overcome those limitations in the near future. The authors' insights on this regard would be highly valuable and make the discussion section stronger.

This is a great observation, especially because the research community is at a transition point in its research capacity. Platforms have made data progressively less available; indeed, this research would be impossible to replicate in 2024. On the other hand, the EU DSA offers a potential regulatory intervention to force platforms to make data available. This paper is well timed to help make the case for data availability from platforms. Space constraints limit how many words we could devote to this point (which could be the focus of an entire paper in and of itself), but we set up this point that data unavailability contributes to the dearth of research on platforms in the literature discussion on p. 3, and we note that the declining availability of data is a crisis for this kind of research in our final concluding paragraph of the paper. We note, though, that this is an unusual conclusion for a paper in *Nature*; but we do think it is a scientific crisis of the highest magnitude. We will look to the editor for guidance as to whether we retain this point.

Perhaps irrelevant for some readers, but still worthy to ask in my opinion is that in the introduction, the authors stated (line #59) "Twitter deplatformed a large number of right-wing extremists, including Trump himself." Should we assume from this that the US president is classified as a right-wing extremist here?

Yes, that's what we intended. We think that Trump's extremism is widely understood in both political science and among the general public.

To provide a broader context to the enforcement of Twitter's moderation policy, the authors might want to acknowledge previous deplatforming processes, at various scales, which have taken place before, both in western and non-western countries. See for instance the case of Myanmar, after legacy media raised attention of the use of Facebook to coordinate attacks against Rohingya minorities, or the suspension of multiple accounts following the outbreak of the Covid-19 pandemic, when social media platforms argued to have put in place unprecedented measures to tackle incivility online:
https://transparency.twitter.com/en/reports/rules-enforcement.html#2021-jul-dec

https://www.washingtonpost.com/technology/2021/02/24/facebook-myanmar-coup-genocide/
https://www.nytimes.com/2020/06/11/technology/twitter-chinese-misinformation.html
https://blog.twitter.com/en_us/topics/company/2020/information-operations-june-2020.html
https://about.fb.com/news/2020/12/coronavirus/

The examples you give are indeed important examples of other episodes where social media companies enforced a terms of use policy. On p. 3, in the last sentence of our literature review, we now mention the Facebook Myanmar event and the Twitter Chinese Covid deplatforming event, and we cite the Washington Post and New York Times articles you provided in these links. Also in the literature review and in the discussion we describe the Broniatowski et al paper that describes the Facebook event where they tried to remove vaccine misinformation.

The design problem is alluded to as one of the two potential reasons for the relatively small number of studies examining platform regulation. But the limited access to comprehensive datasets should be pointed out as well.

We have added this point as a sentence in the lit review on p. 3 as a second reason for the lack of studies, and we have added a cite to a Reuters article on social media data access.

The authors currently argue that their results demonstrate the capacity of social media platforms to regulate public discourse. The assumption here is that their study shows the effects of Twitter's decision over public opinion. Yet, the results cannot inform how or when the ideas spread by right-wing extremists shaped people's opinion in or outside Twitter. We do get useful insights limited though to the extent to which misinformation was reduced and the effects of suspension of users over followers sharing behavior.

We very much agree with this point, that the motivation for our paper takes for granted that misinformation has an impact on how individuals reason and develop opinions about policies and politics. We believe that misinformation spread is important in and of itself; however, we also believe that there is much work to be done to evaluate this assumption. In any case, we have deleted both the previous one-sentence summary as well as a paragraph in the conclusion where we made claims that social media companies have the capacity to regulate public discourse.

The authors argue (line #96) that "U.S. tech companies have a First Amendment right to determine what content their platforms amplify and suppress, and who is permitted to post content […]" Here, I question if this is strictly the case. Laws regulating content platforms can carry and who can access the platforms are frequently held to be violating the First Amendment, but isn't this because they see these laws as violating the rights of the social media users not that of the tech companies themselves?

You are right – Section 230 gives platforms protection from liability for the content that is shared and that is the key point. The Persily article that we cite in the manuscript frames section 230 as a First Amendment protection for the platforms, but there is no reason for us to argue this here. We have deleted the sentence that contained the First Amendment mention as it was unnecessary for the paragraph.

Edit for consistency (line #63) June 30th, 2020 and January 31, 2021 to June 30, 2020

Thank you for catching this. We have fixed it.

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

In my previous review of the paper, I found the approach to be interesting and the topic important but thought that the claims around causal impact needed to be handled more carefully because of the confounding of the Jan 6 incident. I find the updated manuscript handled this very well, and think it merits publication by Nature. Most of the positive points in my past review of the previous manuscript still stand. Some updated thoughts follow.

A. The results of this paper include a deep dive into the impact of account suspension performed on Twitter around the Jan 6 incident and their impacts on the amount of misinformation going forward. The updated paper has a nuanced approach to the SRD design as well as the parallel trends assumption in light of the events external to the platform.

B. The relevance is very high in a variety of fields. The work is not methodologically novel, but the data itself is of high interest and some scale, and the paper applies a broad spectrum of analyses of this data. While a similar style of analysis has been done on similar data from another platform (Facebook), the differences in conclusion from this analsysis makes these results quite interesting.

C. The data and methodology here are highly interesting. The extensive analysis and exploration into the SI are of high quality. This dataset has high value to the community and can hopefully be released.

D. With the updated language around the external confounding of Jan 6, I found the statisical analyses appropriate.

E. It is difficult to draw conclusions from observational data, but this paper does a good job of discussion conditional conclusions and boiling down the assumptions needed.

F. n/a

G. Looks good

H. The revised paper is easy to read and digest.

[Reviewer 2 was unable to re-review the paper]

Referee #3 (Remarks to the Author):

I think the manuscript has substantially improved and applaud the authors for the very good revision. I appreciate the work that the authors have put into this updated version of the paper addressing my main concern (also that of reviewer #2) on the violations of the assumptions needed for causal inference. I believe the paper merits publication after the extensive revision undertaken by the authors. I am pleased the authors included the discussion point on the relevance of data access to advance our understanding of the shortcomings of platform moderation policies (and platform interventions) and how legislative advances can secure future research of this caliber.