Changes in YouTube's Content Moderation Policy Had Little Detectable Impact on Election Denial Content

Nilima Pisharody [1,2], Sean T. Norton[2,*], Kevin T. Greene[2], Jacob N. Shapiro[2]

[1]Department of Economics, New York University
[2] Empirical Studies of Conflict Project, Princeton University
[*] Corresponding Author sn2489@princeton.edu

Abstract

While a growing body of research examines the effect of deplatforming, the blanket banning of certain types of content and users, on social media discourse, comparatively little research examines the effect of replatforming, when social media companies remove such blanket bans. We exploit such a policy change, YouTube's June 2nd decision to stop removing content denying the validity of the 2020 US Presidential Election, to examine the effect of replatforming content. Using data from YouTube, Facebook, Telegram conspiracy groups, and Google, we find no evidence that YouTube's policy change increased demand for or supply of election denial content over the short term. These results are consistent across all platforms and three different methodologies: regression discontinuity models, structural topic models, and a Bayesian structural time series model. This suggests that replatforming alone, when conducted after the effective marginalization of the targeted content, has minimal short-term effects on the spread of previously-prohibited content.

Introduction

Pathologies of the information environment represent a far-reaching global problem. While social media has connected people like never before, research suggests it has also served as a recruitment tool for extremists (Gates and Podder 2015), fueled ethnic and political conflict (Fink 2018), threatened the integrity of elections (Allcott and Gentzkow 2017), contributed to the spread of misinformation on public health crises (Cinelli et al. 2020), and damaged the mental health of children and teenagers (Boer et al. 2021).

One way that platforms attempt to deal with the spread of particularly harmful content is through the use of deplatforming: the total removal of specific users and content. While the scale of social media platforms means that some content and users escape such bans, evidence shows that deplatforming is effective at reducing the spread of and engagement with harmful content while also disrupting and degrading the networks of users who create and share such content (Thomas and Wahedi 2023; Innes and Innes 2023). At the same time, deplatforming of specific content or users on one platform may drive harmful content to other platforms with less-restrictive moderation policies, displacing some portion of the harmful content rather than simply eliminating it (Buntain et al. 2023; Mitts, Pisharody, and Shapiro 2022).

Among the most prominent examples of deplatforming occurred in the wake of the January 6th Capitol Riots in the US. After the deadly riots, the largest social media platforms (Facebook, Twitter, Instagram, and YouTube) undertook a massive deplatfoming action, banning thousands of accounts including outgoing US President Donald Trump. Many of these users and their followers fled to alternative social networks with less strict content moderation policies, such as Gab, Telegram, and TruthSocial (Buntain et al. 2023). This policy action would become known as the Great Deplatforming.

However, while a significant body of research exists on the effects of deplatforming, little research examines the effects of replatforming. On June 2nd, 2023 YouTube partially reversed the Great Deplatforming, announcing that while they would continue to ban content that promotes conspiracies about upcoming elections, they would no longer ban content promoting conspiracy theories about the 2020 US Presidential Election or any past elections. After two and a half years of nearly all election denial content being forced on to smaller platforms, it was suddenly replatformed on YouTube, a platform with approximately 239 million American users. Both experts and the press immediately voiced concerns, fearing that it would open a floodgate of misinformation and spur large increases in demand for and supply of election denial content across the social media ecosystem (Scott 2023; Bond 2023; Fischer 2023; Ingram 2023).

While these concerns were widespread, this paper empirically tests whether the replatforming of election denial content on YouTube led to substantively significant short-term increases in the prevalence of election denial content in four contexts: YouTube itself, Facebook, Google search trends, and conspiracy groups on Telegram. These platforms were chosen to study the impact on YouTube itself, the potential of contagion to another mainstream platform (Facebook), the level of interest among the online general public in election denial content (search trends), and activity among the groups most likely to be participate in creating and sharing election denial content (Telegram groups). We examine the prevalence of election denial content on Facebook and YouTube as well as activity levels in Telegram conspiracy groups using regression discontinuity. Additionally, we use structural topic models to analyze discourse on Facebook and Telegram, and a Bayesian structural time series model

to analyze search trends on YouTube and Google. Across all models and data sources, we find that YouTube's policy reversal did not result in any detectable changes to the supply of and demand for election denial content in the one to three months following the policy change. While the longer term impact of YouTube's June 2nd policy change may be of more policy interest than the short-term impact, we cannot causally identify this effect due to limitations in inference techniques and the presence of many confounders associated with the rapidly-evolving US political environment. However, our results suggest that replatforming alone is unlikely to have short-term effects.

Effects of the Original Deplatforming

While it is difficult to make causal claims on the effects of the January 6th deplatforming due to concurrent changes to the US political and media environments, descriptive research on the Great Deplatforming largely confirms two foundational results: it was highly effective on the mainstream platforms that instituted bans on election denial content while it simultaneously pushed many users to lesser-moderated, alt-tech platforms where extremist content was more common.

By analyzing discourse across Twitter, Reddit, and Gab, Buntain et al. (2023) shows a brief spike in hate speech on Twitter immediately following January 6th followed by a sustained 10-15% decrease in hate speech relative to a December baseline [1]. This is consistent with the experimental results in Thomas and Wahedi (2023), which found that targeted removals of the leadership of hate networks on Facebook resulted in a brief short-term backlash followed by sustained decreases in the amount of hate content produced and consumed, increasing overall platform health. Mekacher, Falkenberg, and Baronchelli (2023) also finds that banned Twitter users who migrated to Gettr were more toxic on Gettr than on Twitter, suggesting that the Great Deplatforming was effective in reducing the number of users engaged in toxic political speech on mainstream platforms.

However, the Great Deplatforming also created a surge of interest in Parler, Gab and Telegram, where election-denying content remained unmoderated; in the 2 months following January 6th, mentions of "switching" to Parler, Gab, and Telegram surged Buntain et al. (2023). Similarly, Buntain et al. (2023) finds large spikes in Google Trends interest for these platforms and increased mentions of and links to Rumble (an alt-tech video sharing platform similar to YouTube). Displacement effects were most pronounced on Telegram; Bryanov et al. (2021) finds that large, right-wing communities on Telegram experienced rapid userbase growth in the wake of January 6th. Similarly, Mekacher, Falkenberg, and Baronchelli (2023) use a dataset that matches banned Twitter users to accounts on Gettr, finding that users who were banned from Twitter had significantly higher activity and user retention rates than Gettr users who were not banned on Twitter. Taken together with the evidence from mainstream platforms, the evidence suggests that while the Great Deplatforming substantially reduced toxic and election denial content on mainstream platforms, it displaced some quantity of this content to alternative platforms, allowing many users to remain active in creating and engaging with such content.

This fits with the results of other research on smaller-scale deplatforming incidents. Innes and Innes (2023), studying the deplatforming of two prominent COVID conspiracy activists on Facebook, finds that while deplatforming disrupted the ability of these actors to spread misinformation to and through their network of supporters, it also displaced the conspiracy actors' content to other platforms, creating a path for it to indirectly spread back to Facebook via off-site linking. Rauchfleisch and Kaiser (2021) found that of 111 deplatformed, far-right, YouTube channels only 20 established new presences on alt-tech platform BitChute, where they garnered considerably less engagement and reach. While the scale of the Great Deplatforming is unique, it does not appear as if the effects differ substantially from expectations set by previous research.

Background on Policy Change

YouTube's election misinformation policy (YouTube N.d.) was originally implemented as part of the Great Deplatforming in January 2021, and prohibits content that:

- Attempts voter suppression by misleading people about "time, place, means, or eligibility requirements for voting" or that makes "false claims that could materially discourage voting"
- Makes false claims about candidate eligibility
- Encourages others to interfere with the democratic process, e.g. by disrupting or obstructing voting
- Content that calls into question election integrity by "advancing false claims that widespread fraud, errors, or glitches occurred in certain past elections to determine heads of government. Or, content that claims that the certified results of those elections were false."

The policy remained in force until June 2nd 2023, when YouTube abruptly announced a major policy change; effective immediately, the platform would "stop removing content that advances false claims that widespread fraud, errors, or glitches occurred in the 2020 and other past US Presidential elections" (YouTube 2023). YouTube stated that they were changing this policy because while the ban may have reduced the spread of misinformation, they were concerned that it would also curtail political discussion without reducing real-world harms. Crucially, this policy change only applied to past elections - false claims about the upcoming 2024 US Presidential Election or any other upcoming elections remain banned.

Materials and methods

To determine the short-term effect of YouTube's policy change, we combine data from 4 different sources to analyze 11 different outcomes. Table 1 summarizes these data sources and the methods used to analyze them, while the Data and Methods sections provide further detail.

[1] Changes on Reddit were minimal.

Platform	Outcome	Resolution	Time Range	Test
YouTube	# of videos using denial keywords	Daily	Apr. 30 -July 1	RDD
YouTube	# of unique channels using denial keywords	Daily	Apr. 30 -July 1	RDD
YouTube	Search interest in election denial terms	Weekly	Jan. 1 - Sept. 7	Model
Facebook	# of election denial posts	Daily	Apr. 30 -July 1	RDD
Facebook	% of election denial posts	Daily	Apr. 30 - July 1	RDD
Facebook	Topic prevalence	Daily	Apr. 30 - July 1	STM
Telegram	# of posts in extremist groups	Daily	Apr. 30 - July 1	RDD
Telegram	# of unique posters in extremist groups	Daily	Apr. 30 - July 1	RDD
Telegram	Avg. views per post	Daily	Apr. 30 - July 1	RDD
Telegram	Topic prevalence	Daily	Apr. 30 - July 1	STM
Google	Search interest in election denial keywords	Weekly	Jan. 1 - Sept. 7	Model

Table 1: Platforms, outcomes, time resolution of data, time range of data, and test method. All dates are in 2023 unless otherwise specified. RDD stands for regression discontinuity design, STM stands for structural topic model, and "model" indicates the Causal Impact model.

Data

This section describes the process used to select our election denial keywords and provides descriptive detail on our data sources. For a detailed discussion on the limitations and potential biases of our data, please see the Limitations subsection within the Results section.

Election Denial Keywords To enable classification of election denial posts on Facebook and identify election denial-related search terms, we first built a dictionary of election-denial related keywords. To build the dictionary, we identified domains highly active in spreading election denial content from three sources: "The Big Lie and Big Tech" (Baldassaro, Harbath, and Scholtens 2021), a report by the Carter Center, "Mail in Voter Fraud: Anatomy of a Disinformation Campaign" (Benkler et al. 2020), a report by Harvard's Berkman Klein Center, and sites rated as low reliability by MediaBias/FactCheck. We then used SerpAPI to gather articles related to election denial published prior to June 2nd 2023, from which we sampled 25 articles. Additionally, we sampled 200 posts from Facebook groups that posted previously posted election denial content and 200 posts from the election denial-focused Telegram group "Stop the Steal." The use of articles from these domains allows us to identify the language most used in news or news-related content that spread election denial conspiracies, whereas the use of of Facebook groups that previously shared election denial content and one of the most prominent public Telegram groups in election denial conspiracies (Stop the Steal) allows us to identify the language that those actively discussing election denial content used.

Each of these 425 documents were closely read by human coders and keywords that were unambiguously used to indicate election denial, but not elections generally, were extracted. For instance, we do not include "voter registration" but do include "falsified voter registration." Selecting into a variety of sources where election denial content is likely, and then having human annotators closely read the materials helps us ensure that the keywords we collect relate to election denial content, rather than elections generally. For a list of these keywords, see Appendix Table 8.

Social Media Data YouTube data was obtained through YouTube's Research API. We used our list of election denial keywords to search for videos posted between April 30th and July 2nd 2022. We then downloaded metadata for each returned video until either no results remained or we reached the fifth page of results, whichever came first. To count the number of unique channels posting videos per day, we counted the number of unique channel IDs associated with videos in each calendar day.

Telegram data was scraped using the official Telegram API from a curated list of public Telegram groups associated with the American far-right and conspiracy communities. This list was built and maintained by the Bridging Divides Initiative at Princeton. These channels range from those run to provide updates to members of far-right activist groups, such as Patriot Front and the Proud Boys, election conspiracy centered groups, and channels associated with right-wing social media influencers.

Facebook data was obtained by sampling public pages using the CrowdTangle API. We first searched the API for all groups/pages that used the words "election" or "voter" in May 2023. We sampled 1,000 pages/groups from a list of all unique groups returned and then scraped all posts from May 3rd, 2023 to July 2nd, 2023. Facebook posts were classified as being election denial content if they contained one of these election denial keywords.

Google and YouTube search trends data was obtained through the Google Trends API using the gtrendsR package. We downloaded trends for each of the election denial terms in our dictionary and averaged them, creating a separate composite indicator of election denial search interest for both platforms. Not every term in the dictionary generated enough search interest to be included in the public Google Trends data. See Appendix Table 8 for a full list of terms included in the composite indicator.

Methods

Regression discontinuity designs (RDDs) were used for daily counts of election denial content on Facebook and Youtube and daily active users and average views per post on Telegram. We chose regression discontinuity because the YouTube policy change represents a "sharp" treatment event on June 2nd, before which election denial content was uniformly banned and after which some election denial content was again permitted. Regression discontinuity estimates a local treatment effect around such a cutoff point by estimating an optimal "bandwidth" of data (in our case, days) surrounding the cutoff, fitting a local polynomial to within-bandwidth data on either side of the cutpoint, and estimating the "gap" between the predicted value of the regressions at the cutoff. In situations where randomization of treatment is impossible, such as platform-wide policy changes, RDD provides a valid causal estimate of the local treatment effect of the intervention even in the absence of a control group. Additionally, estimating the local average treatment effect (LATE) is preferable over estimating the average treatment effect (ATE) in our case; while the amount of election denial content immediately prior to or following the YouTube policy change is plausibly exogenous from other factors, this exogeneity is impossible to maintain as time since treatment increases. To conduct the RDDs, we first grouped all rel-

evant dependent variables at the daily level. We used the number of days before/after June 2nd, 2023 as the running variable. RDDs were estimated in R using the package *rdrobust*. The *rdrobust* package automatically selects the optimal bandwidth and provides bias-corrected and robust confidence intervals and p-values (Calonico, Cattaneo, and Titiunik 2015).

Facebook and Telegram group posts were analyzed using a structural topic model (Roberts et al. 2014). The structural topic model (STM) was chosen because it allows us to estimate the impact of covariates, in this case whether a post was before or after the June 2nd policy change, on topic prevalence. This allows us to estimate whether discussion of election denial or other political topics became more prevalent on Facebook and in Telegram conspiracy groups after the replatforming of election denial content on YouTube. This cannot be accomplished with the standard topic model or more advanced, transformer-based methods such as BERTopic (Grootendorst 2022). While it is in theory possible to compare the prevalence of topics before and after the cutoff using a difference-in-means test, this would require a multiple hypothesis test adjustment equal to the number of topics and would not price in relevant modeling uncertainty (e.g. in document-level topic responsibilities). Nonetheless, a robustness check with BERTopic (Appendix tables 12 and 13) revealed that the highest-populated topics were broadly similar to the topics identified by the STM.

Before modeling with the STM, we cleaned the corpora for both platforms by removing a standard list of non-English stopwords, all non-words, and any word not used in more than one document. All non-English posts were also removed from the corpus. Since social media posts are generally short, making it difficult to calculate the document-level statistics necessary for fitting and interpreting topic models, we grouped all posts at the group-day level. We selected k (the number of topics) by first fitting many models and selecting the models that perform best at the coherence-exclusivity frontier, as recommended in Roberts et al. (2014) (see Appendix Figures 16 and 17). We then selected the best performing models for human coding; the coder used both the keywords identified by the topic model and the 25 most representative documents in each topic to determine whether the topics were cohesive and interpretable. For both platforms, only one candidate model returned a set of topics which were all interpretable and cohesive. All models and effects were estimated using the *stm* package in R (Roberts et al. 2014).

Google and YouTube search trends data was analyzed using the Causal Impact model as described in Brodersen et al. (2015) and implemented in the *CausalImpact* R package. The CausalImpact model is a Bayesian structural time series model that enables inference on interventions when no direct counterfactual, such as a control group, is available or possible. It constructs a synthetic control trend using a provided set of control time series, such as other search queries, that are not affected by the treatment. The model is robust to the potential inclusion of irrelevant controls because it uses a slab-spike prior to choose only the controls which are useful in approximating the pre-treatment trend. By comparing the trend of the dependent variable to the synthetic control in the post-intervention period, the model allows us to acquire a semi-parametric estimate of the treatment effect. Since Google and YouTube Trends data is aggregated at the calendar week level, we use a weekly seasonality component in the structural time series model. To determine whether the effect is non-spurious, we use the model-provided credible intervals and posterior probability of an effect. Crucially, this model was initially built to analyze the impact of advertising interventions on Google search traffic, making it uniquely appropriate to this use case. Election denial terms are available in Appendix Table 8 and the terms used to construct the synthetic control are available in Appendix Table 9.

Platform	Outcome	Statistical Significant	Substantively Significant	Direction
YouTube	# of videos using denial keywords	✖	✖	
YouTube	# of unique channels using denial keywords	✖	✖	
YouTube	Search interest in election denial terms	✖	✖	
Facebook	# of election denial posts	✖	✖	
Facebook	% of election denial posts	✖	✖	
Facebook	Topic prevalence	✖	✖	
Telegram	# of posts in extremist groups	✖	✖	
Telegram	# of unique posters in extremist groups	✖	✖	
Telegram	Avg. views per post	✖	✖	
Telegram	Topic prevalence	✓	✓	-
Google	Search interest in election denial keywords	✖	✖	

Table 2: Summary table of results. Where results are statistically significant, $+$ indicates that the direction of the effect is positive and $-$ that the direction is negative.

Results

Table 2 summarizes the data sources, outcome, and results. Each subsection provides more detail on the relevant analyses.

YouTube: Effects on Supply and Demand of Election Denial Videos

Using the procedure detailed in the Methods section, we created a daily count of videos returned by searching election denial keywords, displayed in Figure 1 and Figure 2.

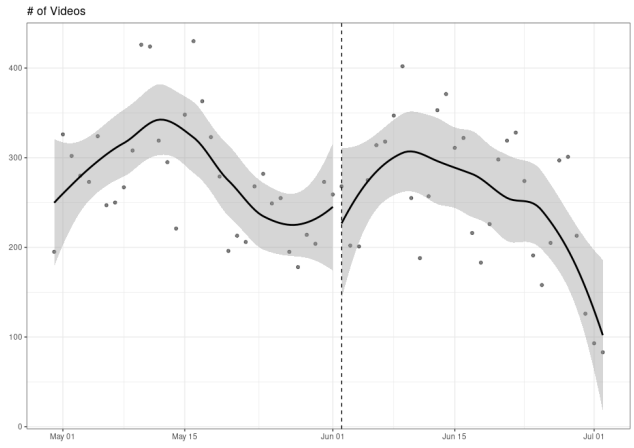

Figure 1: Daily number of videos posted related to election denial search terms with LOESS trend lines. The dashed line indicates June 2nd, the date of the policy change.

	# Videos		# Unique Channels	
Estimator	Coefficient	p-value	Coefficient	p-value
Conventional	−46.365	0.207	−18.149	0.497
Robust		0.142		0.365
Bandwidth:	MSE optimal		MSE optimal	
Kernel:	Triangular		Triangular	
* p < 0.05				

Table 3: YouTube: Sharp RDD estimates for the number of daily videos returned by election denial term searches (left) and the number of unique channels responsible for returned videos (right). Effects are not statistically-significant for both outcomes across all estimators.

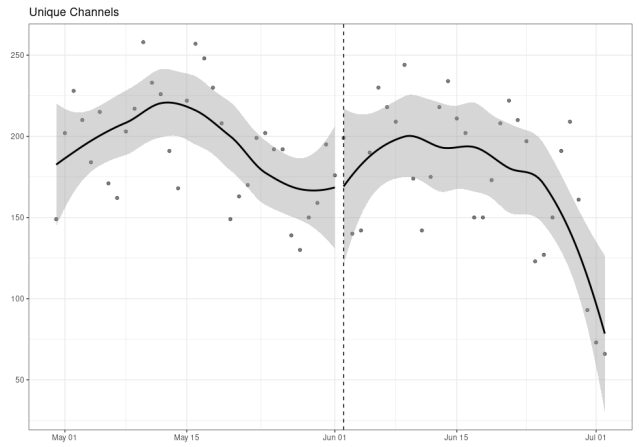

Figure 2: Daily number of unique channels posting videos related to election denial search terms with LOESS trend lines. The dashed line indicates June 2nd, the date of the policy change.

The RDD analyses do not find a statistically significant short-term increase in the number of videos returned by election denial search terms following the June 2nd policy change nor does it find any short-term change in the daily number of channels posting such videos (Table 3). As such, we determine that the policy change was not associated with an increase in the short-term supply of election denial videos.

While there is no evidence that the policy change increased the number of YouTube videos or channels posting videos related to election denial search terms, it is possible that the YouTube policy change increased demand for this information. We test this using Google and YouTube search trends. Using our dictionary of election denial search terms, we used the Google Public API to get weekly trends for each term from January 1st 2023 to September 9th 2023. These trends were combined into a composite indicator by taking the mean of all trends, constructing a measure of election denial Google and YouTube search activity from January 1st, 2023 to September 29th, 2023. We then used a Bayesian structural time-series model, the Causal Impact model (see Methods), to evaluate the impact of the policy change on

	Average	Cumulative
Actual	12	200
Prediction	13	221
95% CI	[10, 15]	[177, 262]
Absolute effect	-1.3	-21.5
95% CI	[-3.7, 1.4]	[-62.5, 23.3]
Relative Effect	-8.7%	-8.7%
95% CI	[-24%, 13%]	[-24%, 13%]
Posterior prob. of a causal effect:	82%	

Table 4: Google searches: average and cumulative differences in election denial search activity between the synthetic counterfactual and the observed data, with 95% credible intervals.

election denial search activity (Brodersen et al. 2015). Figure 3 shows the trend of this composite indicator over time with cutpoints at the 2022 US congressional elections and the June 2nd YouTube policy change.

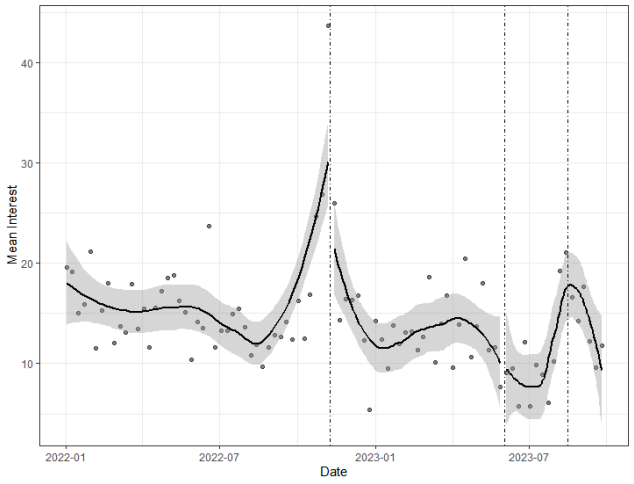

Figure 3: Google: composite indicator of election denial search activity, Jan. 1 2022 - Sept. 29 2023 with LOESS trend lines. The leftmost dashed vertical line indicates the US 2022 midterm elections, the 2nd line to the right indicates the June 2nd 2023 YouTube policy change, and the final line on the right indicates the Georgia indictment of Donald Trump on August 15th 2023.

Interest in election denial remains generally low over time. Notably, the 2022 midterm elections are associated with a major spike in interest in election denial search terms; this coincides with increased traditional and social media attention on election conspiracy claims.

Using data from Google Trends and the Causal Impact model, we find that there is an 82% posterior probability of a -1.2 unit absolute decrease and -21.5 unit cumulative decrease in Google search interest for election denial terms (Table 4 and Figure 4). The 95% credible intervals for both the absolute and relative effects include zero, indicating that

the effect is indistinguishable from zero.

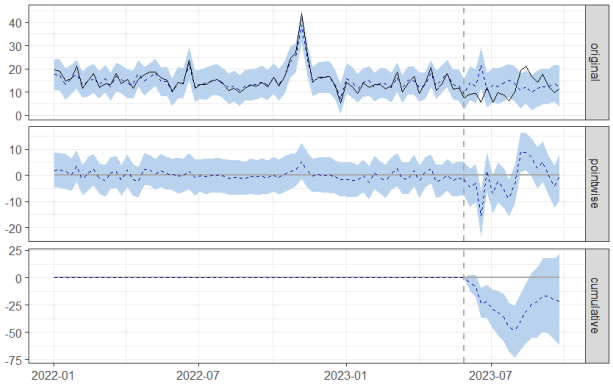

Figure 4: Causal Impact model, Google. The vertical dashed line is the date of the policy change. The original panel compares the observed data (solid line) to the synthetic counterfactual data (dashed). The pointwise panel shows the difference between observed and counterfactual, i.e. the effect at each time point. Cumulative sums the pointwise estimate from the date of the intervention.

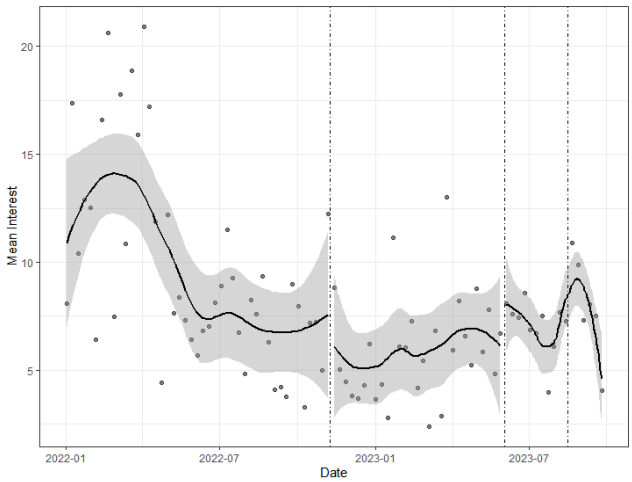

Figure 5: YouTube: composite indicator of election denial search activity, Jan. 1 2023 - Sept. 29 2023 with LOESS trend line. The leftmost dashed vertical line indicates the US 2022 midterm elections, the 2nd line to the right indicates the June 2nd 2023 YouTube policy change, and the final line on the right indicates the Georgia indictment of Donald Trump on August 15th 2023.

Interest in election denial search terms on YouTube has a lower baseline than interest in Google searches, averaging 8.071 to Google's 14.4. It displays a similar spike surrounding the 2022 midterm elections.

The Causal Impact model finds a 76% posterior probability of 0.78 unit absolute and 13.29 unit cumulative increase in election denial related search interest following June 2nd

	Average	Cumulative
Actual	7.4	125.6
Prediction	6.6	112.3
95% CI	[2.9, 8.9]	[49.1, 151.6]
Absolute effect	0.78	13.29
95% CI	[-1.5, 4.5]	[-26, 76.5]
Relative Effect	25%	25%
95% CI	[-17%, 155%]	[-17%, 155%]
Posterior prob. of a causal effect:	76%	

Table 5: YouTube searches: average and cumulative differences in election denial search activity between the synthetic counterfactual and the observed data, with 95% credible intervals.

(Table 5 and Figure 6). As with the Google model, the 95% credible interval for both effects contains zero, indicating the the effect is indistinguishable from zero.

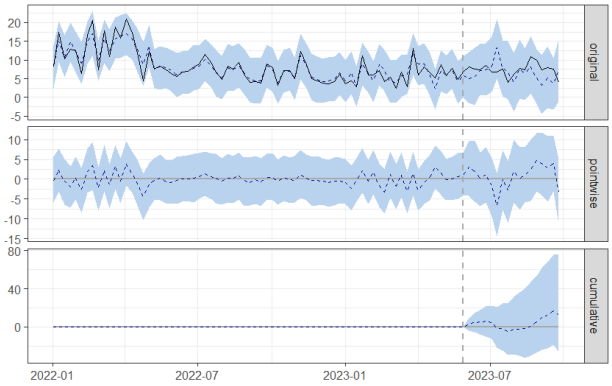

Figure 6: Causal Impact model, YouTube. The vertical dashed line is the date of the policy change. The original panel compares the observed data (solid line) to the synthetic counterfactual data (dashed). The pointwise panel shows the difference between observed and counterfactual, i.e. the effect at each time point. Cumulative sums the pointwise estimate from the date of the intervention.

Facebook: Election Denial Demand and Discourse

Our Facebook data consists of 766,168 posts from May 3rd, 2023 to June 30th, 2023, sampled using the procedure detailed in the Data section. Posts were coded as election denial content if they contained an election denial keyword. Figures 1 and 2 show trends in both the total number of posts and the number of election denial posts, respectively. Election denial posts represent a very small share of all sampled Facebook content, averaging .081% over the sampled time frame. Figures 7 and 8 display the total daily post count and the total number of daily election denial posts.

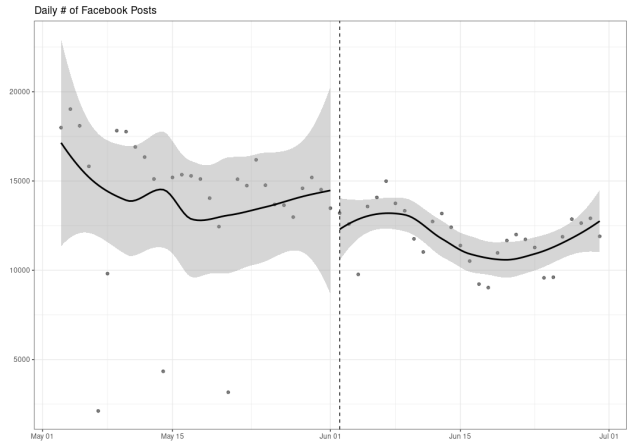

Figure 7: Daily Facebook posts in the dataset with LOESS trend lines. The dashed line indicates June 2nd, the date of the policy change.

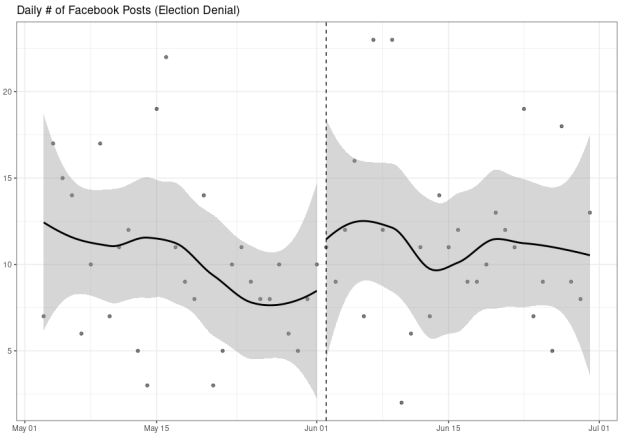

Figure 8: Daily Facebook election denial posts in the dataset with LOESS trend lines. The dashed line indicates June 2nd, the date of the policy change.

As with our YouTube data, we estimated regression discontinuity design (RDD) models with a sharp cutoff of June 2nd 2023. Models were estimated for both the number of election denial posts and the prevalence (percentage) of election denial content at the daily level (Table 6). We find no statistically significant effect of the policy change on daily total or prevalence of election denial posts.

To determine whether the policy change resulted in shifts in the topics discussed in our Facebook data, we use structural topic models (Roberts et al. 2014), as detailed in the Methods section. Both models use a single covariate, *post*, which indicates whether the group-day document was created prior to June 2nd 2023 or later (inclusive of June 2nd).

	# Denial Posts		% Denial Posts	
Estimator	Coefficient	p-value	Coefficient	p-value
Conventional	0.812	0.792	0.000 25	0.323
Robust		0.963		0.570
Bandwidth:	MSE optimal		MSE optimal	
Kernel:	Triangular		Triangular	
* p < 0.05				

Table 6: Facebook: Sharp RDD estimates for the daily number (left) and daily percentage (right) of election denial posts on Facebook. Effects are not statistically-significant for both outcomes and estimators.

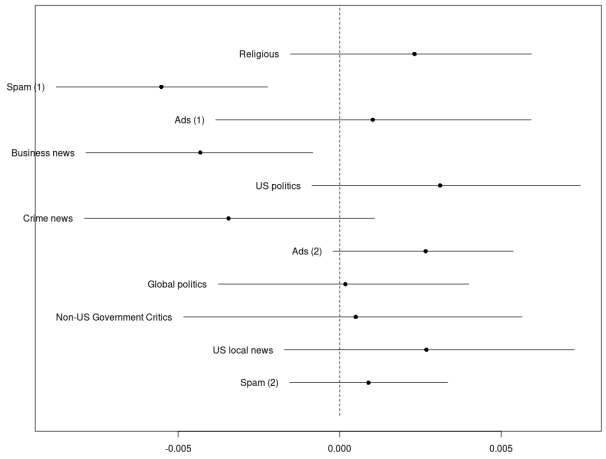

Figure 9: Structural topic model, Facebook: Change in topic prevalence after June 2nd YouTube policy change. We do not observe any significant changes in prevalence of topics possibly related to election denial content.

Election denial content is extremely rare in the Facebook data, representing only 0.0008 % of all posts on public pages in our sample. As such, the Facebook model does not clearly identify any topics related specifically to election denial. However, we see that the two topics most related to politics and most likely to contain election-denial content - US politics, global politics, and US local news - do not experience any change in prevalence post-June 2nd (see Figure 9). Combined with the RDD results on the number and prevalence of election denial posts on Facebook (Table 6), this indicates that the YouTube policy change had no significant effect on Facebook discourse.

Telegram: Policy Impact on Conspiracy Spreaders

We also estimate RDDs for our Telegram data. However, this dataset is qualitatively different than the Facebook dataset; while the Facebook data is a sample of public pages from the CrowdTangle API, the Telegram data is scraped from 90 specifically-targeted conspiracy and far-right groups over the period May 3rd to July 2nd, 2023 (see Materials and

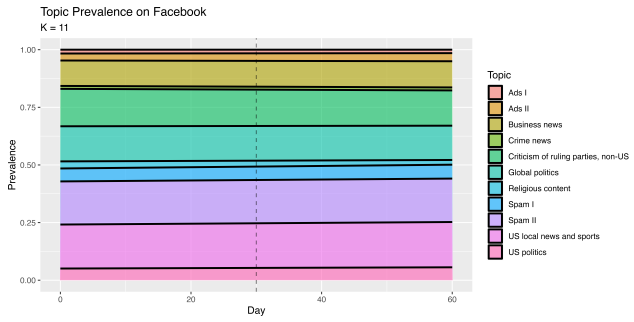

Figure 10: Facebook: topic prevalence over time.

Methods, all groups/channels are public). Rather than estimating the percentage of election denial content, we instead seek to determine whether the June 2nd YouTube policy change activated these groups, increasing the number of posts (Figure 11), number of unique users creating posts (Figure 12), or the average number of post views (Figure 13). We estimate RDDs for these three quantities at the daily level, finding no statistically significant change in any metric (Table 7).

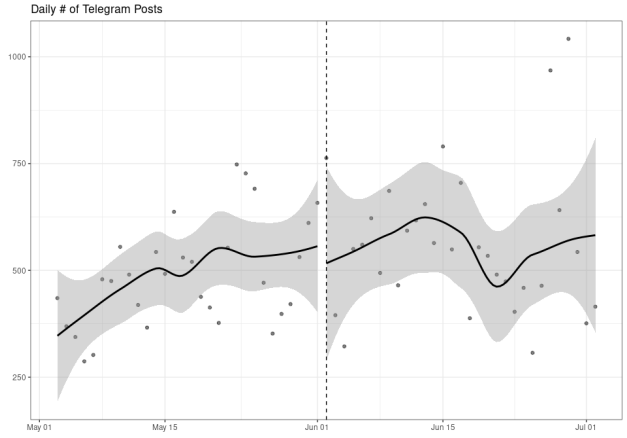

Figure 11: Total posts per day in the Telegram dataset with LOESS trend lines. The dashed line indicates June 2nd, the date of the policy change.

	# Posts		# Active Users		Avg. Views	
Estimator	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
Conventional	−155.358	0.465	−8.278	0.459	−2400.030	0.161
Robust		0.320		0.414		0.170
Bandwidth:	MSE optimal		MSE optimal		MSE optimal	
Kernel:	Triangular		Triangular		Triangular	
* p < 0.05						

Table 7: Telegram: Sharp RDD estimates for the daily number of posts (left), daily number of active users (center), defined as users who create a post, and the daily average number of post views (right). There are no statistically significant effects across all three outcomes.

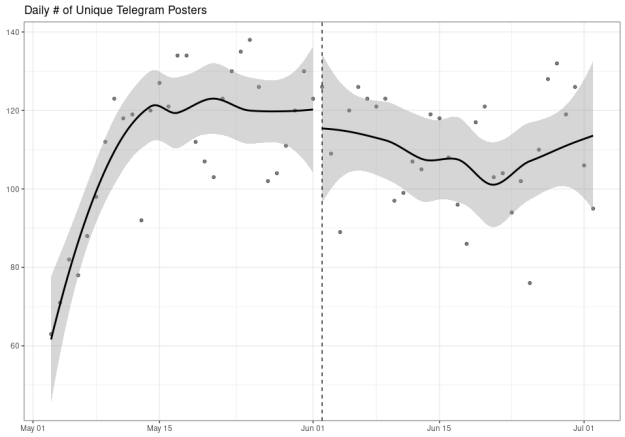

Figure 12: Total unique active users per day (posted content) in the Telegram dataset with LOESS trend lines. The dashed line indicates June 2nd, the date of the policy change.

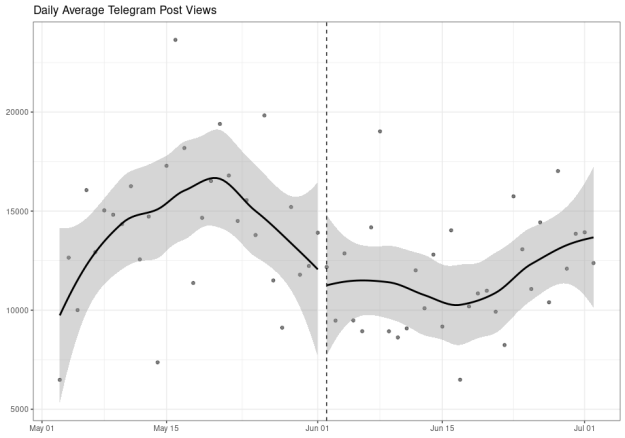

Figure 13: Daily average views per post in the Telegram dataset with LOESS trend lines. The dashed line indicates June 2nd, the date of the policy change.

To determine whether the policy change resulted in shifts in the topics discussed across our dataset, we use a structural topic model, following the same procedures used to model the Facebook data (Roberts et al. 2014).

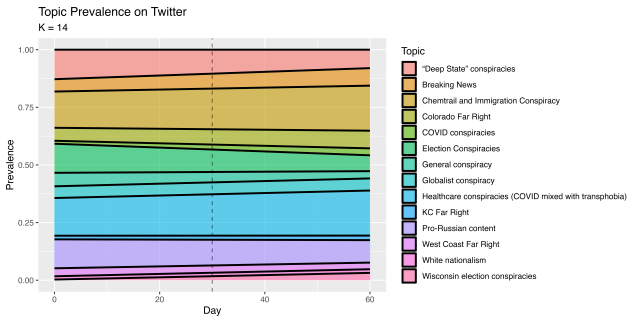

Figure 15: Telegram: topic prevalence over time.

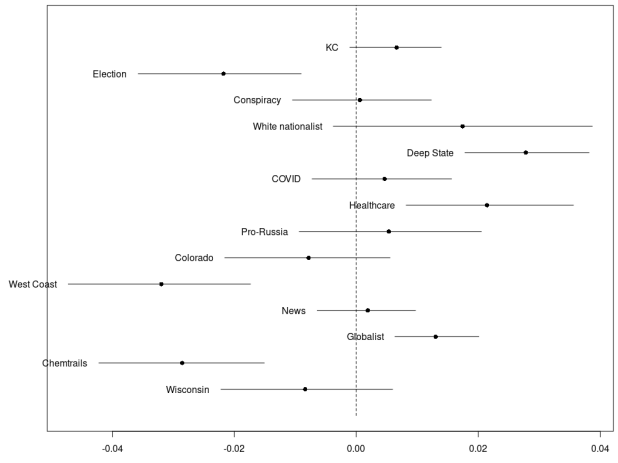

Figure 14: Structural topic model, Telegram: Change in topic prevalence after June 2nd YouTube policy change. Election denial content becomes slightly less prevalent after June 2nd.

As Figure 14 demonstrates, Telegram groups discussed the topic most associated with election conspiracies less in the aftermath of the June 2nd policy Ccange. The only topics that show statistically significant increases in prevalence after June 2nd are deep state, globalist, and healthcare conspiracy topics. There is no evidence that the replatforming of 2020 election conspiracies on YouTube increased demand for or supply of election conspiracy content in the tracked far-right and conspiracy Telegram groups, who we would expect to be most active in creating and consuming election denial content across all platforms where such content is permissible.

Figures 10 and 15 display topic prevalence over time for both models. Discourse on Facebook is considerably stable with little to no changes in topic prevalence after June 2nd. Discourse in Telegram conspiracy groups is somewhat less stable, but the decrease in election-related content after June 2nd, a trend which began prior to the policy change, is clear.

Limitations

The limitations of the data fall into three broad categories: potential bias due to the use of keywords to identify election denial content, potential bias due to sampling from a restricted, public population of social media data, and measurement bias in the search trends data. We discuss each of these limitations and their potential implications in turn. In sum, we believe that the wide variety of data and analytical methods used in this work greatly reduce the risks of bias associated with any individual data source.

Keywords were chosen using criteria designed to unambiguously identify content that contained election denial conspiracies rather than basic discussions of elections and voting. However, keywords cannot distinguish between actual election denial content and non-conspiracy discussion of election denial content, e.g. by news media or fact-checkers who directly quote denial content. If election denial content had increased in the wake of the YouTube policy change, we would have expected non-conspiracy discussion to also increase as social and traditional media reacted to the increased prevalence of election denial content. If such an increase had occurred, our methods could not have determined what share of our data represented real election denial content versus reactions and more detailed analysis, e.g. of video transcripts, would have been necessary. However, we observe no increase of content using election denial keywords across all data sources, indicating that neither election denial content nor reactions to election denial content increased in the 1-3 months following YouTube's policy change. Since there is no increase in keyword use, it is not critical to distinguish between denial content and reactions to denial content. For there to be a null result on total keywords if denial content had increased due to the policy change, the rate of discussion of denial content would have had to go down on the date of the policy announcement, which seems unlikely.

Keywords also do not tell the full story; it is possible that despite the broad data sources we used to generate our keyword list that we missed important keywords, thus biasing our data. The use of structural topic models on the Facebook and Telegram corpora is intended to mitigate this bias. Facebook and Telegram data was not selected using keywords. Facebook data was random sampled from a list of any public page that used the words "voter" or "election" in May 2023. This exclusion criteria was intended to exclude groups that never or extremely infrequently discuss politics; previous research has shown that most social media users rarely seek out, see, or interact with political content (Wojcieszak et al. 2022). Sampling from the total population of public Facebook groups would have required an impractically large sample in order to avoid biasing the results towards a null effect. Telegram groups come from a list of far-right and conspiracy channels and are not intended to represent a general social media audience, but rather the audience most likely to spread and interact with election denial content. Regardless of the data source, the structural topic model uses the entire corpora of data to identify topics in an unsupervised manner, i.e. it is not seeded with our election denial keywords. In both data sources, we observe no statistically-significant

increase in election denial or political topics, meaning the results are unlikely to be driven by keyword selection.

While the use of public Facebook and Telegram data does reduce the concerns of keyword-driven bias, it does impose some additional limitations on our results. The vast majority of content on Facebook and Telegram is private and unavailable to researchers; we should expect some proportion of that private data to contain election denial content. It is possible that while there was no observed increase of election denial content in public data, such an increase does exist in private data. The use of Google and YouTube search trends partially ameliorates this risk; if private discussion of election denial content increased, we would also expect to see an increase in demand for election denial content reflected in search trend data. We observe no such increase. Ultimately, the unavailability of private data combined with the lack of a field-wide ethical and safety framework for research using private data is a problem that all social media research faces. Like other researchers, we cannot completely eliminate the concerns arising from the field's blind spots.

Finally, Google and YouTube search trends data is a relative measurement and not an objective one, i.e. it measures the popularity of a search term relative to all other search terms over a calendar week. The weekly resolution means that temporary spikes in search interest for election denial terms may be washed out by less interest over the course of the week. Given that we find no increase in supply of election denial content at a daily resolution on both YouTube and Facebook, it is unlikely that any such spike occurred. Furthermore, any spike in search interest large enough to attribute substantive significance to would almost certainly be large enough to shift the measurement upward for any given week and would be detectable in our data. We do not detect any such spikes except surrounding the 2022 US midterm elections and the Georgia indictment of former President Donald Trump, suggesting that if a substantively-significant increase in election denial searches occurred after June 2nd, it would have been visible even at weekly resolution.

Discussion

While experts, lawmakers, and the media predicted that YouTube's June 2nd policy reversal would create a flood of election denial content across the social media ecosystem, we find little to no short-term effect of the policy change on the supply of and demand for election denial content (Scott 2023; Bond 2023; Fischer 2023; Ingram 2023). While determining the the exact reasons for this null effect is beyond the scope of the paper, we argue that this possibly reflects both a lack of incentives for election denial content creators to migrate en masse back to YouTube as well as a lack of latent, unmet demand for election denial content on YouTube and other platforms.

If predictions that YouTube's June 2nd policy change would lead to a rapid, cross-platform increase of election denial content had been true, they could only have been true if election denial content creators returned to YouTube after the policy change. Previous research on deplatforming has found that it displaces large numbers of users and content creators to less-moderated platforms, and that the Great Deplatforming had a large displacement effect (Buntain et al. 2023). At the time of the YouTube policy change, many previously-banned election denial content creators had been active on alternative platforms for over 2 years. For these displaced content creators, replatforming created two choices, which we call return and continued substitution:

- *Return*: Deplatformed individuals, seeking the restoration of their original audiences or communities, return to the platforms from which they were banned and begin posting previously-prohibited content. If successful in reactivating old audiences or creating new ones, this could lead to wider spread of the previously deplatformed content on mainstream platforms.
- *Continued substitution*: Deplatformed individuals, having built new audiences on new platforms, do not see the utility in returning to the original platform and remain on alternative platforms with smaller audiences. This maintains the status quo created by the original deplatforming incidents.

Return was likely an attractive option to content creators due to monetization and audience size incentives. Creators of election conspiracy content were forced to smaller, alt-tech platforms as a result of the Great Deplatforming, substantially reducing the sizes of their audience and their potential payouts. Returning to their original, mainstream platform offers the opportunity to rebuild their audiences and their incomes. However, return does not come without costs; re-establishing an audience on a mainstream platform is time-consuming and there is no guarantee that it will result in higher profits or larger audiences than on alternative platforms. Given that maintaining an audience on even one platform is time-consuming for content creators, trying to rebuild an audience on YouTube while simultaneously retaining an audience on an alt-tech platform may simply have been too costly for content creators (Arriagada and Ibáñez 2020). This is especially likely to be true if YouTube's platform affordances and algorithms differ in even minor ways from creator's current platforms, as customizing content to maximize reach on specific platforms is extremely labor-intensive (Duffy and Sawey 2021).

Furthermore, YouTube did not restore previously banned accounts or videos, meaning that content creators returning to YouTube would need to either repost all previously deleted content or start from scratch. For those who have established new audiences on alt-tech platforms, this may have been costly enough to disincentivize them from returning to their original platform. Additionally, the June 2nd YouTube decision was only a partial replatforming; conspiracy content about past elections is now allowed, but content creators still cannot cultivate conspiracies about current or upcoming elections. To avoid being banned or having content deleted a second time, content creators would still need to carefully moderate their new videos. Content creators and users may also not trust mainstream platforms to maintain their replatforming; if the platform abruptly reverses course on replatforming, they could simply be banned again. Since we do not observe a short-term increase in the supply of election denial content, the null results of this study could be explained by

content creator's unwillingness to put necessary the time, effort, and risk into returning to YouTube.

It is also plausible that the partial nature of this replatforming effectively prevents creators from exploiting unmet demand for election denial content as it exists in 2023. If audiences are now significantly more interested in conspiracies about upcoming elections, content creators would not be able to activate this latent demand since such content remains banned. Demand for election denial content may also be in a natural lull given that the YouTube policy change occurred in an inter-election period. If the policy change had occurred more proximate to either the 2022 midterm elections or the 2024 presidential election, the demand needed to kick-start a discursive feedback loop that leads to spread of and discussion of election denial content across multiple platforms may have been more likely to esist. It is possible that the participation of two prominent actors from the 2020 elections in the 2024 elections - Vice President Harris and former President Trump - may create such a feedback loop as the election approaches, but it appears replatforming alone was not sufficient to create this feedback loop.

While we do not find short-term effects associated with YouTube's policy change, we should reasonably expect the policy change's effects to play out over the long term as well. While the long-term effects are of great academic and policy interest, causally identifying the long-term effect of a moderation policy is difficult given the compounding of confounders that should be expected to affect the supply and demand of election denial content. As such, we have focused solely on identifying primarily the short-term effect of the policy change (see Table 1 for details on the time coverage of all data sources); this represents an important limitation in our results. However, given that deplatforming is intended primarily to produce immediate and short-term effects on platform health, it is reasonable to determine whether replatforming has similarly immediate effects. Additionally, both experts and the media predicted a near-immediate negative effect of YouTube's policy change on the information environment, demonstrating that the short-term effect of replatforming is an area in need of analytical rigor and evidence (Scott 2023; Bond 2023; Fischer 2023; Ingram 2023).

The lack of short-term effects in our study also draws an interesting contrast with the deplatforming literature, which finds immediate and large reductions in the prevalence of targeted content (Buntain et al. 2023; Thomas and Wahedi 2023). This suggests that despite replatforming being the inverse of deplatforming at the policy level, any effects occur as the result of different processes. When considering the primary actors in both processes, this makes intuitive sense: deplatforming is a platform-led action designed to quickly eliminate undesirable content and behavior, while replatforming involves the platform itself stepping back, meaning that any effects of replatforming depend on the actions of users themselves. While the deplatforming literature may give us some useful theoretical expectations as to the state of users after the initial deplatforming actions, we should not expect it to provide extensive insight into user behavior in response to replatforming.

Conclusion

While deplatforming has garnered significant scholarly attention over the past decade, considerably less attention has been paid to replatforming of previously banned users and content. Replatforming remains undertheorized and understudied to the point that we lack even a comprehensive accounting of the number and nature of replatforming actions taken by platforms. As social media platforms continue attempting to balance free speech concerns with the business and societal risks of potentially harmful content, we should expect to see further deplatforming and replatforming actions.

Here we provide a systematic investigation of the effects of one such replatforming effort, YouTube's June 2nd policy change which re-allowed conspiracy content surrounding past, but not present or upcoming, elections. However, we find no evidence that the replatforming of election conspiracy content increased the short-term prevalence of election conspiracy discourse on public Facebook pages or in Telegram conspiracy groups. Additionally, we do not find increased demand for election conspiracy content using Google and YouTube search data. At least in the short term, replatforming appears to have had little impact on the broader social media ecosystem, even in the places where we would expect an impact to be most likely (Telegram far-right and conspiracy groups).

There are several limitations to this study that demonstrate the need for further research. This study is short-term in nature, following only the month after the policy change across most outcomes. It is possible that changes to the discourse may occur more gradually or due to events which trigger increased election-related discourse, namely the 2024 US Presidential Election. Given that both current US Presidential candidates also ran on the 2020 presidential tickets, we might expect that despite banning content disputing the 2024 election, re-allowing content disputing the 2020 election may still have negative effects as the election approaches. Additionally, due to the lack of pre-existing research and theory on the effects of replatforming, this work is largely deductive. A larger body of research, similar to the body of research on deplatforming, must be built before theory can stand on solid ground.

However, at present replatforming research is hamstrung by the fact that we are unaware of most replatforming incidents. Social media platforms change their policies often and frequently do so in silence; the YouTube policy change was unusual in that YouTube announced it themselves, very publicly, at the moment of implementation. Funding to build a persistently-maintained database of platform policy changes would be of great utility for the study of both deplatforming and replatforming, allowing researchers to quickly identify and study past, present, and future deplatforming events. There is also considerable room for methodological innovation in studying the long term (> 3 months) impact of replatforming. While identifying long-term effects requires dealing with a multitude of difficult confounders that multiply over time, such challenges can and have been overcome in the social sciences before. Future research on other deplatforming incidents may provide the descriptive detail neces-

sary to build such long-term impact studies.

References

Allcott, H.; and Gentzkow, M. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2): 211–236.

Arriagada, A.; and Ibáñez, F. 2020. "You need at least one picture daily, if not, you're dead": content creators and platform evolution in the social media ecology. *Social Media+ Society*, 6(3): 2056305120944624.

Baldassaro, M.; Harbath, K.; and Scholtens, M. 2021. The Big Lie and Big Tech.

Benkler, Y.; Tilton, C.; Etling, B.; Roberts, H.; Clark, J.; Faris, R.; Kaiser, J.; and Schmitt, C. 2020. Mail-in voter fraud: Anatomy of a disinformation campaign. *Berkman Center Research Publication*, (2020-6).

Boer, M.; Stevens, G. W.; Finkenauer, C.; de Looze, M. E.; and van den Eijnden, R. J. 2021. Social media use intensity, social media use problems, and mental health among adolescents: Investigating directionality and mediating processes. *Computers in Human Behavior*, 116: 106645.

Bond, S. 2023. YouTube will no longer take down false claims about U.S. elections. *NPR*.

Brodersen, K. H.; Gallusser, F.; Koehler, J.; Remy, N.; and Scott, S. L. 2015. Inferring causal impact using Bayesian structural time-series models. *Annals of Applied Statistics*, 9: 247–274.

Bryanov, K.; Vasina, D.; Pankova, Y.; and Pakholkov, V. 2021. The other side of deplatforming: Right-wing telegram in the wake of trump's twitter ouster. In *International Conference on Digital Transformation and Global Society*, 417–428. Springer.

Buntain, C.; Innes, M.; Mitts, T.; and Shapiro, J. 2023. Cross-platform reactions to the post-january 6 deplatforming. *Journal of Quantitative Description: Digital Media*, 3.

Calonico, S.; Cattaneo, M. D.; and Titiunik, R. 2015. Rdrobust: an R package for robust nonparametric inference in regression-discontinuity designs. *R J.*, 7(1): 38.

Cinelli, M.; Quattrociocchi, W.; Galeazzi, A.; Valensise, C. M.; Brugnoli, E.; Schmidt, A. L.; Zola, P.; Zollo, F.; and Scala, A. 2020. The COVID-19 social media infodemic. *Scientific reports*, 10(1): 1–10.

Duffy, B. E.; and Sawey, M. 2021. Value, service, and precarity among Instagram content creators. *Creator culture: An introduction to global social media entertainment*, 135–152.

Fink, C. 2018. Dangerous speech, anti-Muslim violence, and Facebook in Myanmar. *Journal of International Affairs*, 71(1.5): 43–52.

Fischer, S. 2023. Scoop: YouTube reverses misinformation policy to allow U.S. election denialism.

Gates, S.; and Podder, S. 2015. Social media, recruitment, allegiance and the Islamic State. *Perspectives on Terrorism*, 9(4): 107–116.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Ingram, M. 2023. The tech platforms have surrendered in the fight over election-related misinformation.

Innes, H.; and Innes, M. 2023. De-platforming disinformation: conspiracy theories and their control. *Information, Communication & Society*, 26(6): 1262–1280.

Mekacher, A.; Falkenberg, M.; and Baronchelli, A. 2023. The systemic impact of deplatforming on social media. *arXiv preprint arXiv:2303.11147*.

Mitts, T.; Pisharody, N.; and Shapiro, J. 2022. Removal of anti-vaccine content impacts social media discourse. In *Proceedings of the 14th ACM Web Science Conference 2022*, 319–326.

Rauchfleisch, A.; and Kaiser, J. 2021. Deplatforming the far-right: An analysis of YouTube and BitChute. *Available at SSRN 3867818*.

Roberts, M. E.; Stewart, B. M.; Tingley, D.; Lucas, C.; Leder-Luis, J.; Gadarian, S. K.; Albertson, B.; and Rand, D. G. 2014. Structural topic models for open-ended survey responses. *American journal of political science*, 58(4): 1064–1082.

Scott, L. 2023. Experts Divided as YouTube Reverses Policy on Election Misinformation.

Thomas, D. R.; and Wahedi, L. A. 2023. Disrupting hate: The effect of deplatforming hate organizations on their online audience. *Proceedings of the National Academy of Sciences*, 120(24): e2214080120.

Wojcieszak, M.; Casas, A.; Yu, X.; Nagler, J.; and Tucker, J. A. 2022. Most users do not follow political elites on Twitter; those who do show overwhelming preferences for ideological congruity. *Science advances*, 8(39): eabn9418.

YouTube. 2023. An update on our approach to US election misinformation. https://blog.youtube/inside-youtube/us-election-misinformation-update-2023/. Accessed: 2024-05-14.

YouTube. N.d. Elections misinformation policies. https://support.google.com/youtube/answer/10835034?hl=en&ref_topic=10833358&sjid=4002580615044614685-NA. Accessed: 2024-05-14.