SOCIAL SCIENCES

Subscriptions and external links help drive resentful users to alternative and extremist YouTube channels

Annie Y. Chen[1], Brendan Nyhan[2]*, Jason Reifler[3], Ronald E. Robertson[4,5], Christo Wilson[5]

Do online platforms facilitate the consumption of potentially harmful content? Using paired behavioral and survey data provided by participants recruited from a representative sample in 2020 (*n* = 1181), we show that exposure to alternative and extremist channel videos on YouTube is heavily concentrated among a small group of people with high prior levels of gender and racial resentment. These viewers often subscribe to these channels (prompting recommendations to their videos) and follow external links to them. In contrast, nonsubscribers rarely see or follow recommendations to videos from these channels. Our findings suggest that YouTube's algorithms were not sending people down "rabbit holes" during our observation window in 2020, possibly due to changes that the company made to its recommender system in 2019. However, the platform continues to play a key role in facilitating exposure to content from alternative and extremist channels among dedicated audiences.

INTRODUCTION

What role do technology platforms play in exposing people to dubious and hateful information and enabling its spread? Concerns have grown in recent years that online communication is exacerbating the human tendency to engage in preferential exposure to congenial information (*1*–*3*). These concerns are particularly acute on social media, where people may be especially likely to view content about topics such as politics and health that is false, extremist, or, otherwise, potentially harmful. The use of algorithmic recommendations and platform affordances such as following and subscribing features may enable this process by helping people to find potentially harmful content and helping content creators build and monetize an audience for it.

These concerns are particularly pronounced for YouTube, the most widely used social media platform in the United States (*4*). Critics highlight the popularity of extreme and harmful content such as videos by white nationalists on YouTube, which they often attribute to the recommendation system that the company itself says is responsible for 70% of user watch time (*5*). Many fear that these algorithmic recommendations are an engine for radicalization. For instance, the sociologist Tufekci (*6*) wrote that the YouTube recommendation system "may be one of the most powerful radicalizing instruments of the 21st century". These claims seem to be supported by reports that feature descriptions of recommendations to potentially harmful videos and accounts of people whose lives were upended by content they encountered online (*7*–*9*).

YouTube subsequently announced changes in 2019 to "reduce the spread of content that comes close to—but does not quite cross the line of—violating our Community Guidelines" (*10*). It claimed that these interventions resulted in a 50% drop in watch time from recommendations for "borderline content and harmful misinformation" (*11*) and a 70% decline in watch time from nonsubscribed recommendations (*12*). However, these claims have not been independently evaluated using behavioral data, nor have the implications or caveats of "nonsubscribed recommendations" been sufficiently explored.

In general, questions remain about the size and composition of the audience for potentially harmful videos on YouTube following these changes, the manner in which people reach those videos, and the role of the recommendation system in that process. Studies show that sites such as Twitter and Facebook can amplify tendencies toward extreme opinions or spread false information (*13*, *14*), although the extent of these effects and the prevalence of exposure are often overstated (*15*–*17*). YouTube may operate differently, though, given its focus on video and the central role of its recommendation system (*18*, *19*). Browsing data have documented the existence of a sizeable audience of dedicated far-right news consumers on YouTube who often reach extremist videos via external links (*20*), but these data lack information about the recommendations shown to users by YouTube or the channels the users follow (a key source of recommendations). Random walk simulations conducted during and after 2019 found that problematic content was reachable, but its prevalence in recommendations fell during this period (*21*). Research conducted after 2019 found that watching videos promoting misinformation still led to recommendations of similar videos on some topics, although their overall prevalence among recommendations was low (*22*–*25*). We build on these studies, seeking to determine the extent to which YouTube's goal of "reduc[ing] recommendations of borderline content and harmful misinformation" has been met using a distinct measurement approach (*12*).

This study advances scientific understanding of the audience for potentially harmful content on YouTube and the manner in which people are exposed to it. We pair individual-level viewer histories and the associated video recommendations shown with survey data from a sample of 1181 U.S. respondents who were weighted to resemble the U.S. adult population on key demographic traits. This research design allows us to examine the association between demographic and attitudinal variables, especially gender and racial resentment, and YouTube consumption behavior. Using these data,

[1]CUNY Institute for State & Local Governance, 10 East 34th St., New York, NY 10016, USA. [2]Dartmouth College, 6108 Hinman, Hanover, NH 03755, USA. [3]University of Exeter, Stocker Road, Exeter EX4 4PY, UK. [4]Stanford Internet Observatory, Stanford University, 616 Jane Stanford Way, Stanford, CA 94305, USA. [5]Northeastern University, 360 Huntington Ave, Boston, MA 02115, USA.
*Corresponding author. Email: nyhan@dartmouth.edu

we address three limitations of prior research in the field. First, prior work has not taken YouTube users' channel subscriptions into account, a key indicator of user demand for specific types of content and a major factor in what recommendations are shown to users. We address this point by inferring the channels that our participants subscribe to and stratifying our analysis of recommendations along this axis. Second, existing work has either relied on data from controlled experiments and random walks—which lack ecological validity—or browsing histories that lack data on video recommendations. Our dataset offers the ecological validity that comes from directly observing user behavior on YouTube, providing the first direct evidence of the extent to which real-world algorithmic recommendations push people toward potentially harmful content. Third, prior fears about the frequency of "rabbit holes" are based on anecdotes and lack a precise definition. We address this problem by constructing a specific set of rules to define a rabbit hole event. This definition builds on and reaffirms prior work (*21*, *24*–*26*), and we applied it to our dataset to measure the prevalence of radicalization rabbit holes among U.S. YouTube users in 2020.

Our sample of 1181 participants is recruited from a sample of 4000 YouGov panelists, including oversamples of two groups whom we identified as especially likely to be exposed to potentially harmful video content: (i) people who previously expressed high levels of gender and/or racial resentment and (ii) those who indicated they used YouTube frequently. Participants voluntarily agreed to install a custom browser extension in Chrome or Firefox that monitored their web browsing behavior. The study was conducted from 21 July to 31 December 2020 (i.e., after the 2019 changes to YouTube's algorithm); respondents were enrolled in data collection for a median of 133 days. (See Materials and Methods below for further details on measurement. We provide descriptive statistics on study participants and their browser activity data availability and aggregate consumption patterns in the Supplementary Materials.)

We report two key findings. First, we replicate findings from Hosseinmardi *et al.* (*20*) concerning the overall size of the audience for alternative and extreme content and enhance their validity by examining participants' attitudinal variables. Although almost all participants use YouTube, videos from alternative and extremist channels are overwhelmingly watched by a small minority of participants with high levels of gender and racial resentment. Within this group, total viewership is heavily concentrated among a few individuals, a common finding among studies examining potentially harmful online content (*27*). Similar to prior work (*20*), we observe that viewers often reach these videos via external links (e.g., from other social media platforms). In addition, we find that viewers are often subscribers to the channels in question. These findings demonstrate the scientific contribution made by our study. They also highlight that YouTube remains a key hosting provider for alternative and extremist channels, helping them continue to profit from their audience (*28*, *29*) and reinforcing concerns about lax content moderation on the platform (*30*).

Second, we investigate the prevalence of rabbit holes in YouTube's recommendations during the fall of 2020. We rarely observe recommendations to alternative or extremist channel videos being shown to, or followed by, nonsubscribers. During our study period, only 3% of participants who were not already subscribed to alternative or extremist channels viewed a video from one of these channels based on a recommendation. On one hand, this finding suggests that unsolicited exposure to potentially harmful content on YouTube in the post-2019 era is rare, in line with findings from prior work (*24*, *25*). On the other hand, even low levels of algorithmic amplification can have damaging consequences when extrapolated over YouTube's vast user base and across time (*20*). Further, it may be the case that the susceptible population was already radicalized during YouTube's pre-2019 era. Last, given the limitations of our study, our results must be interpreted as a lower bound on rabbit hole events, which suggests that YouTube may still need to do more to remove "borderline" content from recommendations.

RESULTS

Study participants completed a public opinion survey and installed a browser extension that recorded their browser activity (n = 1181; see Materials and Methods for details on sampling and recruitment). The browser extension passively logged user page views, including the full uniform resource locator (URL) and a timestamp and collected hypertext markup language (HTML) snapshots when users viewed YouTube videos, allowing us to examine the video recommendations that participants received. This combination of passive monitoring and HTML snapshots provides us with the ability to measure not only what respondents watched but also what YouTube showed them before that action.

Exposure levels

Although 91% [95% confidence interval (CI), 89.6 to 92.8] of study participants visited YouTube, the vast majority did not view any alternative or extremist channel videos. Only 15.4% (95% CI, 13.4 to 17.5) of the sample for whom we have browser activity data (n = 1181) viewed any video from an alternative channel, and only 6.1% (95% CI, 4.8 to 7.5) viewed any video from an extremist channel. By comparison, 43.5% (95% CI, 40.7 to 46.3) viewed at least one video from a mainstream media channel. (See Materials and Methods for how channel types were defined and how view history and watch time were defined.) Videos from mainstream media channels account for 3.57% (95% CI, 3.54 to 3.60) of videos watched in our sample—a figure that falls between recent estimates that 2.9 to 11% of videos watched on YouTube are news (*20*, *31*). The corresponding numbers for videos from alternative and extremist channels are 2.96% (95% CI, 2.93 to 2.99) and 0.51% (95% CI, 0.50 to 0.52), respectively [similar to estimates from 2019 (*20*)].

The audience for alternative and extremist channels is skewed toward people who subscribe to the channel in question or one like it, which we determine by inspecting whether the subscription button is activated when a participant views a video from that channel (see Materials and Methods for more details). Among the set of people who saw at least one extremist channel video during the study period, for instance, 51.7% (95% CI, 39.7 to 63.6) watched a video from an extremist channel to which they subscribed. Similarly, 39.0% (95% CI, 31.6 to 46.4) of alternative channel viewers watched at least one video from an alternative channel to which they subscribed.

Figure 1 illustrates this point in a different way by disaggregating video views according to both channel type and subscription status. We observe that 60.8% (95% CI, 60.2 to 61.5) of views for videos from alternative channels and 54.7% (95% CI, 53.1 to 56.2) of

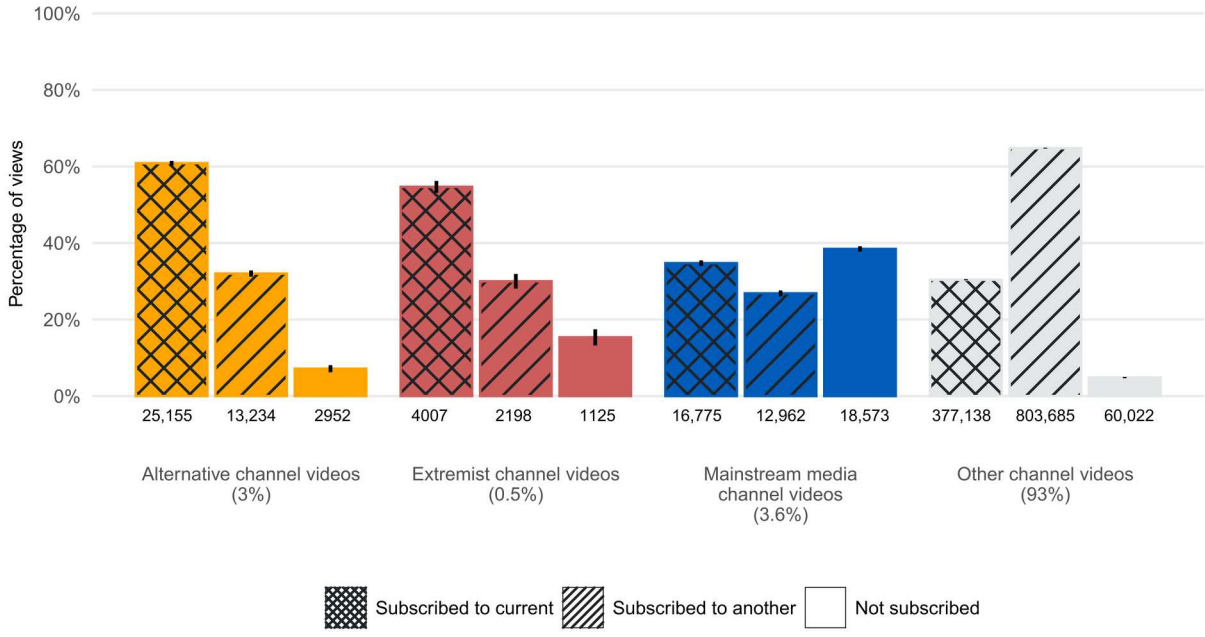

Fig. 1. Distribution of video views by subscription status and channel type. Weighted percentages of views for videos from each type of channel that come from people who are subscribed to that channel (crosshatches), who subscribe to one or more different channels of the same type but not the channel currently being viewed (hatches), and who do not subscribe to any channel of that type (no hatches). Each estimate includes the corresponding 95% CI (unweighted N = 1,097,849). Total view counts are displayed at the bottom of each bar. Total views for videos of that type as a percentage of all views are displayed under the channel labels.

views for videos from extremist channels come from subscribers to the channel in question. If we instead define subscribers to include all people who subscribe to at least one channel of the type in question, then the proportion of views from subscribers increases to 92.9% (95% CI, 92.6 to 93.1) for alternative channels and 84.7% (95% CI, 83.8 to 85.5) for extremist channels. These patterns for alternative and extremist channels are distinct from mainstream media channels, which receive 38.4% (95% CI, 37.7 to 39.1) of their views from people who do not subscribe to any channel in the category.

Among participants who viewed at least one video from either an alternative channel or an extremist channel, the time spent watching videos of that type was relatively low and concentrated among subscribers. Mean time spent watching alternative videos among people who viewed at least one video from an alternative channel was 25.7 (95% CI, 13.6 to 37.8) min/week, with means of 62.2 (95% CI, 33.2 to 91.3) min/week for subscribers to one or more alternative channels and 0.2 (95% CI, 0.1 to 0.4) min/week for nonsubscribers. Similarly, mean time spent watching extremist videos among participants who viewed at least one video from an extremist channel was 8.1 (95% CI, 3.5 to 12.7) min/week for extremist channel videos, which was divided between 14.6 (95% CI, 5.1 to 24.0) min/week for subscribers and 0.04 (95% CI, 0.01 to 0.07) min/week for nonsubscribers. The comparison statistics are 11.9 (95% CI, 7.3 to 16.5) min/week for mainstream media channel videos and 214.2 (95% CI, 169.2 to 259.2) min/week for videos from other channels. As noted above, however, these data are highly skewed: The median time spent watching was 1.1 (95% CI, 0.4 to 3.5) min/week for alternative channel videos among alternative channel video viewers and 0.6 (95% CI, 0.2 to 5.5) min/week for extremist channel videos among extremist channel video viewers. These results mirror those from Hosseinmardi *et al.* (*20*), who observed the same ordering, in terms of video watch time, for anti-woke (i.e., alternative), far right (i.e., extreme), and mainstream news sources, from most to least watched.

Viewership of potentially harmful videos on YouTube is heavily concentrated among a few participants, mirroring patterns observed on YouTube over the 2016–2019 time frame (*20*), Twitter and untrustworthy websites (*32*, *33*), and news content generally (*31*, *34*). As Fig. 2 indicates, 1.7% (95% CI, 0.0 to 5.6) of participants account for 80% of total time spent on videos from alternative channels. This imbalance is even more severe for extremist channels, where 0.6% (95% CI, 0.0 to 4.6) of participants were responsible for 80% of total time spent on these videos. Skew is similar when we examine view counts (fig. S16) rather than time spent on videos—1.9% (95% CI, 0.0 to 5.8) and 1.1% (95% CI, 0.0 to 5.0) of participants were responsible for 80% of alternative and extremist channel viewership, respectively. We observe a similar pattern of concentration for mainstream media consumption—only 3.8% (95% CI, 0.0 to 7.7) of participants account for 80% of the total views. (We provide a more detailed analysis of the viewership patterns of these "superconsumers" in the Supplementary Materials.)

Correlates of exposure

We next evaluate demographic and attitudinal factors that are potentially correlated with time spent watching videos from alternative, extremist, and mainstream media channels. We focus specifically on hostile sexism, racial resentment, and negative feelings toward Jews—three factors that may make people vulnerable to the types of messages offered by alternative and extremist channels, which often target women, racial and ethnic minorities, and Jews (*35*, *36*). Negative attitudes toward these outgroups may make

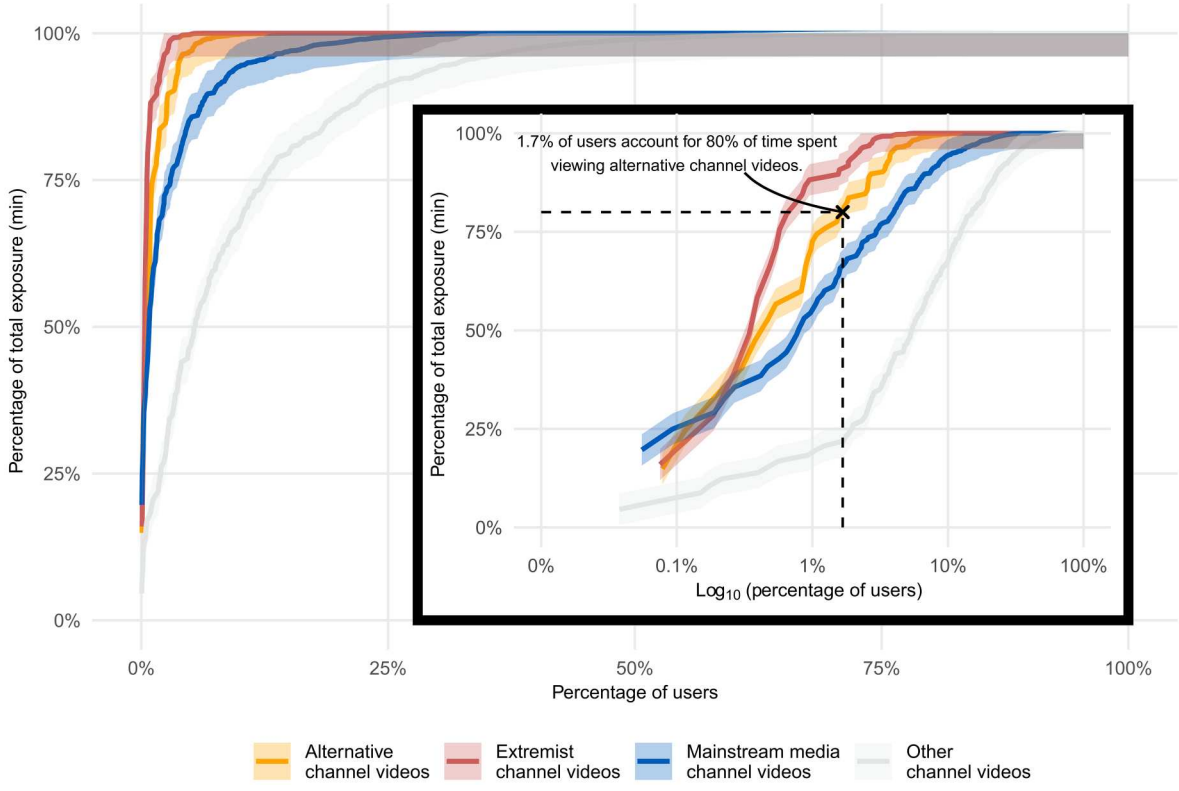

Fig. 2. Concentration of exposure to alternative and extremist channels. Weighted empirical cumulative distribution function showing the percentage of participants responsible for a given level of total observed video viewership of alternative and extremist channels on YouTube (in minutes). Each estimate includes the corresponding 95% CI (unweighted N = 1181). Inset graph shows the same data using a log scale for the weighted empirical cumulative distribution function.

people vulnerable to the types of messages offered by alternative and extremist channels. We therefore estimate the statistical models reported below on the subset of 851 respondents for whom prior scale measures of hostile sexism and racial resentment are available from the 2018 Cooperative Congressional Election Study. (Details on survey wording and measurement, including the wording for these scales, are provided in Materials and Methods below; feelings toward Jews are measured using a feeling thermometer.)

We estimate models measuring the association between the average time per week that respondents spent on videos from alternative, extremist, or mainstream media channels and the measures listed above as well as relevant demographic characteristics: age, sex (male or not male), race (white or non-white), and indicators for different levels of education above high school (some college, bachelor's, or post-grad). Results of the quasi-poisson models that we estimate, which account for the skew in video watch time, are shown in Fig. 3. (See fig. S9 for equivalent results for the number of views of videos from alternative and extremist channels.)

The results indicate that prior levels of hostile sexism are significantly associated with time spent on videos from alternative channels (b = 1.7; 95% CI, 1.0 to 2.4) and time spent on videos from extremist channels (b = 1.6; 95% CI, 0.4 to 2.8) but not time spent watching mainstream media channels (b = 0.0; 95% CI, −0.6 to 0.6). This relationship, which is consistent with the commenter overlap observed between men's rights antifeminist channels and alt-right channels on YouTube (*37*), is not observed for prior levels of racial resentment when controlling for hostile sexism. However, both hostile sexism and racial resentment are positively associated with time spent on videos and number of views of videos from alternative and extremist channels when entered into statistical models separately (see tables S6 and S7). Last, we find no association between feelings toward Jews and viewership of any of these types of channels.

Figure 4 illustrates the relationship between prior levels of hostile sexism and time spent per week watching videos from alternative or extremist channels using the model results described above. When hostile sexism is at its minimum value of 1, expected levels are 0.4 [95% prediction interval (PI), 0.1 to 2.8] min/week spent watching alternative channel videos and 0.08 (95% PI, 0.002 to 3.077) min/week for extremist channel videos. These predicted values increase to 383.0 (95% PI, 75.9 to 1933.1) and 51.0 (95% PI, 7.4 to 353.0) min/week, respectively, when hostile sexism is at its maximum value of 5 (with the greatest marginal increases as hostile sexism reaches its highest levels).

Recommendations and YouTube rabbit holes

Critics of YouTube have emphasized the role of its algorithmic recommendations in leading people to potentially harmful content. We therefore measure which types of videos YouTube recommended to participants and how often those recommendations were followed. Next, we specifically count how often people follow recommendations to more extreme channels to which they do not subscribe in a manner that is consistent with the rabbit hole narrative. Last, we disaggregate YouTube recommendations and

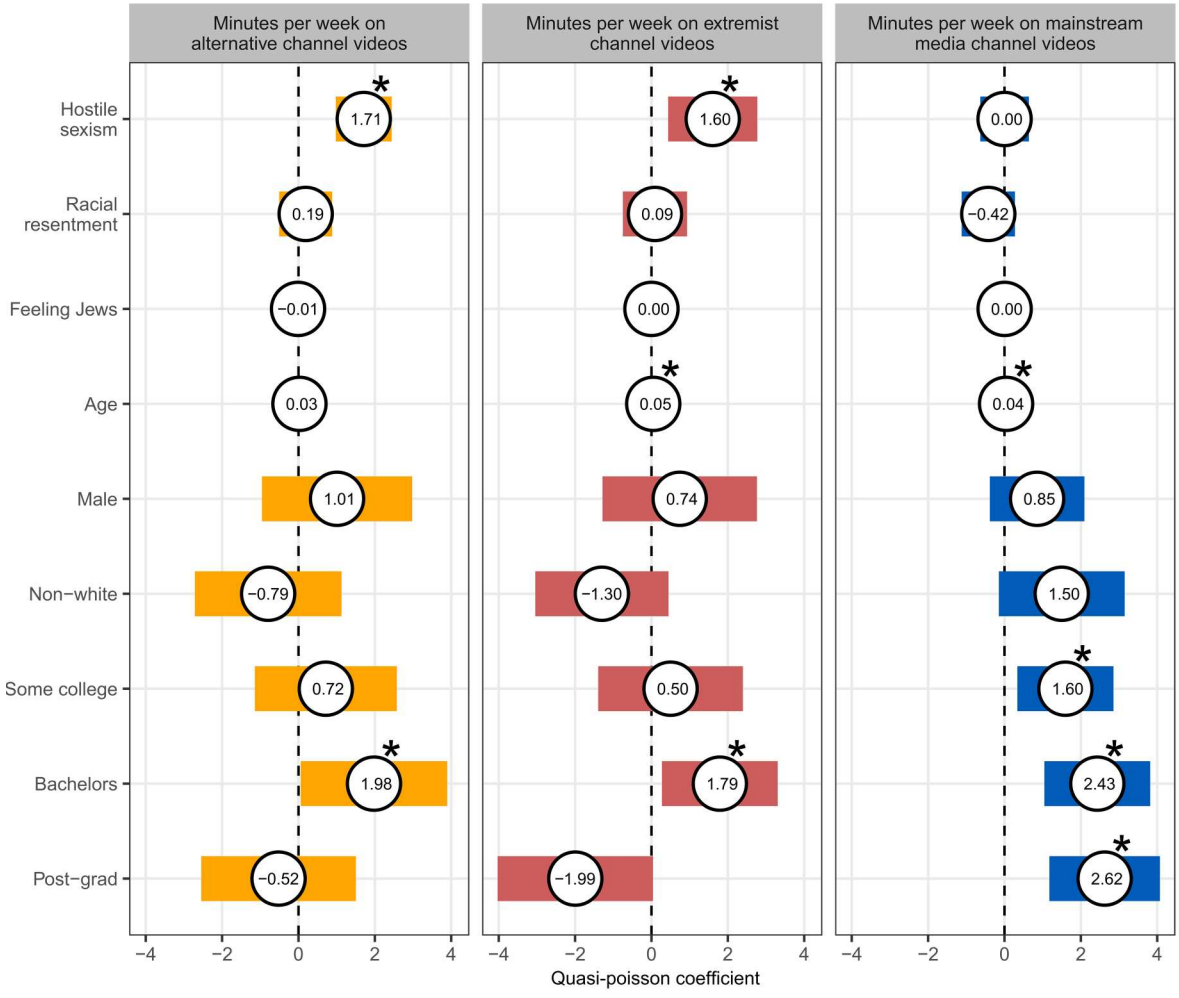

Fig. 3. Predictors of video watch time. Quasi-poisson regression coefficients for correlates of the amount of time respondents spent on videos from alternative, extremist, and mainstream media channels in minutes per week. Figure includes 95% CIs calculated from robust SEs (unweighted $N = 851$). All results incorporate survey weights. Asterisks indicate coefficients that are significant at the $P < 0.05$ level. See table S2 for regression table.

following behavior based on subscription status. In general, we find that recommendations to alternative and extremist channel videos are rare and frequently shown to and followed by people who already subscribe to those channels.

We disaggregate the recommendations shown to participants by the type of video on which the recommendation appears, which appears to play a large role in determining what YouTube recommends. As Fig. 5A shows, there are relatively few recommendations to alternative and extremist videos. As Fig. 5B shows, recommendations to alternative and extremist channel videos are very rare when watching videos from mainstream media or other types of channels, which, together, make up 96.4% (95% CI, 96.3 to 96.4) of views in our sample. Recommendations to alternative and extremist channel videos are much more common, however, when people are already viewing videos from alternative and extremist channels, which make up 2.96% (95% CI, 2.93 to 2.99) and 0.51% (95% CI, 0.50 to 0.52) of views, respectively. Just under half (47.9%; 95% CI, 47.6 to 48.3) of recommendations when viewing an alternative channel video point to another alternative channel video, while 41.1% (95% CI, 40.3 to 41.8) of recommendations follow the same pattern for extremist channel videos. Substantively similar patterns of recommendations have been observed in random walk studies on YouTube (*24*, *26*).

Figure S6 provides corresponding statistics for the proportion of recommendations followed by channel type. As expected, the people who are already watching alternative and extremist channel videos are especially likely to follow recommendations to other alternative or extremist channel videos. Among people who were watching alternative channel videos, 53.7% (95% CI, 49.1 to 58.4) of recommendations followed were to alternative or extremist channel videos [compared to 50.3% (95% CI, 50.0 to 50.6) of recommendations shown]. Correspondingly, 73.8% (95% CI, 67.7 to 79.9) of recommendations followed from extremist channel videos were to other extremist or alternative channel videos [versus 54.3% (95% CI, 53.6 to 54.9) of recommendations shown]. The probability of following a recommendation to such a video by people not already watching an alternative or extremist channel video was negligible. (We disaggregate recommendations and follows by recommendation rank in figs. S17 and S18.)

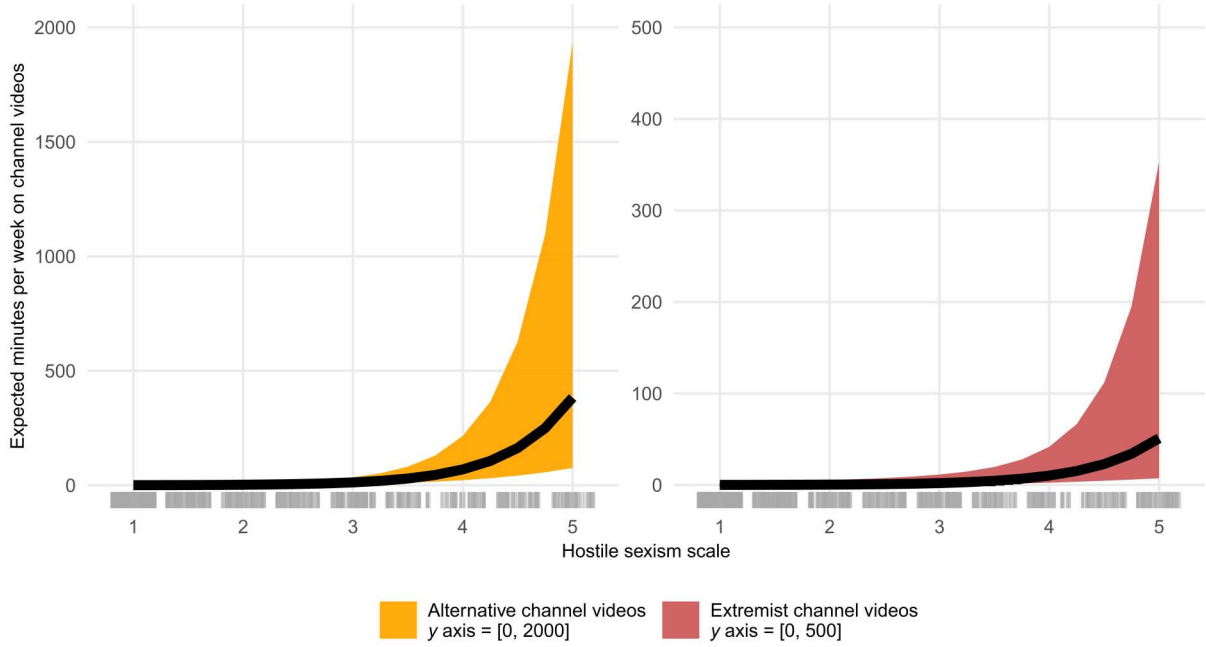

Fig. 4. Hostile sexism as predictor of alternative and extremist channel viewing. Predictions are estimated from the quasi-poisson regression models in Fig. 3 holding other covariates at their median (continuous variables) and modal (categorical variables) values. Colored bands represent 95% robust CIs. All results incorporate survey weights.

Next, we more directly test how often YouTube video recommendations create rabbit holes in which people are shown more extreme content than they would otherwise encounter. Specifically, we define four conditions that must be met to constitute a rabbit hole and report how often these criteria are met when applied sequentially:

1) A participant followed a recommendation to an alternative or extremist channel video: 0.17% (95% CI, 0.16 to 0.18) of all video visits among 7.3% (95% CI, 5.7 to 8.8) of participants;

2) The recommendation that the participant followed moved them to a more extreme channel type (i.e., {mainstream media, other} → {alternative} or {mainstream media, other, alternative} → {extreme}): 0.07% (95% CI, 0.06 to 0.08) of all video visits among 5.4% (95% CI, 4.1 to 6.8) of participants;

3) The participant does not subscribe to the channel of the recommended video: 0.02% (95% CI, 0.016 to 0.023) of all video visits among 4.7% (95% CI, 3.4 to 6.0) of participants;

4) The participant does not subscribe to any channels of the same type (i.e., alternative or extremist) as the recommended video: 0.01% (95% CI, 0.007 to 0.011) of all video visits among only 3.0% (95% CI, 2.0 to 4.0) of participants.

On the basis of these strict criteria, we observe very few cases of rabbit hole events. As noted above, the set of events that meet all four criteria for alternative and extremist channel videos represents only 0.01% of all video visits and was observed among only 3.0% of participants. The set of these sequences that specifically ended in exposure to an extremist channel video represented only 0.002% (95% CI, 0.001 to 0.003) of all visits and was only observed among 1.0% (95% CI, 0.3 to 1.8) of participants. (We provide qualitative accounts of three such sequences in the Supplementary Materials and an analysis showing no trend toward greater exposure to alternative or extremist channel videos in longer YouTube sessions.)

We observe that recommendations to videos from alternative and extremist channels are frequently shown to channel subscribers—the same group that is most likely to follow those recommendations. As Fig. 6 demonstrates, people who subscribe to at least one alternative channel received 53.1% (95% CI, 52.9 to 53.3) of all alternative channel video recommendations and represented 67.2% (95% CI, 63.9 to 70.5) of the cases in which a participant followed a recommendation to an alternative channel video. This skew was somewhat smaller for extremist channel videos—subscribers to one or more extremist channels saw 44.7% (95% CI, 44.1 to 45.2) of recommendations to videos from extremist channels and made up 49.0% (95% CI, 43.2 to 54.9) of the cases in which respondents followed a recommendation to watch such a video. These figures are generally larger than those observed for mainstream media channels or other types of channels.

Internal and external referrers

Last, we replicate and expand an analysis conducted by Hosseinmardi *et al.* (*20*) that measures the process by which people come to watch alternative and extremist videos on YouTube. As in prior work, we denote the page that people viewed immediately before a video being opened (within an existing browser tab or within a new tab) as the "referrer" and distinguish between "on-platform" referrers (a YouTube channel page, the YouTube homepage, a YouTube search page, or another YouTube video) and "off-platform" referrers that are not part of the YouTube domain such as search engines, webmail sites, mainstream social media sites (e.g., Facebook, Twitter, and Reddit), or alternative social media sites (e.g., Parler, Gab, and 4chan). The complete list of external referrers in each category can be found in table S10. Details on how we identify referrers are provided in Materials and Methods below.

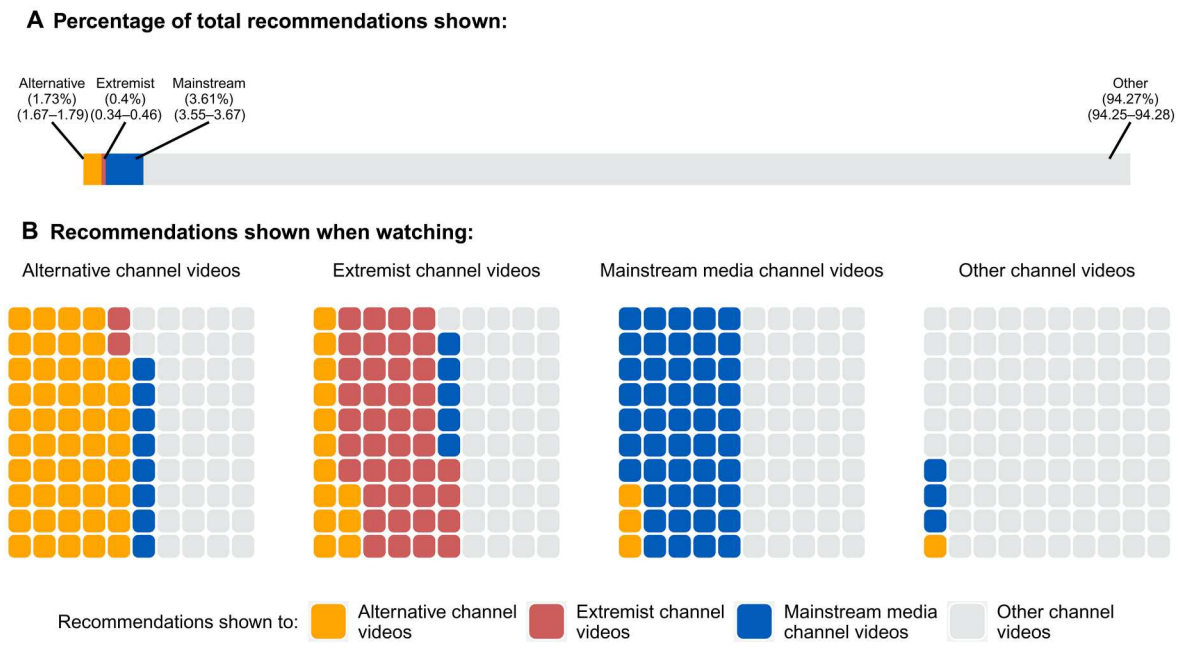

Fig. 5. Recommendation frequency by type of channel being watched. Number of colored tiles shown are proportional to the proportion of recommendations shown for each type of video when watching videos from alternative, extremist, mainstream media, or other channels. (**A**) summarizes the percentage of recommendations shown by channel type and (**B**) disaggregates the recommendations shown by channel type for the video being watched. Results are based on the full set of recommendations that we could extract from each video and incorporate survey weights. Each category estimate includes the corresponding 95% CI (unweighted N = 8,303,137).

We find that off-platform referrers are responsible for approximately half of all views of alternative and extremist channel videos, a finding that is roughly consistent with YouTube's statement that "borderline content gets most of its views from other platforms that link to YouTube" (*38*). Our finding is slightly higher than the 36 to 41% external referrers for alternative and extreme videos observed by Hosseinmardi *et al.* (*20*), but we include referrals from non-YouTube search engines in our total, while Hosseinmardi *et al.* (*20*) do not. That said, as we show in Fig. 7, 52.4% (95% CI, 51.9 to 52.9) and 46.6% (95% CI, 45.2 to 47.7) of referrals to alternative and extremist channel videos, respectively, were from off-platform sources, which is only somewhat higher than off-platform referrals for videos from mainstream media (41.7%; 95% CI, 41.3 to 42.2) or other channels (41.1%; 95% CI, 41.0 to 41.2).

With respect to on-platform referrers, we observe frequent within-category referrals by video type, with 19.6% (95% CI, 19.2 to 20.0) of referrals to alternative channel videos coming from other alternative channel videos, 21.3% (95% CI, 20.3 to 22.4) of referrals to extremist channel videos coming from other extremist channel videos, and 25.6% (95% CI, 25.2 to 26.0) of referrals to mainstream media channel videos coming from other mainstream media channel videos. This is broadly consistent with results from random walk studies on YouTube that have examined recommendations between different types of videos (*24*, *26*). We observe 3.8% (95% CI, 3.3 to 4.3) of referrals to extremist channel videos coming from alternative channel videos, but only 0.8% (95% CI, 0.7 to 0.9) of referrals to alternative channel videos coming from extremist channel videos, which suggests that it is rare for our participants to move from more to less extreme content in this manner. Last, we observe that alternative, extremist, and mainstream media channel videos all receive roughly equal referrals from videos in other channels (10.0 to 12.8%) and other on-platform sources (15.1 to 19.1%). Overall, these results are also broadly similar to those Hosseinmardi *et al.* (*20*), who found that 36 to 39% of referrals to alternative and extreme videos came from other videos, while 21 to 23% of referrals came from other on-platform sources.

Figure 7 reports the proportion of views to each type of YouTube channel video (alternative, extremist, mainstream media, and other) from each type of referrer. This analysis allows us to determine which types of referrers are unusually (un)common across channel types. On-platform, we note that the YouTube homepage, YouTube search, and other YouTube videos are relatively less frequent sources of referrals to alternative and extremist channel videos than videos from mainstream media channels and other channels. In contrast, channel pages are a more common referral source to alternative and extremist channel videos. Similar to quantitatively similar findings by Hosseinmardi *et al.* (*20*), this highlights that participants arrive at alternative and extremist videos from a variety of referrers, not only YouTube recommendations.

Among off-platform referrers, social media platforms stand out as playing an especially important role in referring people to alternative and extremist channel videos. Participants are disproportionately more likely to reach alternative channel videos via mainstream social media sites and to reach extremist channel videos via alternative social media sites compared with videos from other types of channels. For instance, 9.3% (95% CI, 8.6 to 10.0) of extremist channel video views were preceded by a visit to an alternative social media site despite their limited reach. Platforms such as Gab and 4chan may attract extremist users in part due to their lax content moderation policies. These results supplement those from

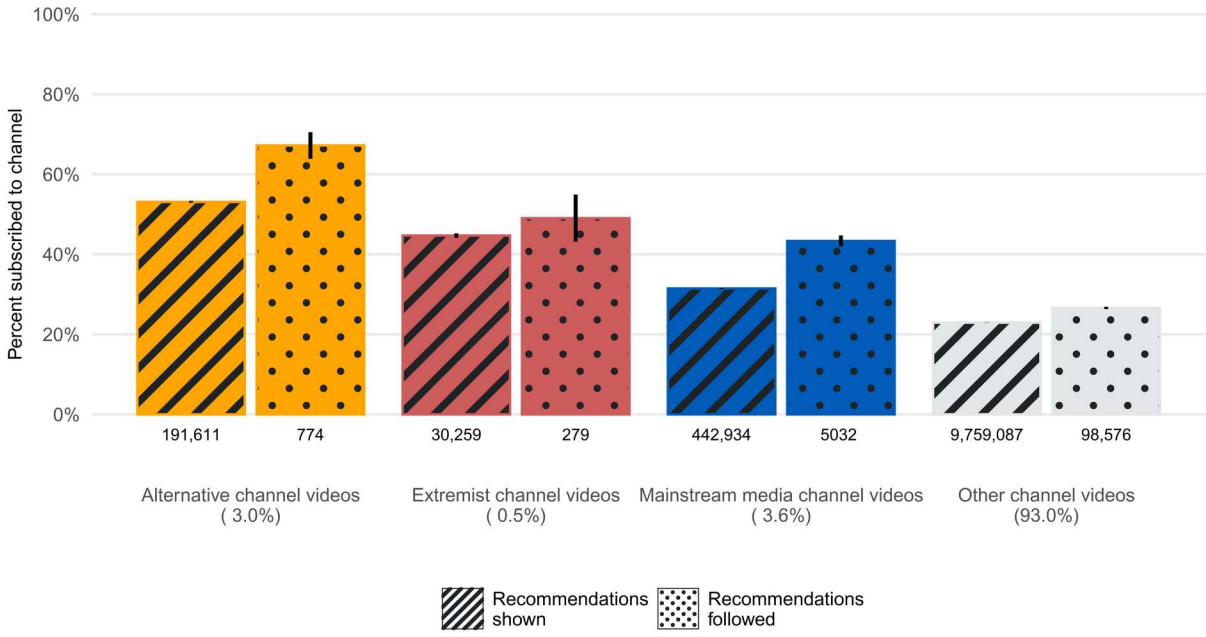

Fig. 6. YouTube recommendations by subscription status and channel type. The weighted percentage of recommendations shown and followed to people who subscribe to one or more channels of each type [including 95% CIs for both, although these are sometimes not visible because of the sample size of the recommendations shown data (unweighted N = 8,303,137)]. The weighted percentage of views of each type of video is shown in parentheses under the labels.

Hosseinmardi *et al.* (*20*), who found that alternative and extreme news websites generated many of the off-platform referrals to the corresponding types of videos on YouTube.

DISCUSSION

Using web browsing data collected in 2020, we provide behavioral measures of exposure to videos from alternative and extremist channels on YouTube. These data enable us to measure exposure to potentially harmful content on the platform and to analyze the role of YouTube's algorithms in facilitating exposure to that content after reported changes to the recommendation system in 2019.

Our data indicate that many alternative and extremist channels remain on the platform and attract a small but active audience of individuals who expressed high levels of hostile sexism and racial resentment in survey data collected in 2018. These participants frequently subscribe to the channels in question, generating more frequent recommendations. By continuing to host these channels, YouTube facilitates the growth of problematic communities (many channel views originate in referrals from alternative social media platforms where users with high levels of gender and racial resentment may congregate) and enables creators of alternative and extreme content to profit from shared YouTube advertising revenue or indirectly via affiliated stores and donation campaigns (*28*, *29*).

In the data we collected in 2020, YouTube's recommendation algorithm plays a secondary role in facilitating exposure to potentially harmful content. We observe that recommendations to videos from alternative and extreme channels are far more common when people are already watching those videos or subscribed to those channels relative to videos from mainstream news and non-news channels. We also observe that people rarely follow recommendations to videos from alternative and extreme channels when they are watching videos from mainstream news and non-news channels.

These results have two key implications for future research. Methodologically, our results highlight the importance of jointly measuring what people see and do on platforms rather than just one side of those interactions. In practice, these quantities can diverge quite markedly (*39*). Substantively, our results indicate that research into human behavior on social media platforms should devote greater attention to the often dominant role of small minorities of people with extreme views in the audiences for potentially harmful content.

While these results complicate the narrative of pervasive radicalization via rabbit holes on YouTube, our study does not imply that there never was a radicalization problem on YouTube or that the status quo is normatively unproblematic. Our data do not allow us to evaluate the previous state of the platform; YouTube's algorithms may have recommended videos from alternative and extremist channels more frequently before the changes made in 2019. Furthermore, given the limitations of our study (see below), our findings should be interpreted as estimating lower bounds on rabbit hole exposures in 2020 on YouTube. In addition, even very low rates of rabbit hole recommendations may be enough to expose large numbers of vulnerable people to harm, especially when extrapolated over YouTube's entire viewership and over the course of years.

It is important to note several other limitations of the study:

1) Although our browser extension sample is large and diverse and we weight our results to national benchmarks, it is not fully representative and does not capture YouTube consumption among users of browsers other than Chrome and Firefox or on mobile

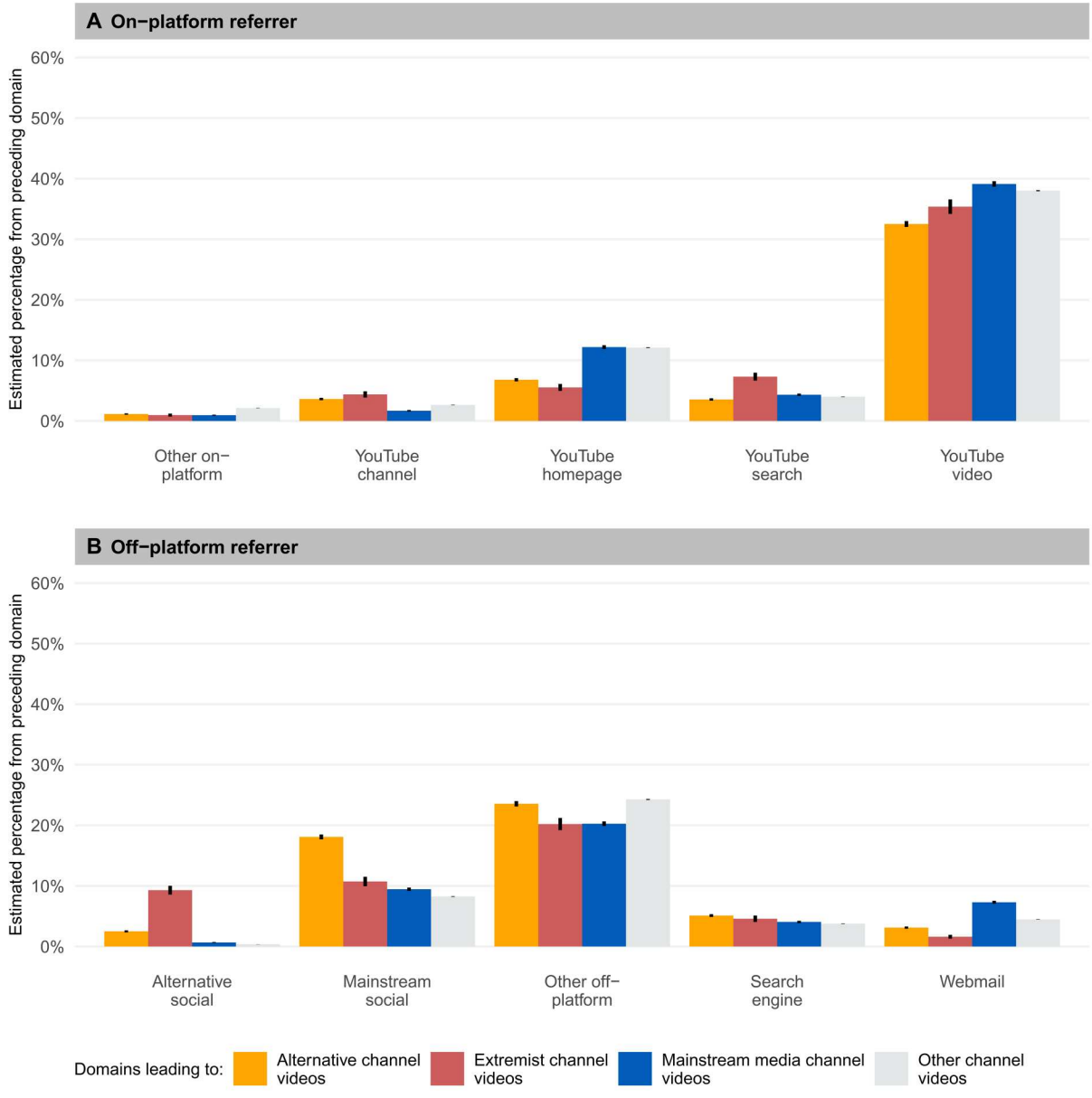

Fig. 7. Relative frequency of referrals to YouTube videos by channel and referrer type. Weighted proportion of referrals to YouTube videos of each channel type by referrer type for on-platform referrers (**A**) and off-platform referrers (**B**). All data include 95% CIs, though these are sometimes not visible because of the sample size of the referrals data (unweighted $N = 1{,}013{,}692$)]. Other on-platform platform referrals such as YouTube playlists and personal user pages were grouped into a separate category. Similarly, off-platform domains that do not fit into any of the labeled categories in (B) are grouped together. A list of all domains included in each group can be found in the Supplementary Materials.

devices. Any outside study of a platform also faces challenges in recruiting large numbers of heavy consumers of fringe content.

2) YouTube users who were susceptible to potentially harmful content may have already suffered from its effects before changes to the platform's algorithms in 2019. We are therefore unable to make causal claims based on our data—participant's preexisting gender and racial resentment may have caused them to seek out congruent content on YouTube, but, in some cases, YouTube's algorithmic recommendations may have introduced them to such content and increased feelings of resentment even before our prior survey measures of hostile sexism and racial resentment were recorded (November/December 2018). Exposure to YouTube's algorithms before the changes in 2019 could also reduce our ability to detect new rabbit hole events during the study period in 2020 as some people who are likely to follow problematic recommendations might already be subscribed to these types of channels.

3) Our results only cover U.S. users; they should be replicated outside the United States in contexts including Europe and the global South (and with non-English language content).

4) Our results depend on channel-level classifications from scholars and subject matter experts; further research should

examine whether the patterns we observe are robust to alternate measures at the channel and (if possible) video level.

5) Our measures of views, referrals, and subscriptions contain some degree of error. In particular, as with most passive behavioral data, we cannot verify that every user paid attention to the content that appeared on their device in every instance.

Nonetheless, these results underscore the need to apply the tools of behavioral science to measure exposure to extremist content across social media platforms and to determine how these platforms may reinforce (or hinder) those patterns of behavior individually and collectively. As our findings suggest, these problems often center on the way social media platforms enable the distribution of potentially harmful content to vulnerable audiences rather than algorithmic exposure itself.

MATERIALS AND METHODS

Study participants

We contracted with the survey company YouGov to conduct a public opinion survey with 4000 respondents from three distinct populations: a nationally representative sample of 2000 respondents who previously took part in the 2018 Cooperative Congressional Election Survey (CCES) when it was fielded by YouGov; an oversample of 1000 respondents who expressed high levels of racial resentment (*40*), hostile sexism (*41*), and denial of institutional racism (*42*) in their responses to the 2018 CCES; and an oversample of 1000 respondents who did not take part in the 2018 CCES but indicated that they use YouTube "several times per day" or "almost constantly" in their survey response. (The prior measures of racial resentment and hostile sexism, which were collected as part of the 2018 CCES for 3000 of our 4000 respondents, are also used as independent variables in our analysis; see below for details on question wording.) While completing the survey, participants who used an eligible browser (Chrome or Firefox) were offered the opportunity to download a browser extension that would record their browser activity in exchange for additional compensation. A total of 1181 respondents did so (778 from the nationally representative sample, 97 from the high resentment oversample, and 306 from the high YouTube user oversample).

All analyses we report below use survey weights created by YouGov to account for the fact that, in addition to a national sample, we have also specifically recruited participants who fall into one of two oversample groups: (i) those who previously expressed gender and/or racial resentment or (ii) those who are frequent YouTube users. When we apply these weights to all three samples, the total sample is weighted to be nationally representative. Applying these weights to the subset of participants who installed the browser extension helps us to best approximate the characteristics of a nationally representative sample, although the sample is of course not fully representative of the U.S. adult population. We therefore report weighted estimates of the number of users or cases of a behavior and weighted percentages or proportions for maximum clarity. Additional details about respondent demographics and other characteristics are provided in the Supplementary Materials.

Ethics and privacy

Our study methods were approved by the Institutional Review Boards (IRBs) at the authors' respective institutions (Dartmouth CPHS STUDY00032001, Northeastern IRB #20-03-04 and University of Exeter Social Sciences and International Studies Ethics Committee #201920-111). All participants were asked to consent to data collection before completing our survey and again when they installed our browser extension. Participants were fully informed about the data collected by our extension when they were invited to install it and again during installation of the extension. The extension did not collect any data until consent was provided and participants were free to opt out at any time by uninstalling our extension. The extension automatically uninstalled itself from participants' browsers at the end of the study period. (See the Supplementary Materials for the full text of our informed consent notices.)

To protect participants' security and privacy, we adopted a number of best practices. Our participants are indexed by pseudonymous identifiers. Our browser extension used Transport Layer Security (TLS) to encrypt collected data while it was in transit. All participant data are stored on servers that are physically secured by key cards. We use standard remote access tools such as Secure Shell (SSH) to access participant data securely.

We have posted data and code on Dataverse that allows for the replication of all results in this article (linked in the "Data and materials availability" section). All analysis code has also been posted. However, raw behavior data cannot be posted publicly to protect the privacy of respondents.

Data collection and measurement

Our data collection approach focuses on browser activity data, which provide important advantages relative to the history data that are provided by the web browser's WebExtension Application Programming Interface (API). The browser APIs report the time when a given web page was first opened and the time when a user makes a transition from that page to another page (e.g., by clicking a link). To account for duplicate data, we dropped additional page views of the same URL within 1 s of the prior page view on the assumption that the user refreshed the page (*43*). However, the APIs do not report the total dwell time on a given web page taking into account changes in the active browser tab. For example, if someone opens web page A in a tab, then opens web page B in another tab, and then switches their browser tab back to A, the browser history APIs will not register this shift in attention, making it difficult to obtain accurate estimates of time spent on a given web page. Our passive monitoring records all changes in the active tab, allowing us to overcome this issue. (In the Supplementary Materials, we validate our browser activity data against browser history data from the extension.)

In this article, we describe YouTube "views," "consumption," and "exposure" using the browser activity data described above. As with any passive behavioral data, we cannot verify that every user saw the content that appeared on their device in every instance.

We measured the amount of time a user spent on a given web page by calculating the difference between the timestamp of the page in question and the next one they viewed. This measure is imperfect because we do not have a measure of eye gaze or a proxy for active viewing. Although some participants might rewind and rewatch videos more than once, we are more concerned about our measure overstating watch time due to users leaving their browser idling. We therefore refine this measure by capping our measure of time spent at the length of the video in question (obtained from the YouTube API).

We measure which channels users subscribed to by extracting additional information from the HTML snapshots of the videos they watched. Specifically, we parsed the subscribe button from each HTML snapshot, which reads "subscribe" when the participant was not subscribed to the video channel at the time the video was watched and "subscribed" when they were already subscribed. Because we must use this indirect method to infer channel subscriptions, we do not know the full set of channels to which participants subscribe. In particular, not all recommended videos in our dataset were viewed by participants. As a result, we could not determine the subscription status for all recommended videos.

We denote the web page that a participant viewed immediately before viewing a YouTube video as the referrer. We are unable to measure HTTP referrer headers using our browser extension, so, instead, we rely on browser activity data to identify referrers to YouTube videos. Using prior browsing history is a common proxy used to analyze people's behavior on the web (*33*, *44*).

All analyses of the percentage of recommendations seen or followed are based on the full set of recommendations that we could extract from each video. The mean number of recommended videos captured was 17.9, and the median was 20, which aligns with the default number of recommendations shown on a YouTube video (*20*) at the time our study was conducted.

Channel definitions and measurement

Following studies of information consumption online that rely on ratings of content quality at the domain level (*32*, *33*), we construct a typology of YouTube channel types to measure participant exposure. Given that YouTube has tens of millions of channels and that the types of content we are interested in a relatively rare, it is necessary to rely on the judgment of experts to help us identify alternative, extremist, and mainstream media channels. We use the resulting channel lists to classify all videos to which our participants are exposed as coming from an alternative channel, an extremist channel, a mainstream media channel, or some other type of channel ("other"). The process by which these channel lists were defined and compiled is described further below; the Supplementary Materials provide more detail on the procedures used by these experts to label channels.

In our typology, alternative channels discuss controversial topics through a lens that attempts to legitimize discredited views by casting them as marginalized viewpoints (despite the channel owners often identifying as white and/or male). Our list combines the 223 channels classified by Ledwich and Zaitsev (*26*) as Men's Rights Activists or Anti-Social Justice Warriors, the 141 Intellectual Dark Web and Alt-lite channels from Ribeiro *et al.* (*24*), and the 24 channels from Lewis' Alternative Influence Network (*35*). After removing duplicates, our alternative channel list contains 322 channels, of which 68 appeared on two source lists, and nine appeared on three. Example alternative channels in our typology include those hosted by Steven Crowder, Tim Pool, Laura Loomer, and Candace Owens. Joe Rogan's is the most prominent alternative channel in our typology (it appears on all three source lists), accounting for 11.6% (95% CI, 11.3 to 12.0) of all visits and 26.0% (95% CI, 26.0 to 26.1) of all time spent on alternative channel videos.

Our list of extremist channels consists of those labeled as white identitarian by Ledwich and Zaitsev (*26*) (30 channels), white supremacist by Charles (*45*) (23 channels), alt-right by Ribeiro *et al.* (*24*) (37 channels), extremist or hateful by the Center on Extremism at the Anti-Defamation League (16 channels), and those compiled by journalist Aaron Sankin from lists curated by the Southern Poverty Law Center, the Canadian Anti-Hate Network, the Counter Extremism Project, and the white supremacist website Stormfront (157 channels) (*46*). After removing duplicates, our extremist channel list contains 290 channels, of which 36.2% appeared on two or more source lists. Example extremist channels include those hosted by Stefan Molyneux, David Duke, Mike Cernovich, and Faith J. Goldy.

As the examples above suggest, the potentially harmful alternative and extremist channels identified by scholarly and subject matter experts are predominantly from the (far) right in the United States. Other forms of extremism exist, of course, especially outside the United States (e.g., Islamic extremism).

Following prior research, we define both alternative and extremist channels as potentially harmful (*2*, *26*, *35*, *45*). Of the 302 alternative and 213 extremist channels that were still available on YouTube as of January 2021 (i.e., they had not been taken down by the owner or by YouTube), videos from 208 alternative and 55 extremist channels were viewed by at least one participant in our sample. We are not making these lists publicly available to avoid directing attention to them but are willing to privately share them with researchers and journalists upon request.

To create our list of mainstream media channels, we collected news channels from Buntain *et al.* (*47*) (65 mainstream news sources), Ledwich and Zaitsev (*26*) (75 mainstream media channels), Stocking *et al.* (*48*) (81 news channels), Ribeiro *et al.* (*24*) (68 popular media channels), Eady *et al.* (*49*) (219 national news domains), and Zannettou *et al.* (*50*) (45 news domains). We manually found the corresponding YouTube channels via YouTube search when authors only provided websites (*24*, *36*, *49*). In cases where news organizations have multiple YouTube channels (e.g., Fox News and Fox Business), all YouTube channels under the parent organization were included. Any channels appearing in fewer than three of these sources were omitted. Last, we also included channels that were featured on YouTube's www.youtube.com/channel/UCYfdidRxbB8Qhf0Nx7ioOYw News hub from 10 February to 5 March 2021.

The resulting list of mainstream media channels was then checked to identify those that meet all of the following criteria:

1) They must publish credible information, which we define as having a NewsGuard score greater than 60 (www.newsguardtech.com) and not being associated with any "black" or "red" fake news websites listed in Grinberg *et al.* (*32*).

2) They must meet at least one criteria for mainstream media recognition or distribution, which we define as having national print circulation, having a cable TV network, being part of the White House press pool, or having won or been nominated for a prestigious journalism award (e.g., Pulitzer Prize, Peabody Award, Emmy, George Polk Award, or Online Journalism Award).

3) They must be a United States–based organization with national news coverage.

Our final mainstream media list consists of 127 YouTube channels. We then placed all YouTube channels in our dataset that did not fall into one of these three categories (alternative, extremist, or mainstream media) into a residual category that we call "other." (These may include alternative, extremist, or mainstream media that were missed by the processes described above.)

Survey measures of racial resentment and hostile sexism

We measure anti-Black animus with a standard four-item scale intended to measure racial resentment (*40*). For example, respondents were asked whether they agree or disagree with the statement "It's really a matter of some people just not trying hard enough: If blacks would only try harder, they could be just as well off as whites." Responses are provided on a five-point agree/disagree scale and coded such that higher numbers represent more resentful attitudes. Respondents' racial resentment score is the average of these four questions. Responses to these questions are taken from respondent answers to the 2018 CCES (as noted above, participants were largely recruited from the pool of previous CCES respondents).

We operationalized hostile sexism using two items from a larger scale that was also asked on the 2018 CCES (*41*). For example, one of the questions asks whether respondents agree or disagree with the statement "When women lose to men in a fair competition, they typically complain about being discriminated against." Responses are provided on a five-point agree/disagree scale and coded such that higher numbers represent more hostile attitudes.

All other question wording is provided in the survey codebook in the Supplementary Materials. Racial resentment and hostile sexism measures were also included in our 2020 survey; responses showed a high degree of persistence over time [$r = 0.92$ (95% CI, 0.91 to 0.92)] for racial resentment, $r = 0.79$ (95% CI, 0.78 to 0.81) for hostile sexism. The two measures, which we refer to as measuring "resentment" or identifying "resentful" users per, e.g., Banda and Casses (*51*) and Schaffner (*52*), were highly correlated with each other as well ($r = 0.84$).

Supplementary Materials

This PDF file includes:

Figs. S1 to S18
Tables S1 to S10
Sample details and additional results
Session trajectories
Channel labeling criteria
Ethics and consent language
Survey codebook

REFERENCES AND NOTES

1. C. Sunstein, *Republic.com*(Princeton Univ. Press, 2001).
2. N. J. Stroud, Polarization and partisan selective exposure. *J. Commun.* **60**, 556–576 (2010).
3. M. Gentzkow, J. M. Shapiro, Ideological segregation online and offline. *Q. J. Econ.* **126**, 1799–1839 (2011).
4. B. Auxier, M. Anderson, *Social Media Use in 2021* (Pew Research Center, 2021); www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/.
5. J. E. Solsman, *YouTube's AI is the Puppet Master over Most of What You Watch* (CNET, 2018); www.cnet.com/news/youtube-ces-2018-neal-mohan/.
6. Z. Tufekci, *YouTube, The Great Radicalizer* (New York Times, 2018); www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html.
7. J. Nicas, *How YouTube Drives People to the Internet's Darkest Corners* (Wall Street Journal, 2018); www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478.
8. K. Roose, *The Making of A YouTube Radical* (New York Times, 2019); www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html.
9. J. McCrosky, B. Geurkink, *YouTube Regrets: A Crowdsourced Investigation into YouTube's Recommendation Algorithm* (Mozilla Foundation, 2021); https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf.
10. YouTube, *Continuing Our Work to Improve Recommendations on YouTube* (YouTube, 2019); https://blog.youtube/news-and-events/continuing-our-work-to-improve/.
11. YouTube, *Our Ongoing Work to Tackle Hate* (YouTube, 2019); https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate/.
12. YouTube, *The Four Rs of Responsibility, Part 2: Raising Authoritative Content and Reducing Borderline Content and Harmful Misinformation* (YouTube, 2019); https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce.
13. C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. B. Fallin Hunzaker, J. Lee, M. Mann, F. Merhout, A. Volfovsky, Exposure to opposing views on social media can increase political polarization. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 9216–9221 (2018).
14. S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
15. E. Bakshy, S. Messing, L. A. Adamic, Exposure to ideologically diverse news and opinion on Facebook. *Science* **348**, 1130–1132 (2015).
16. D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, J. L. Zittrain, The science of fake news. *Science* **359**, 1094–1096 (2018).
17. A. Guess, J. Nagler, J. Tucker, Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Sci. Adv.* **5**, eaau4586 (2019).
18. K. Munger, J. Phillips, Right-wing YouTube: A supply and demand perspective. *Int. J. Press Polit.* **27**, 186–219 (2022).
19. M. Yesilada, S. Lewandowsky, Systematic review: YouTube recommendations and problematic content. *Internet Policy Rev.* **11**, 1652 (2022).
20. H. Hosseinmardi, A. Ghasemian, A. Clauset, M. Mobius, D. M. Rothschild, D. J. Watts, Examining the consumption of radical content on YouTube. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2101967118 (2021).
21. M. Faddoul, G. Chaslot, H. Farid, A longitudinal analysis of YouTube's promotion of conspiracy videos. arXiv:2003.03318 [cs.CY] (2020)
22. E. Hussein, P. Juneja, T. Mitra, Measuring misinformation in video search platforms: An audit study on YouTube. *Proc. ACM Hum. Comput. Interact.* **4**, 1–27 (2020).
23. K. Papadamou, A. Papasavva, S. Zannettou, J. Blackburn, N. Kourtellis, I. Leontiadis, G. Stringhini, M. Sirivianos, Disturbed YouTube for kids: Characterizing and detecting inappropriate videos targeting young children. *Proc. Int. AAAI Conf. Web Soc. Media* **14**, 522–533 (2020).
24. M. H. Ribeiro, R. Ottoni, R. West, V. A. Almeida, W. Meira Jr., Auditing radicalization pathways on youtube, in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, Barcelona, Spain, 27 to 30 January 2020, pp. 131–141.
25. J. Bisbee, M. Brown, A. Lai, R. Bonneau, J. Nagler, J. A. Tucker, Election fraud, YouTube, and public perception of the legitimacy of president Biden. *J. Online Trust Saf.* **1**, 1 (2022).
26. M. Ledwich, A. Zaitsev, Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *First Monday* **25**, 10.5210/fm.v25i3.10419, (2020).
27. R. E. Robertson, Uncommon yet consequential online harms. *J. Online Trust Saf.* **1**10.54501/jots.v1i3.87, (2022).
28. C. Ballard, I. Goldstein, P. Mehta, G. Smothers, K. Take, V. Zhong, R. Greenstadt, T. Lauinger, D. M. Coy, Conspiracy brokers: Understanding the monetization of youtube conspiracy theories, in *Proceedings of the ACM Web Conference 2022*, Association for Computing Machinery Inc., Barcelona, Spain, 28 June to 1 July 2022, pp. 2707–2718.
29. The International Fact-Checking Network, *An Open Letter to YouTube's CEO from the World's Fact-Checkers* (2022); www.poynter.org/fact-checking/2022/an-open-letter-to-youtubes-ceo-from-the-worlds-fact-checkers/.
30. M. S. Locatelli, J. Caetano, W. Meira Jr., V. Almeida, Characterizing vaccination movements on YouTube in the United States and Brazilm, in *Proceedings of the ACM Conference on Hypertext and Social Media*, Association for Computing Machinery, Barcelona, Spain, 28 June to 1 July 2022, pp. 80–90.
31. T. Yang, S. González-Bailón, *Online Media Boosts Exposure to News but Only for A Small Minority Of Hyper-Consumers* (SSRN, 2021); https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3954565.
32. N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, Fake news on Twitter during the 2016 U.S. presidential election. *Science* **363**, 374–378 (2019).
33. A. M. Guess, B. Nyhan, J. Reifler, Exposure to untrustworthy websites in the 2016 US election. *Nat. Hum. Behav.* **4**, 472–480 (2020).
34. J. Allen, B. Howland, M. Mobius, D. Rothschild, D. J. Watts, Evaluating the fake news problem at the scale of the information ecosystem. *Sci. Adv.* **6**, eaay3539 (2020).
35. B. Lewis, *Alternative Influence: Broadcasting the Reactionary Right on YouTube* (Data & Society, 2018); https://datasociety.net/library/alternative-influence/.
36. S. Zannettou, J. Finkelstein, B. Bradlyn, J. Blackburn, in *Proceedings of the International AAAI Conference on Web and Social Media*, Association for the Advancement of Artificial Intelligence, Atlanta, Georgia, 8 to 11 June 2019, vol. 14, pp. 786–797.

37. R. Mamié, M. Horta Ribeiro, R. West, Are anti-feminist communities gateways to the far right? Evidence from Reddit and YouTube, in *Proceedings of the 13th ACM Web Science Conference 2021*, Association for Computing Machinery, Virtual event, 21 to 25 June 2021, pp. 139–147.
38. C. Goodrow, *On YouTube's Recommendation System* (YouTube, 2021); https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/.
39. R. E. Robertson, J. Green, D. J. Ruck, K. Ognyanova, C. Wilson, D. Lazer, Users choose to engage with more partisan news than they are exposed to on google search. *Nature* **618**, 342–348 (2023).
40. D. R. Kinder, L. M. Sanders, *Divided by Color: Racial Politics and Democratic Ideals* (University of Chicago Press, 1996).
41. P. Glick, S. T. Fiske, Hostile and benevolent sexism: Measuring ambivalent sexist attitudes toward women. *Psychol. Women Q.* **21**, 119–135 (1997).
42. C. D. DeSante, C. Watts Smith, Less is more: A cross-generational analysis of the nature and role of racial attitudes in the twenty-first century. *J. Theor. Polit.* **82**, 967–980 (2020).
43. A. M. Guess, (Almost) everything in moderation: New evidence on Americans' online media diets. *Am. J. Polit. Sci.* **65**, 1007–1022 (2021).
44. M. Wojcieszak, E. Menchen-Trevino, J. F. Goncalves, B. Weeks, Avenues to news and diverse news exposure online: Comparing direct navigation, social media, news aggregators, search queries, and article hyperlinks. *Int. J. Press Polit.* **27**, 860–886 (2021).
45. C. Charles, *(Main)streaming Hate: Analyzing White Supremacist Content and Framing Devices On YouTube* (University of Central Florida, 2020); https://stars.library.ucf.edu/etd2020/27/.
46. A. Sankin, *YouTube Said It was Getting Serious about Hate Speech. Why is It Still Full of Extremists?* (Gizmodo, 2019); https://gizmodo.com/youtube-said-it-was-getting-serious-about-hate-speech-1836596239.
47. C. Buntain, R. Bonneau, J. Nagler, J. A. Tucker, YouTube recommendations and effects on sharing across online social platforms. *Proc. ACM Hum. Comput. Interact.* **5**, 1–26 (2021).
48. G. Stocking, P. Van Kessel, M. Barthel, K. Eva Matsa, M. Khuzam, *Many Americans Get News on YouTube, Where News Organizations and Independent Producers Thrive Side by Side* (Pew Research Center, 2020); www.pewresearch.org/journalism/2020/09/28/many-americans-get-news-on-youtube-where-news-organizations-and-independent-producers-thrive-side-by-side/.
49. G. Eady, R. Bonneau, J. A. Tucker, J. Nagler, *News Sharing on Social Media: Mapping the Ideology of News Media Content, Citizens, and Politicians* (OSF, 2020); https://osf.io/preprints/ch8gj/.
50. S. Zannettou, T. Caulfield, E. De Cristofaro, N. Kourtelris, I. Leontiadis, M. Sirivianos, G. Stringhini, J. Blackburn, The web centipede: Understanding how web communities influence each other through the lens of mainstream and alternative news sources, in *Proceedings of the 2017 Internet Measurement Conference* (ACM, 2017), pp. 405–417.
51. K. K. Banda, E. C. Cassese, Hostile sexism, racial resentment, and political mobilization. *Polit. Behav.* **44**, 1317–1335 (2022).
52. B. F. Schaffner, Optimizing the measurement of sexism in political surveys. *Polit. Anal.* **30**, 1–17 (2021).

Acknowledgments: We are grateful to the Russell Sage Foundation, Anti-Defamation League, Carnegie Corporation of New York, the National Science Foundation, John Smith Guggenheim Memorial Foundation, and the European Research Council (ERC) for financial support; to S. Luks at YouGov for survey assistance; to K. Rhee for research assistance; to A. Guess for helping design this project in its initial stages; and to T. Mitra, J. B. Phillips, D. Rothschild, G. Stringhini, and S. Zannettou for comments and feedback. We also thank V. A. F. Almeida, S. Ansolabehere, M. H. Ribeiro, A. Sankin, B. Schaffner, R. West, and A. Zaitsev for sharing their data with us or making it publicly available. This research used equipment funded by NSF grant IIS-1910064. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders. In general, all conclusions and errors are our own. **Funding:** This study was supported by the Russell Sage Foundation, Anti-Defamation League, Carnegie Corporation of New York, Guggenheim Foundation, and the National Science Foundation (grant IIS-1910064). This project (to J.R.) received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 682758). **Author contributions:** B.N., J.R., and C.W. designed the study. All the authors wrote the original manuscript. B.N., J.R., and C.W. revised the manuscript. R.E.R. collected the browser data. A.Y.C. and R.E.R. analyzed the data. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Data and code necessary to replicate the results in this study have been posted on Dataverse (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/UC1XM1) and GitHub (https://github.com/aychen5/youtube-extremism-replication). Please note that the Dataverse materials will be the permanent record of data and code. Consistent with guidance from *Science Advances*, please be aware that GitHub libraries may be modified in ways that complicate reproducibility. To protect the privacy of respondents, no raw digital trace data are included. The data provided are aggregated at the respondent level and only provide key demographic variables and channel counts, which guards against reidentification. All other data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 5 July 2022
Accepted 25 July 2023
Published 30 August 2023
10.1126/sciadv.add8080