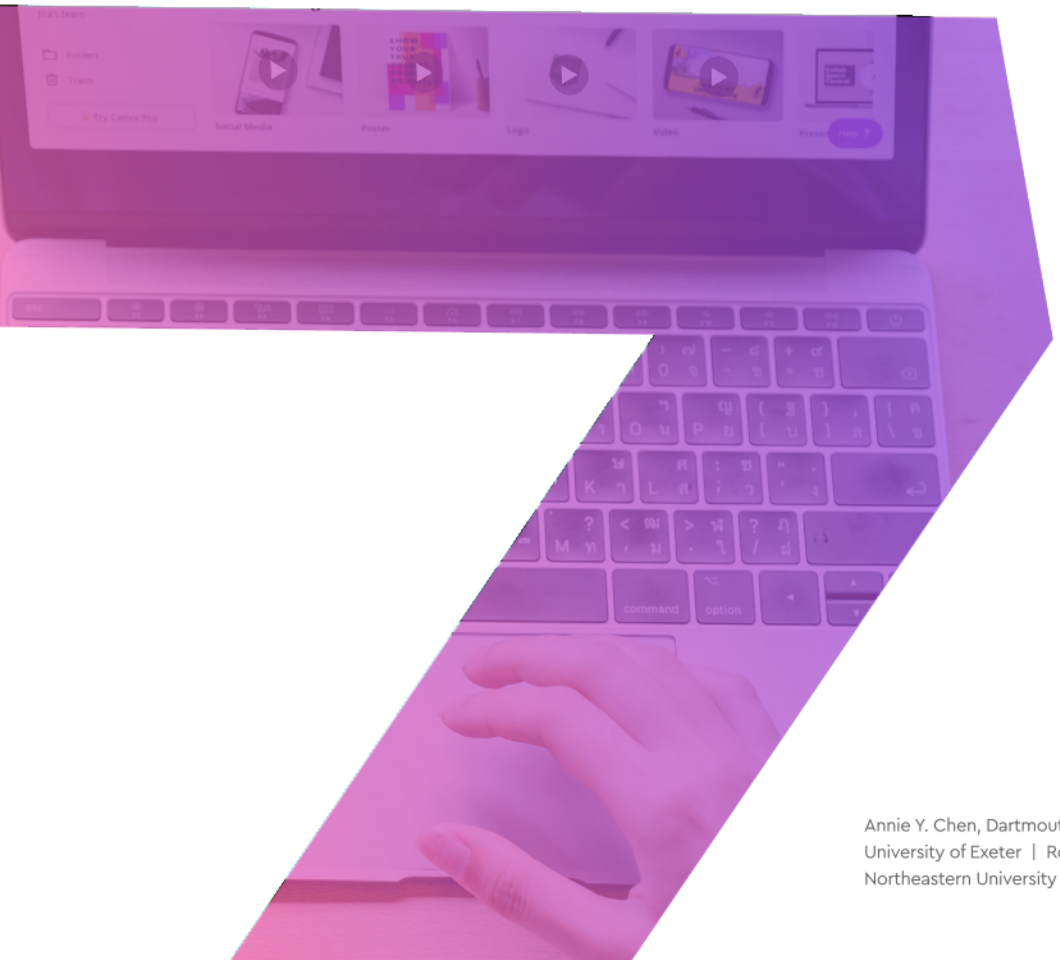
The Belfer Fellowship Series

ADL | CENTER FOR TECHNOLOGY & SOCIETY

EXPOSURE TO

ALTERNATIVE & EXTREMIST CONTENT ON YOUTUBE



Annie Y. Chen, Dartmouth College | Brendan Nyhan, Dartmouth College | Jason Reifler, University of Exeter | Ronald E. Robertson, Northeastern University | Christo Wilson, Northeastern University

OUR

MISSION

To stop the defamation of the Jewish people and to secure justice and fair treatment to all.

The Belfer Fellowship

The Belfer Fellowship was established by the Robert Belfer Family to support innovative research and thought-leadership on combating online hate and harassment for all. Fellows are drawn from the technologist community, academia, and public policy to push innovation, research and knowledge development around the online hate ecosystem. ADL and the Center for Technology and Society thank the Robert Belfer Family for their dedication to our work, and their leadership in establishing the Fellows program.

ABOUT

CENTER FOR TECHNOLOGY & SOCIETY

In a world riddled with cyberhate, online harassment and misuses of technology, the Center for Technology & Society (CTS) serves as a resource to tech platforms and develops proactive solutions. Launched in 2017 and headquartered in Silicon Valley, CTS aims for global impacts and applications in an increasingly borderless space.

It is a force for innovation, producing cutting-edge research to enable online civility, protect vulnerable populations, support digital citizenship and engage youth. CTS builds on ADL's experience over more than a century building a world without hate and supplies the tools to make that a possibility both online and offline.

ADL®
FIGHTING HATE FOR GOOD

ADL (Anti-Defamation League) fights antisemitism and promotes justice for all. Join ADL to give a voice to those without one and to protect our civil rights.

AUTHORS

Annie Y. Chen, Dartmouth College
Brendan Nyhan, Dartmouth College
Jason Reifler, University of Exeter
Ronald E. Robertson, Northeastern University
Christo Wilson, Northeastern University

TABLE OF
CONTENTS

EXECUTIVE

SUMMARY

How harmful is YouTube? Critics worry that it plays an outsized role among technology platforms in exposing people to hateful or extreme ideas, while the platform claims to have substantially reduced the reach of what it calls "borderline content and harmful misinformation."[1] However, little is publicly known about who watches potentially harmful videos on YouTube, how much they watch, or the role of the site's recommendations in promoting those videos to users.

To answer these questions, we collected comprehensive behavioral data measuring YouTube video and recommendation exposure among a diverse group of survey participants. Using browser history and activity data, we examined exposure to extremist and white supremacist YouTube channels as well as to "alternative" channels that can serve as gateways to more extreme forms of content.

Our data indicate that exposure to videos from extremist or white supremacist channels on YouTube remains disturbingly common. Though some high-profile channels were taken down by YouTube before our study period, approximately one in ten participants viewed at least one video from an extremist channel (9.2%) and approximately two in ten (22.1%) viewed at least one video from an alternative channel.[2] Moreover, when participants watch these videos, they are more likely to see and follow recommendations to similar videos.

9.2%

Approximately one in ten participants viewed at least one video from an extremist channel.

22.1%

Approximately two in ten viewed at least one video from an alternative channel.

Consumption was concentrated among a highly engaged subset of respondents. Among those who watched at least one video of a given type, the mean numbers of videos watched were 64.2 (alternative) and 11.5 (extremist). Moreover, consumption of these videos was most frequent among people with negative racial views.

Approximately one in five people who previously reported high levels of racial resentment watched at least one video from an alternative channel (19.4%) and nearly one in six watched a video from an extremist channel (14.7%).[3] In total, people with high racial resentment were responsible for more than 90% of views for videos from alternative and extremist channels.

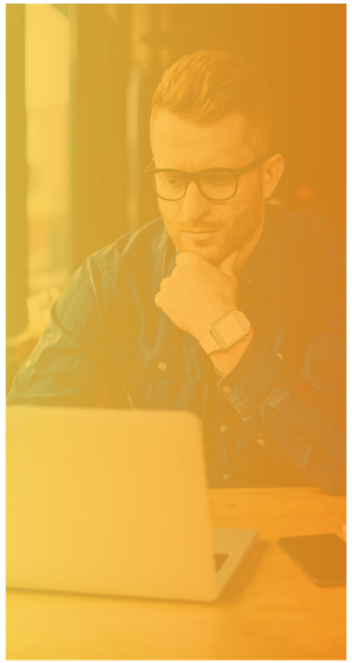
In addition, participants often received and sometimes followed recommendations for videos from alternative and extremist channels, especially on videos from those channels. Though YouTube says it made "over 30 different changes to reduce recommendations" of potentially harmful content,[4] 37.6% of recommendations on videos from alternative channels and 29.3% of recommendations on videos from extremist channels were to other videos of the same type. As a result, 6.6% of participants followed at least one recommendation to a video from an alternative channel and 2.1% followed at least one to a video from an extremist channel. (We cannot say why YouTube's algorithm opts to surface these videos; according to YouTube, some may be selected because users subscribe to the channels in question.)

Overall, we do not find clear evidence that people with neutral or mixed views on issues such as race frequently view videos from alternative or extremist channels on YouTube.

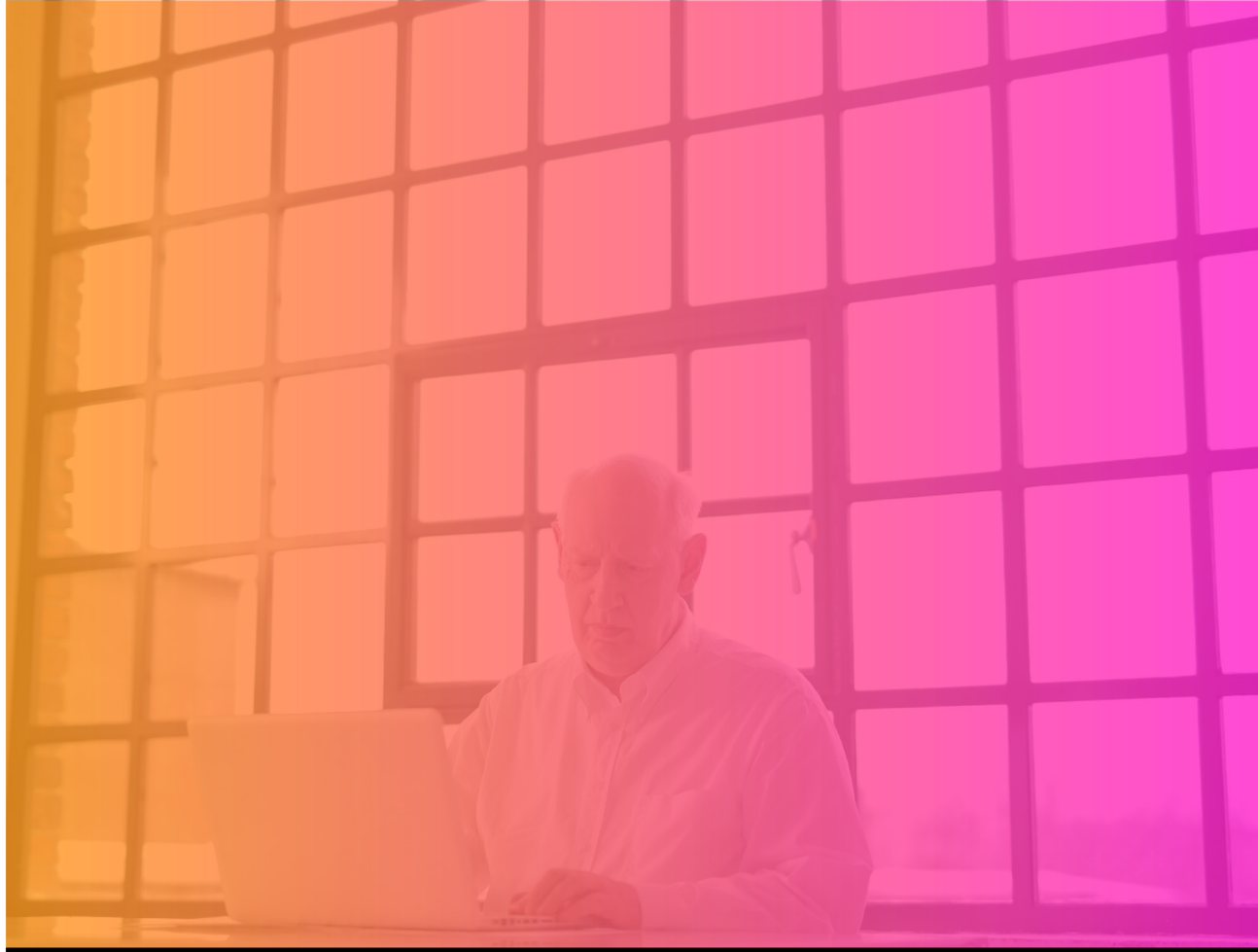
Consumption of this potentially harmful content is instead concentrated among Americans who are already high in racial resentment, the group that is seemingly most vulnerable to its influence. Moreover, despite recent changes to the YouTube algorithm, the site still frequently recommends videos from alternative or extremist channels when people watch a video from those channels. As a result, many racially resentful people are not only watching large numbers of videos from alternative or extremist channels, but also are shown recommendations for more such videos when they do so, further increasing exposure to potentially harmful content.

CONCERNS ABOUT YOUTUBE AS AN ENGINE OF EXTREMISM

YouTube is an important potential vector for harmful and extremist content online. According to the Pew Research Center, YouTube is the most commonly used social media platform in the U.S.[5] However, the site often gets less attention than Facebook and Twitter, which are frequently criticized for spreading and popularizing hateful and virulent content.



YouTube's design and architecture suggest numerous reasons for concern. First, YouTube is an open platform that depends on user-generated content and thus allows people with fringe or extremist views to compete directly with established media and information sources. Second, the financial incentives that YouTube provides based on viewership and watch time may encourage creators to appeal to people with extreme views and provoke controversy. Third, YouTube's algorithm makes recommendations based in part on past user behavior.[6] These recommendations can influence user behavior, especially because the top recommendation is played after the current video concludes by default. Fourth, video requires transcription and takes longer for humans to review than text, posing difficult content moderation challenges.



Finally, the use of video may cause viewers to form parasocial relationships with people whose videos they frequently watch,[7,8] potentially increasing the effects of exposure to those videos.

To understand the type of problematic and extremist content that YouTube's architecture may help spread, Lewis[9] provides a detailed descriptive analysis of 65 channels dubbed the Alternative Influence Network (AIN). Members of the AIN range from self-declared white nationalists to those who identify primarily as conservative or libertarian thinkers. These channels are unified by "a general opposition to feminism, social justice, or left-wing politics" and a rejection of the traditional news media. These channels are densely interconnected and may introduce viewers to extremist ideas (e.g., white supremacy, conspiracy theories, etc.) from guests and other channels in the network.

A prominent concern in recent years is how YouTube's algorithm, which relies heavily on user engagement, helps extend the reach of these "alternative" communities. Like other social media platforms, YouTube makes money through advertisements on its platform and therefore seeks to keep users on the site as long as possible to generate more ad revenue. To this end, its recommendation algorithms seek to maximize the time users spend watching videos. In 2018, YouTube estimated that approximately 70% of watch time on the site was driven by its recommender system.[10]

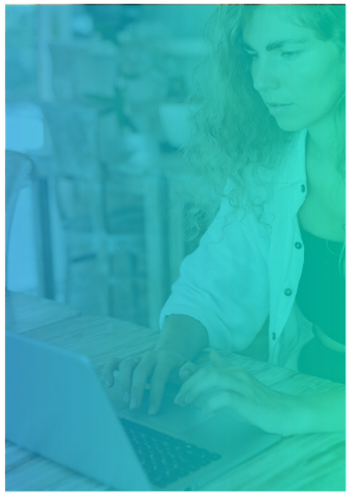
When watching a video, viewers see a set of recommended videos in an adjacent sidebar. When watching a video from "alternative" channels, the recommendations shown there for what to watch next may be especially likely to be extremist or harmful videos since they tend to feature other videos uploaded to the same channel, videos with similar content, or videos that have a similar audience. Faddoul, Chaslot and Farid find, for instance, that there is a clear positive correlation between watching conspiracy videos and being recommended conspiracy videos on YouTube.[11] Critics fear that people who watch one such video may receive recommendations for other potentially harmful content and descend into a "rabbit hole" of increasingly extreme videos that could be radicalizing.[12, 13]

One journalistic account finds that YouTube consistently recommended far-right or far-left videos even to those who sought mainstream news sources.[14] This system has led critics to allege that YouTube's recommender system not only amplifies conspiratorial and extremist content, but generates pathways to radicalization.[15]

There is, however, only very limited evidence that YouTube causes radicalization via the recommendation algorithm. In an expanded analysis of the AIN typology, Ribeiro et al. identify three ideological communities on YouTube: the "Intellectual Dark Web" (IDW) who "discuss controversial subjects like race and I.Q. without necessarily endorsing extreme views," the "alt-right" who "sponsor fringe ideas like that of a white ethnostate," and the "alt-lite" who "deny to embrace white supremacist ideology, although they frequently flirt with concepts associated with it."[16]

Ribeiro et al. find considerable overlap in their user bases; about half of the users who commented on alt-right videos also commented on alt-lite and IDW channels. Their analyses also suggest that users who initially only commented on videos of milder forms of contrarian media (alt-lite and IDW) went on to consume alt-right content.

However, this analysis cannot establish that YouTube's recommender system is responsible for the commenter trajectories they observe. Consistent with this point, Ribeiro et al. show that a random walk through recommended videos that starts on an alt-lite video only reaches an alt-right video by its fifth step 1 in 2,000 times.



Similarly, Ledwich and Zaitsev analyze non-personalized recommendations among different types of videos from a set of more than 800 politically focused channels with large numbers of subscribers.[17] These recommendations, which were scraped anonymously and thus do not take into account prior viewing behavior, tend to direct people out of categories with more extreme political views such as "conspiracy," "anti-SJW," and "white identitarian" and into more conventional categories such as "partisan right" (e.g., Fox News) and non-political topics. However, it is unknown how their results would change if user behavior were taken into account.

One complicating factor is that both the recommendation algorithm and consumption patterns on the site may change over time. Munger and Phillips find that consumption of far-right videos has declined since mid-2017, which they attribute to an increased supply of conservative videos from mainstream sources.[18] After changes in the YouTube algorithm in January 2019, the site claimed U.S. watch time of "borderline content and harmful misinformation" declined by 70% on average for content from non-subscribed recommendations.[19] Buntain et al. further identify a negative trend of sharing AIN videos on Twitter and Reddit after the change (though sharing conspiracy theories did not change on Twitter and increased on Reddit).[20] Faddoul, Chaslot and Farid also find evidence of a decline in conspiracy recommendations afterward.[21]

In addition, few of these studies account for the way YouTube's algorithms personalize recommendations for users based on their past behavior, which is often emphasized in radicalization claims and could compound the effects of exposure to potentially harmful content. For instance, Hussein, Juneja and Mitra find that watching dubious conspiracy theory videos increases recommendations of more such videos on most topics.[22] Similarly, Papadamou et al. find that personalization on YouTube increases recommendations of pseudoscientific videos.[23]

Most recently, Hosseinmardi et al. present an analysis of YouTube use in desktop browsing data collected from a nationally representative Nielsen panel of more than 300,000 Americans.[24] Though they find that news consumption on YouTube is rare overall, consumption of far-right content increased significantly over the 2016–2019 period in their data.

Moreover, Hosseinmardi et al. find that median watch times per video and "stickiness" in consumption preferences on YouTube both exceed the levels observed among other types of news audiences. Their analysis of web browsing data shows that consumers of far-right YouTube content also consume other far-right web content and frequently visit far-right YouTube videos via links from external sites rather than the algorithm.

METHODOLOGY

In this report, we link representative survey and online behavior data to provide the first ecologically valid, individual-level measures of exposure to potentially harmful and extremist YouTube content.

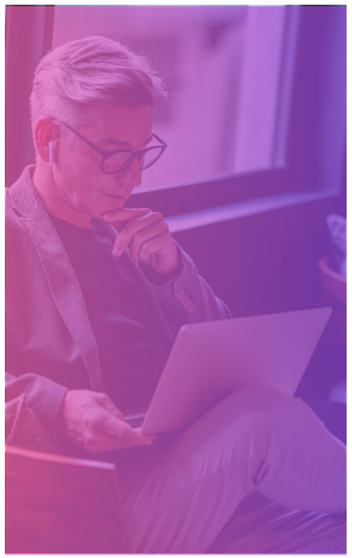
Nationally representative survey

We conducted a nationally representative survey measuring demographic characteristics and political attitudes such as age, race, education, partisanship, ideology, and political knowledge and interest. We also asked questions about people's awareness of algorithm use on technology platforms, their satisfaction with search results on YouTube, and their perceptions of the viewpoints they are exposed to on YouTube.

Survey data were collected from members of the YouGov America online panel. YouGov, a national polling firm, then uses matching and weighting procedures to approximate a nationally representative sample. We restricted eligibility to users of Google Chrome and Mozilla Firefox (which account for approximately two-thirds of installed U.S. desktop browsers according to StatCounter) to increase the likelihood of finding eligible participants for our YouTube study, which employs extensions for those browsers. We drew respondents primarily from the set of YouGov panel members who took part in a 2018 omnibus study called Cooperative Congressional Election Survey, which provided data on respondents' prior political attitudes. We also recruited an oversample of users who reported using YouTube several times per day. Data were collected from July 21, 2020 to September 22, 2020.

YouTube viewing data

With participants' consent, we collected data on their visits to YouTube, the



videos they watch and engage with, and the algorithmic recommendations shown to them on the site. These data points, which are collected via an extension that users installed in Google Chrome or Mozilla Firefox browsers on their desktop or laptop computers, allowed us to measure YouTube use in the population and determine how it varied across different demographic and political subgroups.[25] By joining data on the recommendations users saw with the actual viewing choices they made, we can provide a unique portrait of pathways to potentially harmful content on YouTube.

The data cover browser history records starting prior to the date of installation of the browser extension, as well as data measured directly by the browser extension after installation (usually the day the respondent completed the survey). We refer to these as browser history data and browser activity data, respectively. Browser history data is collected through an API built into Chrome and Firefox which documents every URL visit with a timestamp and other metadata. Browser activity data is a direct measure of how users switched among their open browser tabs and what users saw via HTML snapshots (for a subset of domains only).

We combined these data types by leveraging their respective advantages. Browser history offers an expanded pre-installation user history. Browser activity data offers HTML snapshots (which capture YouTube's recommendations and other website features) and tab shifts (which provide greater validity for estimates of attention such as video watch time). We collected browser activity data from July 21, 2020 (date of first extension install) to October 21, 2020 and obtained browser history data from April 22, 2020 (three months back for the first installer) to October 21, 2020. As we discuss below, we measured video watch counts based on browser history (covering a longer period of user behavior) but used activity data from the browser extension to analyze the recommendations paired with those videos and responses to those recommendations. We note that we cannot determine why recommendations were made to users, as they are determined within the proprietary YouTube algorithm. In particular, YouTube has chosen to recommend videos from channels to which users subscribe. Our data do not include the channel subscriptions of study participants.

Identifying problematic or harmful content on YouTube

To examine exposure to extremist content, we first need to identify what content on YouTube is problematic or harmful. In this report, we focus specifically on exposure to two types of YouTube channels:

1. Alternative YouTube channels that potentially serve as gateways to more extreme forms of content. This list of 322 channels is drawn from the Data & Society report on the so-called "Alternate Influence Network" (AIN) on YouTube,[26] the "Intellectual Dark Web" and "alt-lite" channels identified by Ribeiro et al.,[27] and the YouTube channels tagged as "anti-SJW" or focused on men's rights in Ledwich and Zaitsev.[28] This set of channels includes those of conservative commentators Laura Loomer, Steven Crowder, Candace Owens, and Michelle Malkin.[29]

2. Extremist or white supremacist channels. This list of 290 channels comprises those identified as white supremacist by Charles,[30] "alt-right" by Ribeiro et al.,[31] "white identitarian" by Ledwich and Zaitsev,[32] or as extremist/hate content by ADL's Center on Extremism (COE) or sources consulted by Sankin,[33] which include the Counter Extremism Project, Southern Poverty Law Center, and Hope Not Hate as well as a list of channels that were featured on the white supremacist website Stormfront. Channels in this set include those of the far-right activists Mike Cernovich and Faith J. Goldy.

Of the 612 channels in these two lists, 515 were still active as of January 21, 2021. Among the 97 channels that are no longer active, 6 (3 extremist, 3 alternative) were encountered by our participants during this study, while the remaining 91 (17 alternative, 74 extremist) were either taken down before the start of our study (e.g., Stefan Molyneux and David Duke), terminated between October 21, 2020 (the end of our study) and January 21, 2021, or remained active for at least part of the study period but could not be matched to a video viewed by a study participant.[34] Our analysis includes the 259 active channels (206 alternative, 53 extremist) encountered by one or more study participants as well as six channels that were viewed by at least one participant but were taken down at some point during or immediately after the study.

(We omit complete lists of the channels included in each category from this report due to concerns about amplification but will provide them to researchers or journalists upon request.)

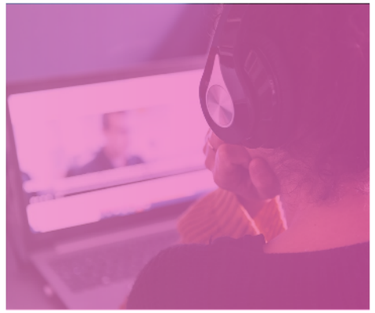
206
Alternative YouTube channels

53
Extremist or white supremacist YouTube channels

Limitations

Our channel-level analysis is some of the most comprehensive to date but still faces limitations that should be noted in considering our findings. First, due to the difficulty of evaluating the content of every video on YouTube, we measured exposure to videos from potentially problematic or harmful channels rather than problematic videos themselves. As such, our data measuring total exposure to potentially problematic or harmful channels will include views of some videos from those channels that are themselves innocuous.

Conversely, we did not measure exposure to extremist content found in individual videos from channels our expert sources do not currently classify as extremist. These limitations are similar to those faced in analyses of exposure to other types of problematic online content such as untrustworthy websites.[35] To address these limitations, we plan to directly examine the content of the videos that people see in future research.



Ethics

Given the sensitivity of the data we collected from our participants, we took several steps to comply with strict ethical norms. This study was approved by the Institutional Review Boards at the authors' respective universities.[36] All participants consented to install our browser extension; no data were collected before we were granted consent. Participants were free to unenroll from our study at any time. TLS encryption was used to secure data transfers from the browser extension to our data collection server. Access to the server was, in turn, restricted to study personnel using standard cryptographic techniques and physical access controls.

All participants in our study are pseudonymous. To the greatest extent possible, we stripped unique identifiers such as Google/YouTube usernames and account identifiers from web data on participants' computers within the browser extension before transmitting the data to our server.

Given the sensitivity of the data we collected from our participants, we took several steps to comply with strict ethical norms. All participants consented to install our browser extension; no data were collected before we were granted consent.

RESULTS

Survey results

In total, we surveyed 4,000 respondents. Of these, 2,000 respondents were selected to be representative of the adult population of the United States by the survey company YouGov, which uses a combination of weighting and matching to approximate the demographics of the U.S. public. We also oversampled two separate groups appropriate to the study—people who self-reported using YouTube several times per day or more (n=1,000) and people who expressed high levels of racial resentment and frequently questioned the prevalence and seriousness of racism and sexism (n=1,000).[37] Not surprisingly, these groups differed on several demographic characteristics.[38]

As we show in Figure 1, the high YouTube (YT) sample is considerably younger than the general population (GP) sample (mean ages of 41 and 50, respectively), while the oversample of people who expressed negative views on race and gender (NVRG) is slightly older (mean age of 54).

Both the YT and NVRG oversamples are disproportionately male (YT: 60.8% male; NVRG: 63.8% male) compared to the GP sample (47.0% male). The YT sample is more racially diverse than the GP sample (54.5% white versus 70.3%), while the NVRG sample is much more white (85.7%). Finally, the GP and YT samples are similar in terms of partisan affiliation, while the NVRG sample is much more Republican (62.1% identify as Republicans compared to 28.8% for the GP sample and 25.5% for the YT sample).

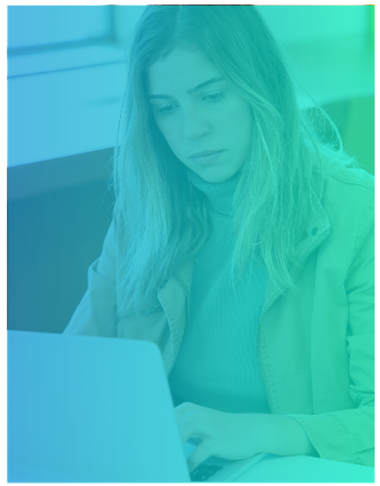
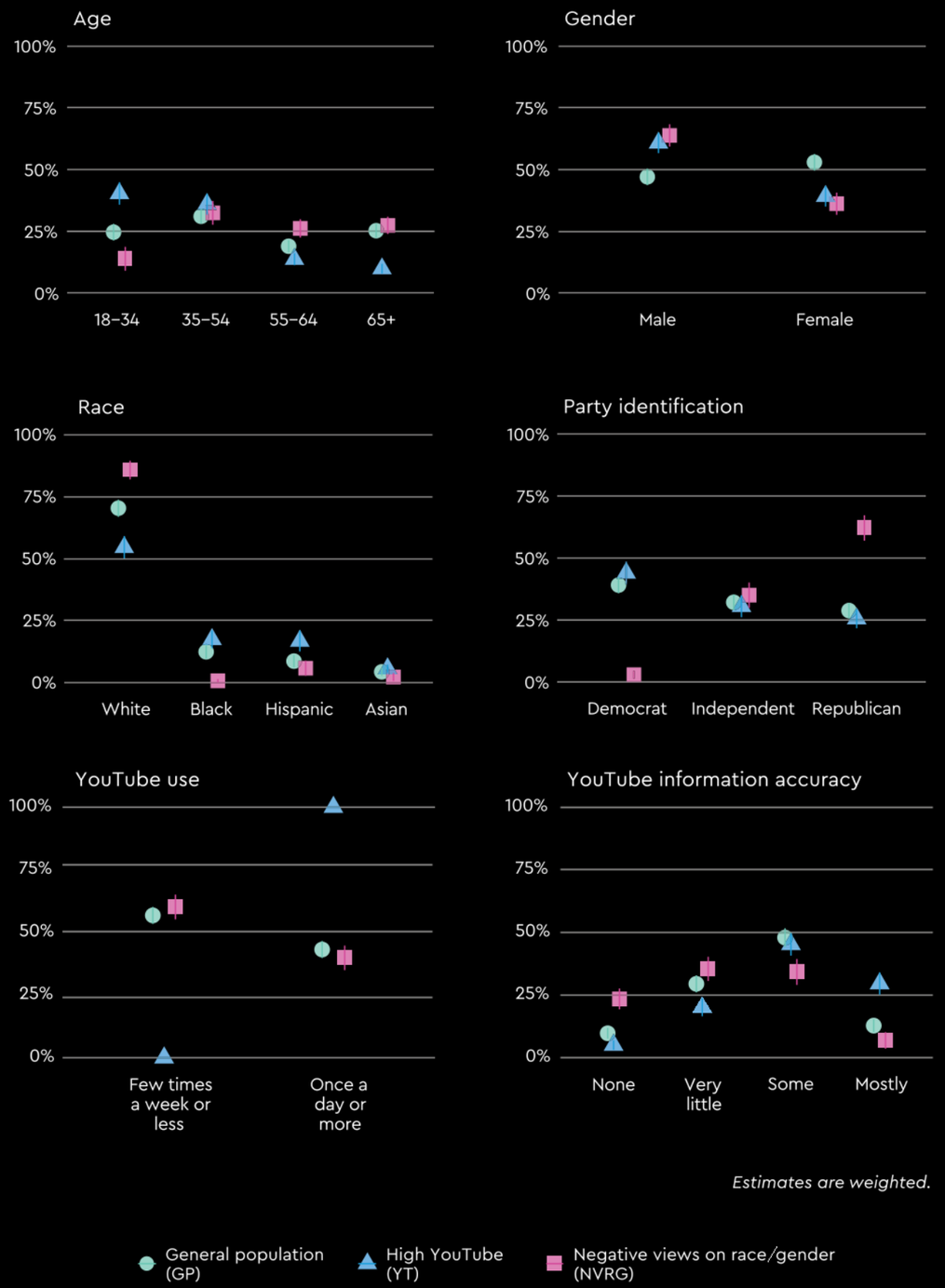


Figure 1: Survey participant characteristics by sample group



In our survey, we asked respondents questions about how often they used YouTube and how much they trust different aspects of the information they see on YouTube. Overall, all three samples reported watching YouTube regularly. By construction, the YT sample exclusively included respondents who said they used YouTube once a day or more: either "almost constantly" (36%) or "several times a day" (64%). For other samples, YouTube use was still fairly high—39.9% of the GP sample reported using YouTube once a day or more, as did 36.0% of the NVRG sample.

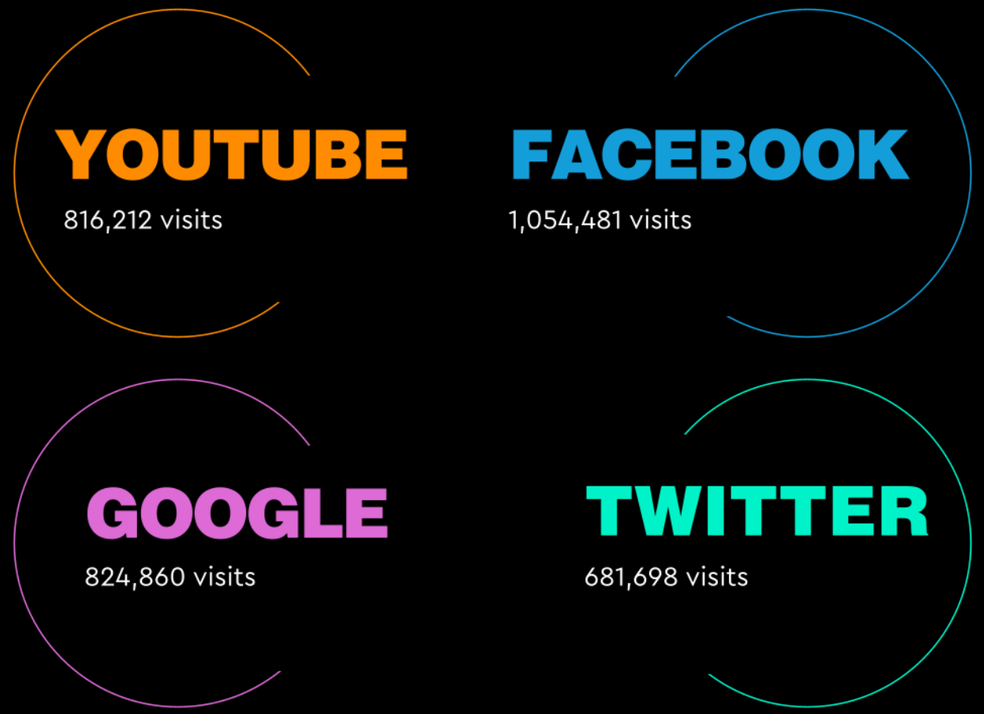
In general, respondents reported moderately high trust in content from YouTube. Only 13% of respondents said that most or all ("all or almost all" or "most") of the information they find using YouTube is "accurate or trustworthy." A plurality of respondents (48.0%) said just "some" of the information they find on YouTube is accurate or trustworthy. The YT sample had the most favorable views towards information found on YouTube, with over one in four (29.6%) saying that it was mostly accurate or trustworthy. The GP and NVRG samples exhibited less trust in YouTube with, respectively, 12.8% and 6.9% saying the same. (Figure A1 in the Appendix provides more information on how respondent views of YouTube vary by YouTube usage levels.) In the analyses that follow, we examine how the consumption of alternative and extremist content on YouTube is associated with survey measures we collected of racial resentment (a four-question scale) and how warmly people feel towards Jews (a 0–100 feeling thermometer).[39]

Web consumption data results

In total, 915 survey respondents accepted our invitation and installed a browser extension to provide data about the websites they visit. Of these, 614 came from our general population sample, 82 from the sample who expressed negative views on race and gender, and 219 from the high YouTube use sample. The respondents who elected to install an extension closely resemble the full sample in their racial and gender composition, but are somewhat younger and more educated. They were also substantially more likely to identify as Democrats. (See Table A1 in the Appendix for further details on the characteristics of our survey respondents and the subset who installed a browser extension.) Finally, participants who installed a browser extension reported using YouTube somewhat more often and expressed somewhat more trust in information from YouTube videos than those who did not install an extension (see Figure A2 in the Appendix).

We consider two types of data from participants: browser history and activity data. Browser history data were collected from the 859 participants for which it was available.[40] When installed, the browser extension collected the past three months of browser history data, providing us with an observation window — the number of days between the first and last data points — that is consistent across browsers (a three-month limit on prior browser history is enforced by Chrome but not present in Firefox). As a result, we have balanced observation windows for users of both browsers that averaged 131 days (SD = 35.2). We also dropped sequential behaviors (e.g., visits to a URL) that occurred less than a second apart, often artifacts of quick page reloads. In such situations, the user does not have time to view the page so we do not count them as separate viewings. After these processing steps, we observed 34.1 million webpage visits from our participants.

We also observed browser activity directly via HTML captured by the extension starting when it was first installed for all of our users' YouTube visits. The observation window for this activity data averaged 64 days (SD = 35.2). Within this window, respondents made 816,212 YouTube visits compared to 1,054,481 visits to Facebook, 824,860 visits to Google, and 681,698 visits to Twitter.



Engagement with potentially radicalizing content on YouTube

Figure 2 summarizes exposure to alternative and extremist YouTube channels in our browser history data. In total, approximately one in five respondents (22.1%) watched a video from an alternative channel, while about one in ten (9.2%) watched a video from an extremist channel.

Almost all of the latter group (85.2%, representing 8.5% of participants) watched videos from alternative channels as well, which is consistent with our expectation that they can serve as potential gateways to extremist content. On average, participants watched 15.0 videos from alternative channels and 2.7 videos from extremist channels. However, these numbers are driven by consumption levels among a small but highly engaged subset of participants. Among those who watched at least one video from either alternative or extremist channels, the mean number of videos watched were 64.2 and 11.5, respectively.

Figure 3 presents another way to visualize the skewed consumption of problematic YouTube videos by plotting the proportion of exposure attributable to users by percentile of consumption. As the figure illustrates, the vast majority of exposure came from a small minority of people. Just 10.9% of participants accounted for 80.5% of exposure to videos from alternative channels. Similarly, only 6.3% of participants accounted for 79.8% of exposure to videos from extremist channels.

Next, we consider which population subgroups are most likely to consume alternative and extremist YouTube channels. We first examined the relationship between expressed levels of racial resentment and video exposure using responses collected in 2018 from a subset of our respondents (i.e., prior to the consumption behavior in question; patterns are similar for 2020 survey measures).

We partitioned the data into three groups that each represent approximately one-third of our sample (terciles), allowing us to compare consumption between people with lower, middle, and higher levels of prior racial resentment.

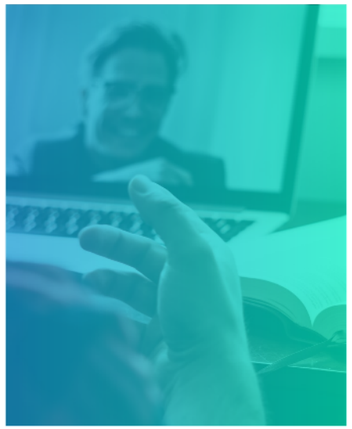
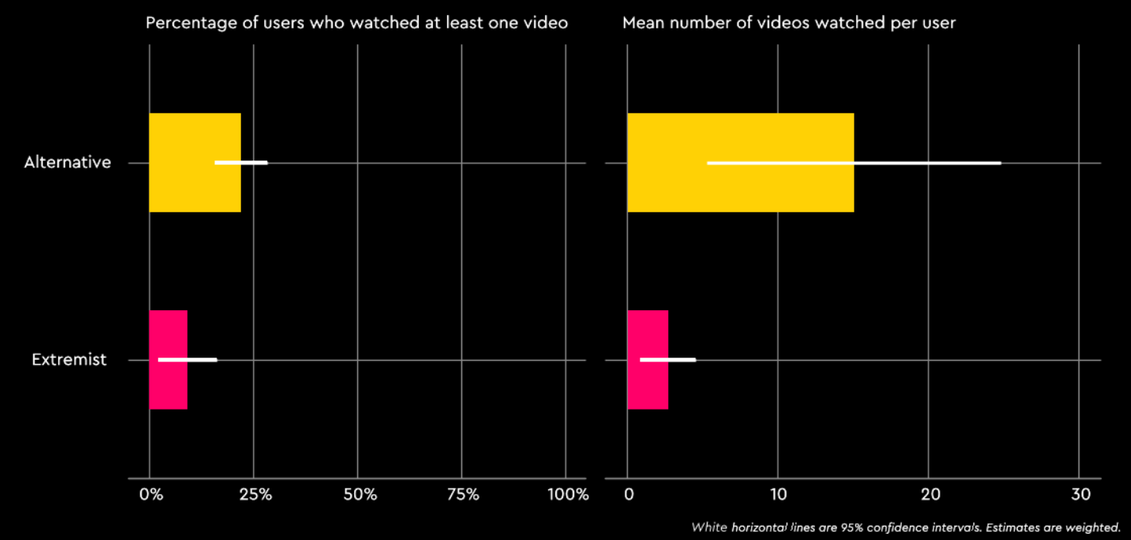
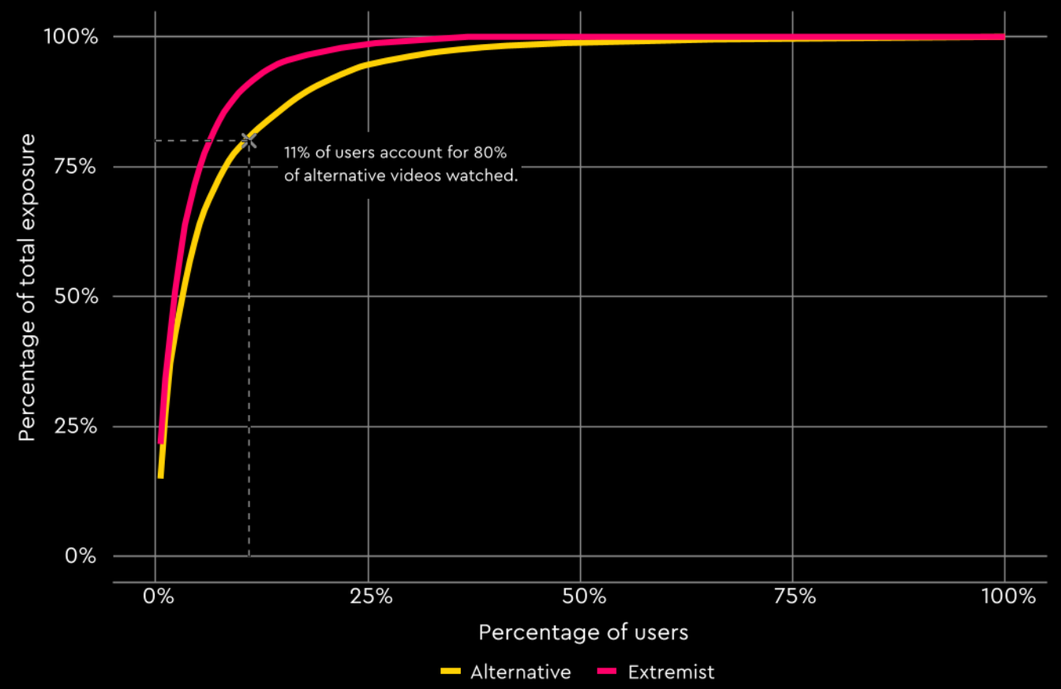


Figure 2: Exposure frequencies and levels for problematic YouTube content



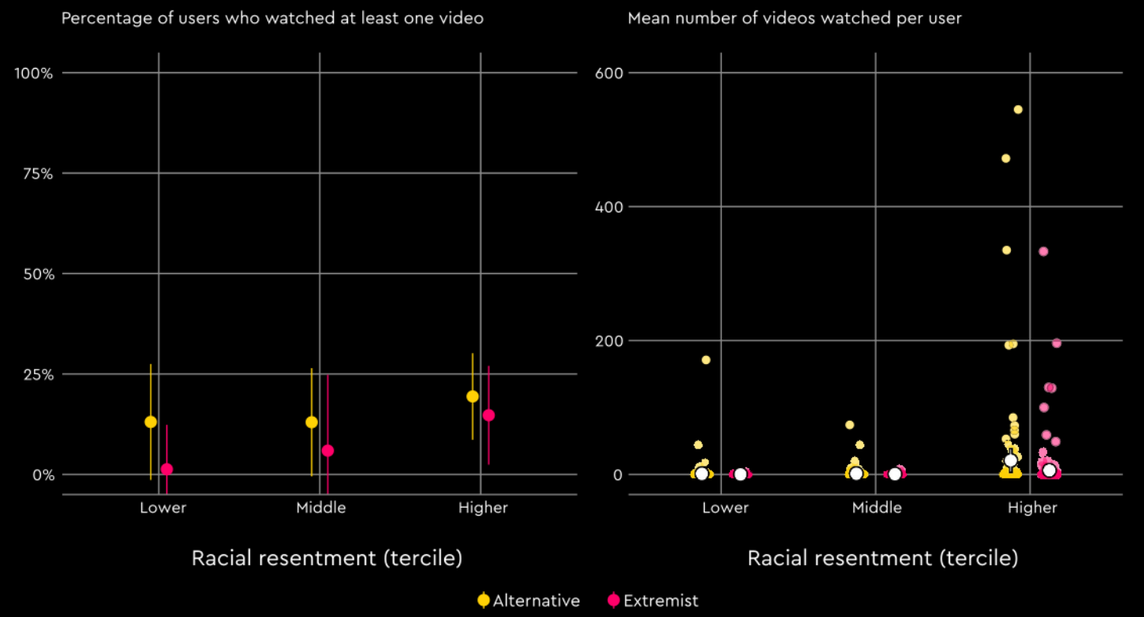
White horizontal lines are 95% confidence intervals. Estimates are weighted.

Figure 3: Concentration of exposure to problematic YouTube content



Estimates are weighted.

Figure 4: Exposure to problematic YouTube content by racial resentment



Estimates are weighted.

As Figure 4 indicates, exposure to videos from alternative and extremist channels increases dramatically for the high tercile (19.4% and 14.7%, respectively) compared to the low (13.1% and 1.3%) and medium groups (13.0% and 5.9%) among the set of respondents for whom prior racial resentment data is available. The mean number of videos watched from alternative and extremist channels is even more skewed when we compare the high racial resentment group (20.6 videos from alternative channels) with the middle and lower groups (1.1 and 0.8, respectively).

Higher racial resentment users watched 5.8 videos from extremist channels on average compared to the middle and lower groups who averaged 0.38 and 0.03, respectively. As these figures suggest, people who are higher in racial resentment are the primary audience for videos from alternative and extremist channels.

Overall, people who exhibited more racial resentment (i.e., those with scores in the top tercile of the sample) were responsible for 93.7% of views for videos from alternative channels and 95.1% of views for videos from extremist channels.

The pattern is less clear for expressed sentiment toward Jews, which we measure in our 2020 survey. Figure 5 presents exposure statistics to videos from alternative and extremist channels by expressed feelings toward Jews on a 0–100 feeling thermometer. We aggregate these scores into 0–40 ("cold"), 41–60 ("medium"), and 61–100 ("warm") groups.

Exposure to alternative and extremist channels tends to be higher among people with cold feelings towards Jews, a group that represented 5.4% of participants—24.9% in this group watched at least one video from an alternative channel and 20.7% watched at least one video from an extremist channel compared to 19.0% and 7.9%, among the participants who expressed medium and warm feelings towards Jews, respectively (67.3% of the sample).

Surprisingly, the opposite pattern emerges in the right-hand panel of Figure 5; the mean number of videos watched from alternative channels is greater for those who express warm feelings towards Jews (16.9) than it is among the medium (14.9) or cold (7.3) groups. Such a relationship could reflect high viewership for pro-Israel channels like those of Ben Shapiro and Stephen Crowder. The average number of videos consumed from extremist channels is also marginally higher for the warm group (3.6) compared to the medium (1.2) and cold (1.6) groups. The relationship between feelings toward Jews and exposure to videos from alternative and extremist channels is thus less clear than it is for racial resentment.

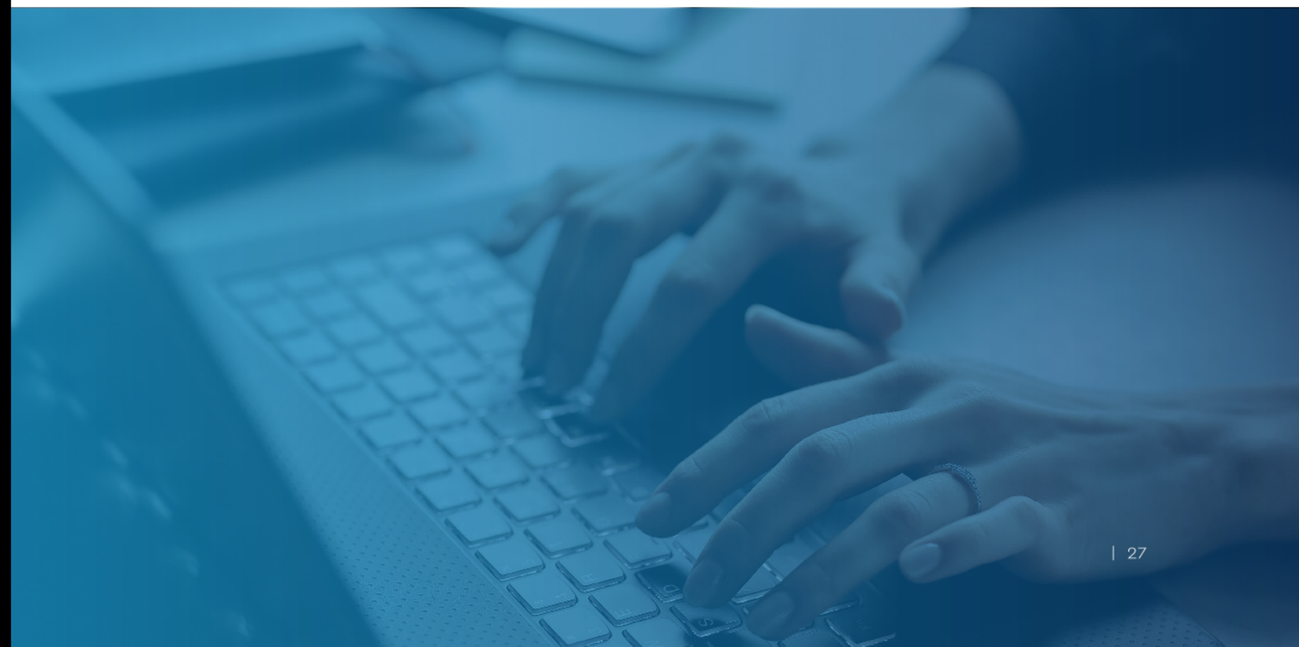
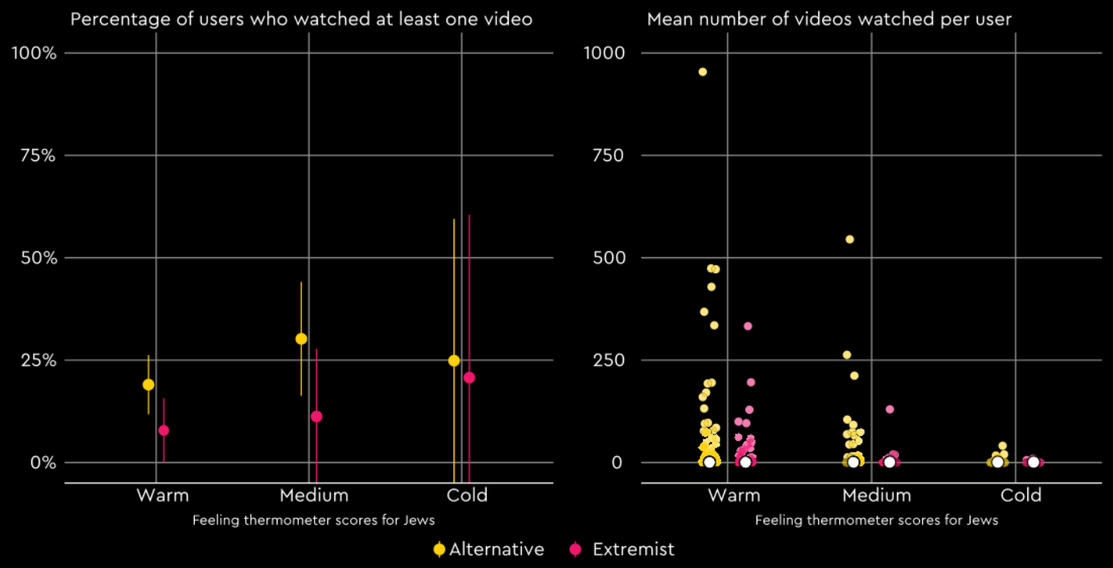


Figure 5: Exposure to problematic YouTube content by feelings toward Jews



Warm indicates a feeling thermometer score for Jews of 61–100 (67.3%), medium indicates a score of 41–60 (27.3%), cold indicates a score of 0–40 (5.4%). Estimates are weighted.

18.9%

Participants appeared to follow a recommendation 18.9% of the time they were shown them.

9.8%

Follows were most common for the top-ranked recommendation, but only took place about 1 in 10 times.

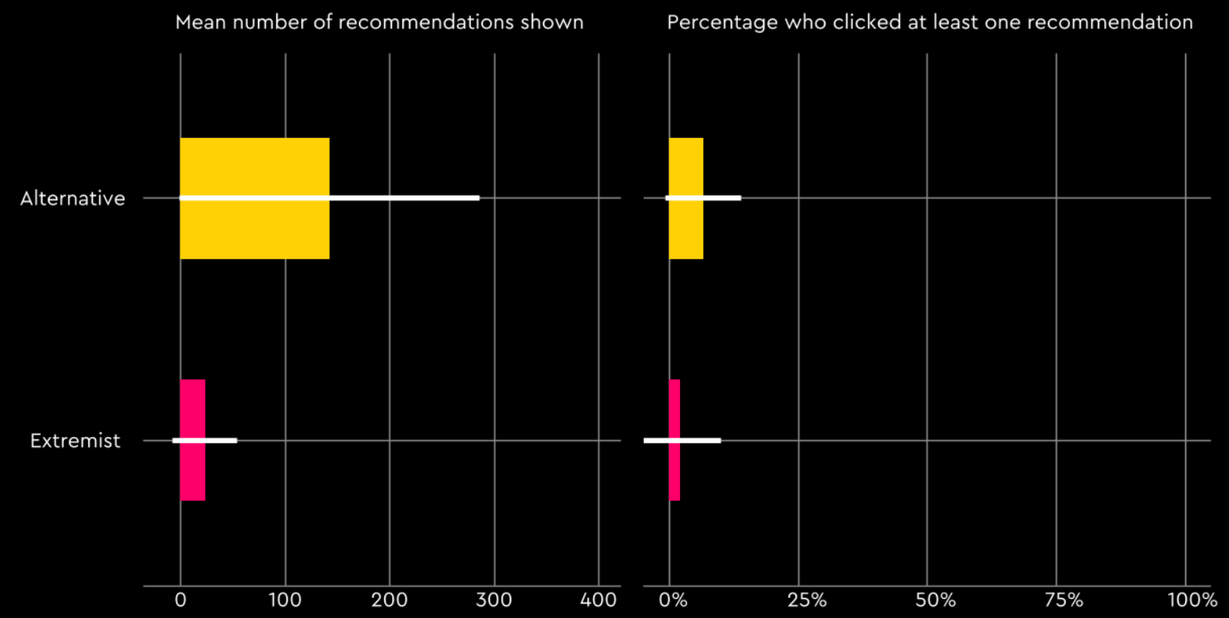
These recommendations frequently included problematic content. As Figure 6 shows, participants frequently received recommendations for videos from alternative and extremist channels over the three months in which we collected browser activity data: on average, participants were recommended 142 videos from alternative channels, and 23 videos from extremist channels. Just under seven percent (6.6%) of the sample followed at least one recommendation to an alternative channel and 2.1% followed at least one to an extremist channel.

Recommendations of Alternative and Extremist Videos

Within our observation window, respondents made 816,212 YouTube visits, including 535,438 visits to video pages. We captured the HTML from 228,038 of those visits and extracted 4,071,364 recommendations that users were exposed in the right hand sidebar—a median of 20 recommendations per video view (mean = 17.9, SD = 9.8). We analyzed the content of these recommendations and subsequent user behavior to evaluate the success of the changes that Google made to YouTube in 2019, when it announced that it had "launched over 30 different changes to reduce recommendations of borderline content and harmful misinformation." It was part of a series of changes that resulted in a "70% average drop in watch time of this content coming from non-subscribed recommendations."[41]

YouTube recommendations can consist of links to other YouTube videos, ads linking to other YouTube videos, and ads linking to external domains. We consider both subscribed and non-subscribed recommendations as it is not currently possible to distinguish between them. Participants appeared to follow a recommendation 18.9% of the time they were shown them. In general, because each video watched typically comes with multiple videos recommended in the right-hand panel (median 20), the likelihood that any single recommended video is followed is quite low. Follows were most common for the top-ranked recommendation, but only took place about 1 in 10 times (9.8%), suggesting that users rarely allow YouTube to auto-play the next video.

Figure 6: YouTube recommendations of alternative and extremist channels

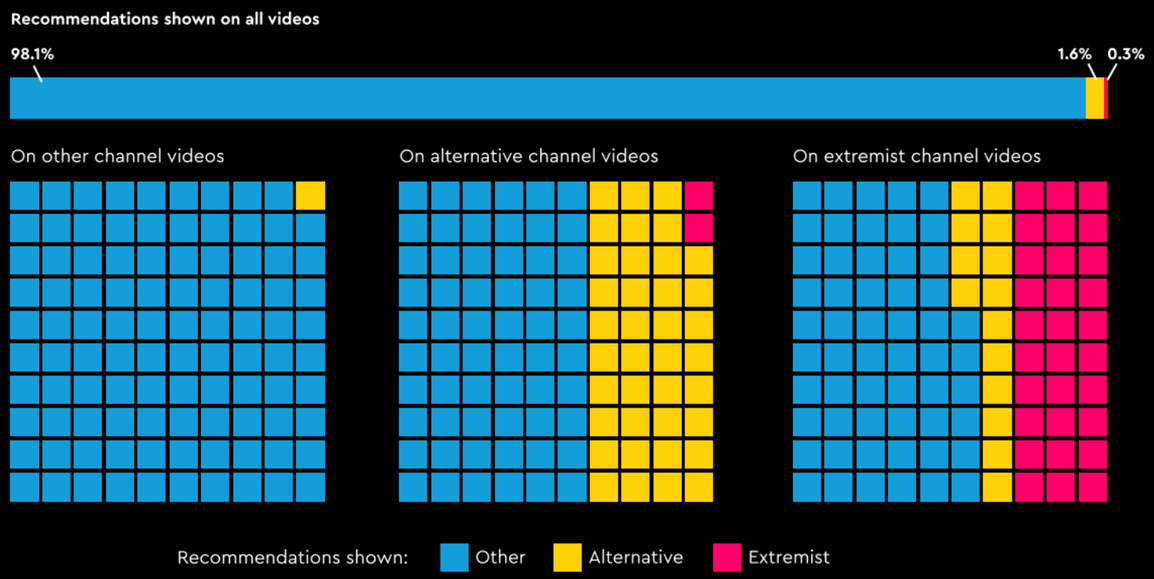


Estimates are weighted.

As shown in Figure 7, of the total number of recommendations users saw, recommendations to videos that are not from alternative or extremist channels (98.1%) greatly exceed those to alternative (1.6%) and extremist channels (0.3%). This follows from the fact that visits to alternative and extremist content are rare to begin with (2.8% and 0.4% of total visits, respectively). Generally, users who started on a video that was not from an alternative or extremist channel were recommended videos from alternative channels in 1.1% of visits and were recommended videos from extremist channels in 0.1% of visits. However, when people watched videos from alternative channels, the proportion of recommendations to potentially harmful content was higher: 37.6% of recommendations were to videos from alternative channels, and 2.3% were to videos from extremist channels. Similarly, the proportion of recommendations to videos from extremist channels is higher when watching videos from extremist channels: 29.3% of recommendations were to other videos from extremist channels, and 14.3% were to videos from alternative channels.

As with the recommendations shown to users, the recommendations that participants followed were overwhelmingly to non-alternative or extremist content.

Figure 7: Recommendations by type of YouTube content visited



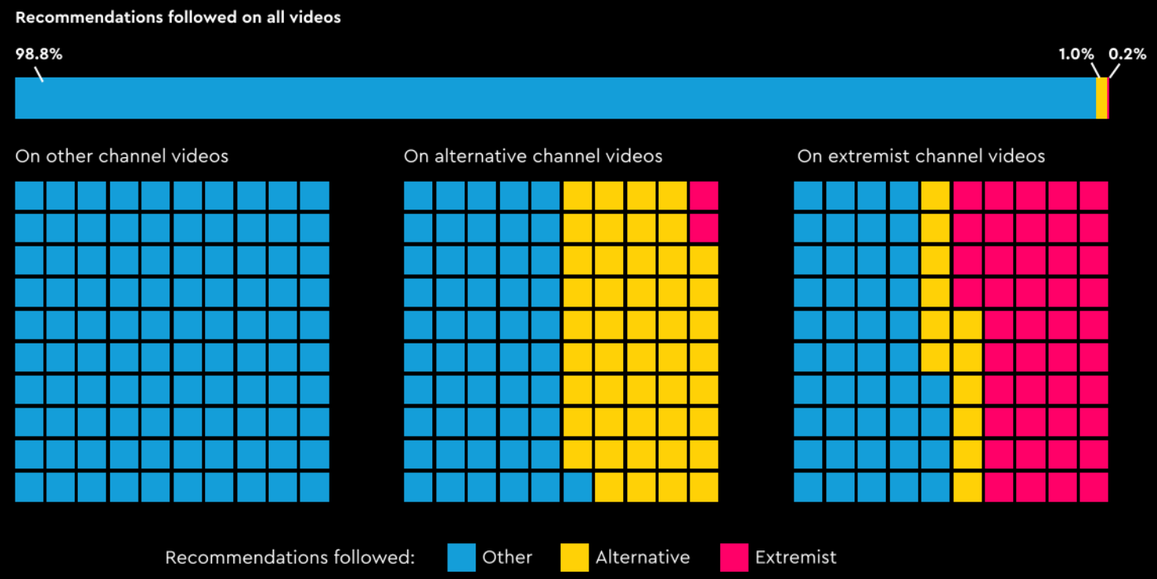
Colored tiles are proportional to the type of recommendation shown after watching other, alternative, or extremist content.

Overall, 98.8% of recommendations that participants followed were to other types of content compared to 1.0% for videos from alternative channels and 0.2% for videos from extremist channels. The proportion of recommendations followed for videos from alternative and extremist channels is thus somewhat lower than the proportion of recommendations participants received for these types of videos.

Figure 8 disaggregates recommendations by whether they were followed. The leftmost panel shows that 99.6% of the recommendations participants followed while viewing content that was not from alternative or extremist channels were to other videos of that type (compared to 0.3% and 0.1% for videos from alternative and extremist channels, respectively). By contrast, 46.8% of the recommendations followed from videos from alternative channels led to other videos from alternative channels and 2.1% to videos from extremist channels.

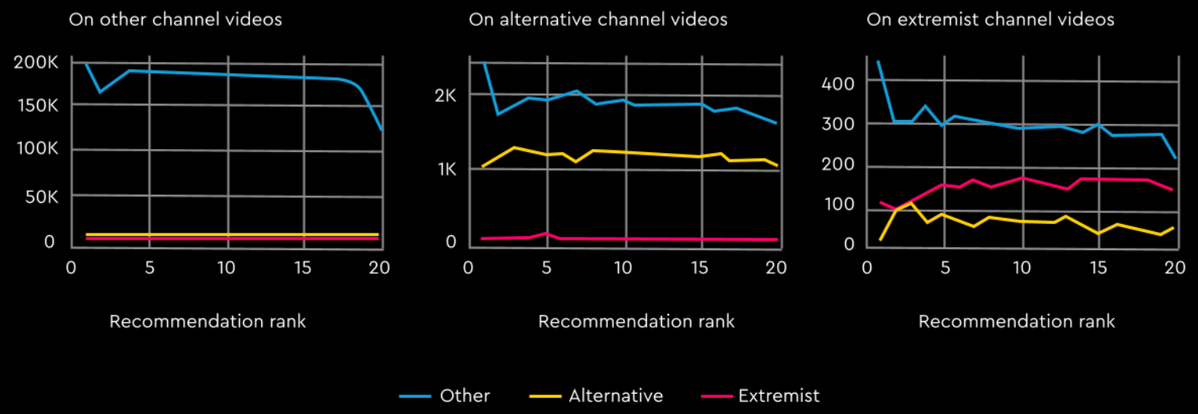
During visits to videos from extremist channels, the share of extremist recommendations followed is equal to the share that was followed to other types of content (both 44.3%), with the remaining 11.4% leading to alternative channels.

Figure 8: Recommendations followed by type of YouTube content visited



Colored tiles are proportional to the type of recommendation followed after watching other, alternative, or extremist content.

Figure 9: YouTube recommendations by content type



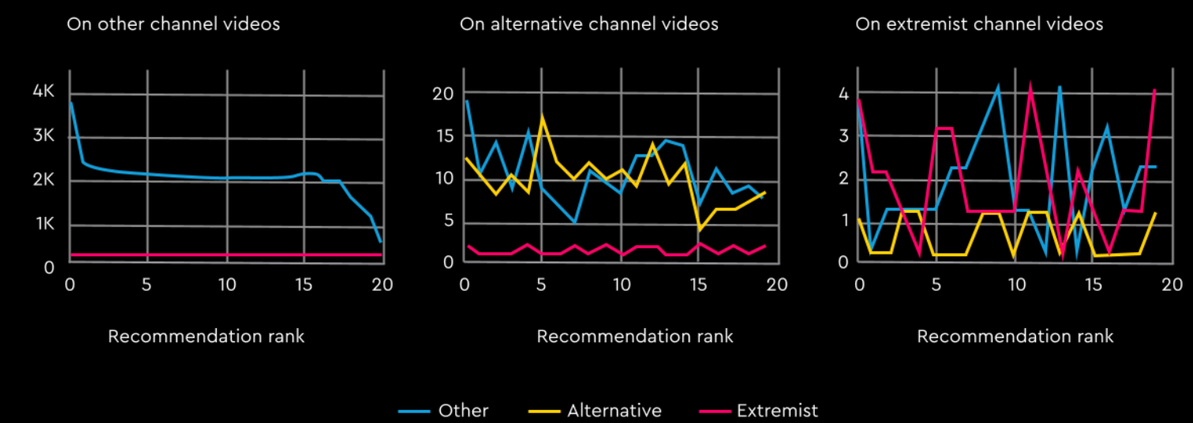
We next consider the relationship between the videos respondents watched and the algorithmically ranked recommendations they were shown. As in the preceding figures, Figure 9 disaggregates the number of video recommendations participants were exposed to by type and rank for visits to videos not from alternative or extremist channels (left panel), visits to videos from alternative channels (center panel), and visits to videos from extremist channels (right panel). For visits to other types of content, the proportion of recommendations to videos from alternative and extremist channels is low (1.1% and 0.1%, respectively), especially at the first recommendation ranking.

However, exposure to such recommendations is more common during visits to videos from alternative (39.9%) and extremist (43.6%) channels, though recommendations for potentially harmful content rarely enter the top-ranked positions which draw the most attention. As the center panel shows, when people watch videos from alternative channels, YouTube is more likely to recommend videos from alternative (37.6%) and extremist (2.3%) channels than it is to recommend other types of content (left panel). Most alarmingly, viewers of videos from extremist channels get almost as many recommendations for videos from alternative and extremist channels as they do for other types of content. The recommendations to videos from extremist channels (29.3%) in this case outnumber those to videos from alternative channels (14.3%).

Exposure to alternative and extremist videos in recommendations is more common during visits to videos from alternative **(39.9%)** and extremist **(43.6%)** channels.

Most alarmingly, viewers of videos from extremist channels get almost as many recommendations for videos from alternative and extremist channels as they do for other types of content.

Figure 10: YouTube recommendation follows by content type



We now consider how many recommendations are followed by content type and recommendation rank. Figure 10 presents the number of video recommendations participants followed by type and rank for visits to content that is neither alternative nor extremist (left panel), visits to videos from alternative channels (center panel), and visits to videos from extremist channels (right panel). In the left panel, higher-ranked recommendations are most likely to be followed (especially the first recommendation, which is auto-played if the user takes no other action), and the proportion of people who followed alternative or extremist recommendations was under 1%. However, when we consider the small number of people who followed recommendations that accompanied videos published by alternative or extremist channels, we find that many followed them to videos from similar sources.

Overall, 46% of recommendations followed when watching a video from an alternative channel went to another video from an alternative channel, and 44% of follows when watching a video from an extremist channel went to other videos from extremist channels—almost as much as the other types of YouTube videos put together.

CONCLUSIONS

Overall, our findings indicate that YouTube plays an important role in exposing people to potentially harmful content. Using comprehensive individual-level behavioral data, we find that exposure to alternative YouTube channels that can serve as gateways to more extreme forms of content and to extremist or white supremacist channels is disturbingly common among a group of Americans. Approximately one in ten participants (9.2%) viewed at least one video from an extremist channel and approximately two in ten (22.1%) viewed at least one video from an alternative channel. When they watch these videos, participants were more likely to see and follow recommendations to similar videos.

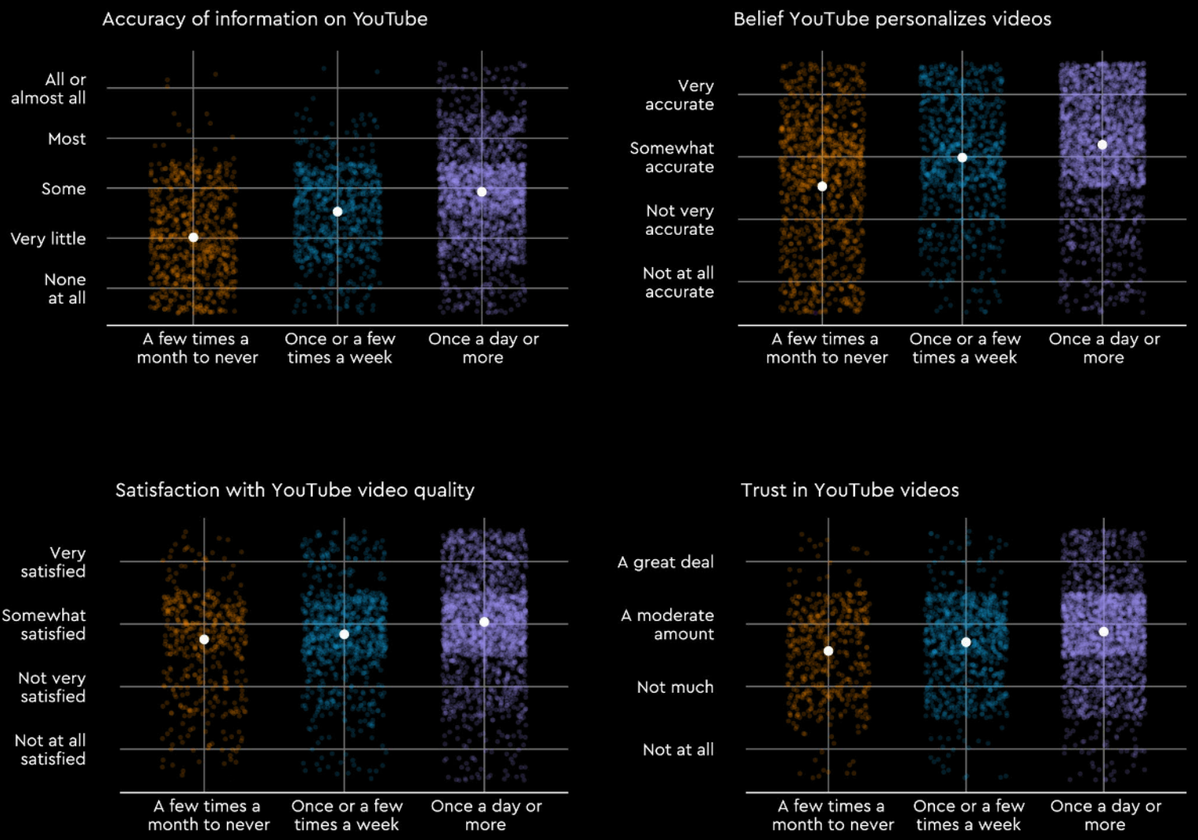
We do not find clear evidence that YouTube frequently exposes people with neutral or mixed views on issues such as race to alternative or extremist content. Instead, the audience for videos from alternative or extremist channels is dominated by people who already have high levels of racial resentment (more than 90% of views for both types). In addition, average consumption levels among these viewers are very high—means of 20.6 videos from alternative channels and 5.8 videos from extremist channels, respectively. Finally, we find that recommendations to potentially harmful videos from other types of videos are rare, but the recommendations that are shown alongside videos from alternative or extremist channels frequently include videos of both types. These recommendations may increase viewership among people who find such videos appealing whether they subscribe to the channels in question or not.

We must of course acknowledge several key limitations of our study. First, we only consider the videos that our participants viewed and the recommendations that they were shown. Future studies will evaluate algorithmic personalization in further detail by comparing personalized and anonymized search results for our participants. Second, though our participants were recruited from a nationally representative survey and are the largest and most diverse sample of this kind, our participants are not fully representative. Third, we acknowledge that our results may be time-sensitive; YouTube's algorithm and behavior on the platform have changed over time in important ways.[42] Fourth, we cannot distinguish between subscribed and non-subscribed recommendations. Finally, we note that all analyses here depend on channel lists compiled by scholars and subject matter experts. Though these are the most complete set of alternative and extremist YouTube channels assembled to date, further research should conduct analysis of exposure at the video level when possible.

Despite these caveats, our study has important implications for both YouTube policies and the study of exposure to potentially harmful content online. Most notably, the results indicate that the much-touted changes in YouTube's handling of "borderline content and harmful misinformation" did not eliminate the problem.[43] A highly active subset of Americans not only continue to watch many videos from alternative and extremist supremacist channels, but are often shown recommendations to further videos from those channels. These findings are cause for significant concern given the high levels of racial resentment that viewers express and the potential effects of the videos they see.

APPENDIX

Figure A1: Perceptions of YouTube by self-reported usage levels



Estimates are weighted.

Figure A2: YouTube usage and opinions by browser extension installation

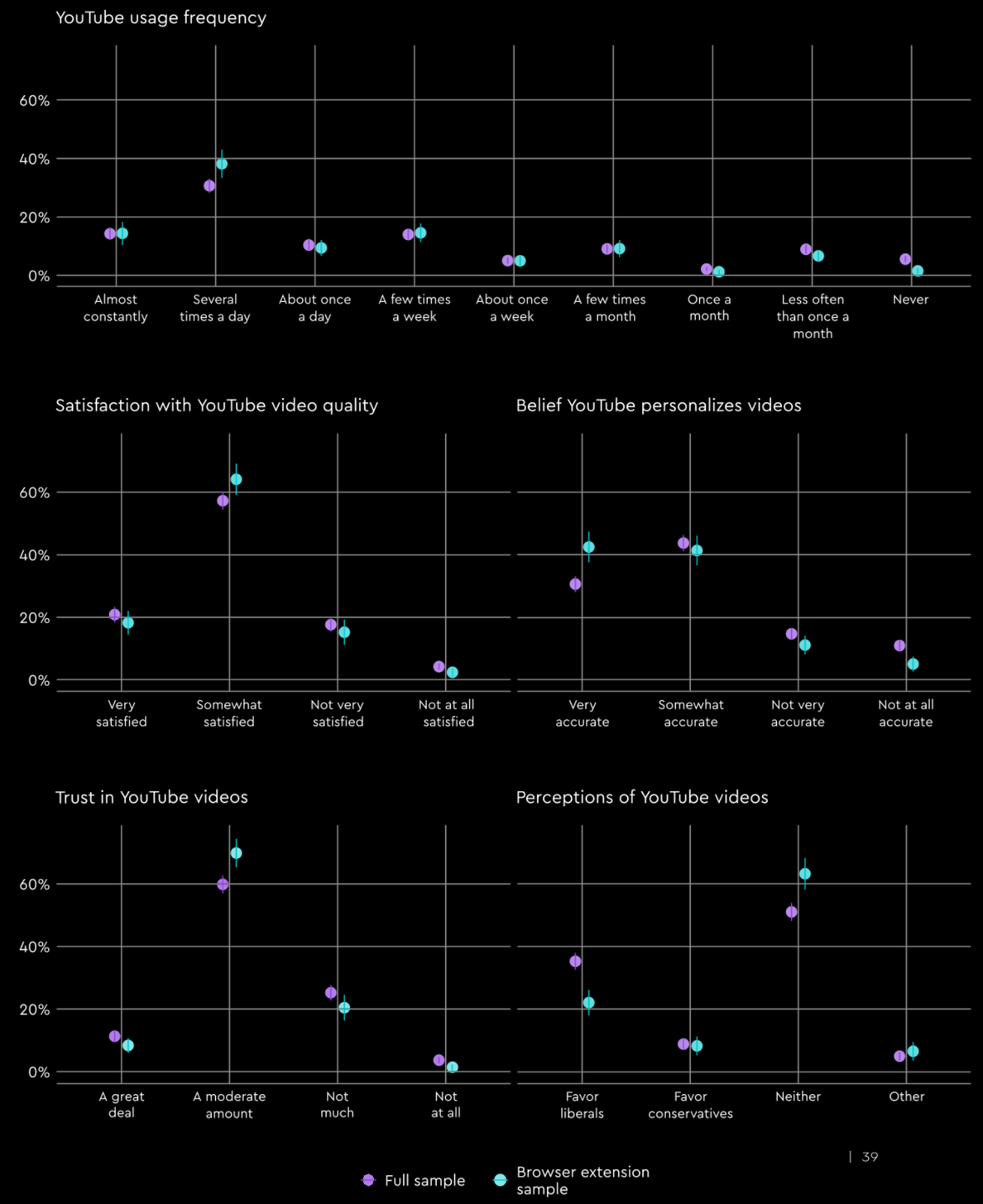


Table A1: Sample demographics by browser extension installation

		Full	Extension
Gender	Male	0.55	0.53
		(0.01)	(0.02)
	Female	0.45	0.47
		(0.01)	(0.02)
Race	White	0.70	0.71
		(0.01)	(0.02)
	Black	0.11	0.14
		(0.01)	(0.02)
	Hispanic	0.10	0.08
		(0.01)	(0.02)
	Asian	0.04	0.03
		(0.01)	(0.02)
2016 Presidential vote	Donald Trump	0.40	0.21
		(0.01)	(0.02)
	Hillary Clinton	0.23	0.39
		(0.01)	(0.02)
Education	High School Graduate	0.36	0.26
		(0.01)	(0.03)
	Some College	0.35	0.35
		(0.01)	(0.02)
	Bachelor's Degree	0.19	0.25
		(0.01)	(0.02)
	Post-Graduate Degree	0.1	0.14
		(0.00)	(0.01)
Party identification	Democrat	0.31	0.50
		(0.01)	(0.03)
	Independent	0.32	0.28
		(0.01)	(0.02)
	Republican	0.36	0.22
		(0.01)	(0.02)
Age	18–34	0.26	0.34
		(0.01)	(0.03)
	35–54	0.33	0.33
		(0.01)	(0.02)
	55–64	0.19	0.16
		(0.01)	(0.01)
	65+	0.22	0.17
		(0.01)	(0.02)
		n=4000	n=915

Weighted statistics are estimated using YouGov survey weights. Standard errors are in parentheses.

Table A2: Full sample demographics with and without survey weights

		Weighted	Unweighted
Gender	Male	0.55	0.54
		(0.01)	(0.02)
	Female	0.45	0.46
		(0.01)	(0.01)
Race	White	0.70	0.76
		(0.01)	(0.01)
	Black	0.11	0.08
		(0.01)	(0.00)
	Hispanic	0.10	0.07
		(0.01)	(0.00)
	Asian	0.04	0.04
		(0.01)	(0.00)
2016 Presidential vote	Donald Trump	0.40	0.40
		(0.01)	(0.01)
	Hillary Clinton	0.23	0.31
		(0.01)	(0.01)
Education	High School Graduate	0.36	0.19
		(0.01)	(0.01)
	Some College	0.35	0.37
		(0.01)	(0.01)
	Bachelor's Degree	0.19	0.26
		(0.01)	(0.02)
	Post-Graduate Degree	0.10	0.18
		(0.00)	(0.01)
Party identification	Democrat	0.31	0.35
		(0.01)	(0.01)
	Independent	0.32	0.32
		(0.01)	(0.01)
	Republican	0.36	0.33
		(0.01)	(0.01)
Age	18–34	0.26	0.16
		(0.01)	(0.01)
	35–54	0.33	0.34
		(0.01)	(0.01)
	55–64	0.19	0.23
		(0.01)	(0.01)
	65+	0.22	0.27
		(0.01)	(0.01)
		n=4000	n=4000

Weighted statistics are estimated using YouGov survey weights. Standard errors are in parentheses.

Table A3: Browser extension sample demographics with and without survey weights

		Weighted	Unweighted
Gender	Male	0.53	0.51
		(0.02)	(0.02)
	Female	0.47	0.49
		(0.02)	(0.02)
Race	White	0.71	0.76
		(0.02)	(0.01)
	Black	0.14	0.09
		(0.02)	(0.01)
	Hispanic	0.08	0.06
		(0.02)	(0.01)
	Asian	0.03	0.04
		(0.01)	(0.01)
2016 Presidential vote	Donald Trump	0.21	0.19
		(0.02)	(0.01)
	Hillary Clinton	0.39	0.51
		(0.02)	(0.02)
Education	High School Graduate	0.26	0.12
		(0.03)	(0.01)
	Some College	0.35	0.34
		(0.02)	(0.02)
	Bachelor's Degree	0.25	0.29
		(0.02)	(0.02)
	Post-Graduate Degree	0.14	0.24
		(0.01)	(0.01)
Party identification	Democrat	0.50	0.56
		(0.03)	(0.02)
	Independent	0.28	0.27
		(0.02)	(0.02)
	Republican	0.22	0.17
		(0.02)	(0.01)
Age	18–34	0.34	0.21
		(0.03)	(0.01)
	35–54	0.33	0.37
		(0.02)	(0.02)
	55–64	0.16	0.23
		(0.01)	(0.01)
	65+	0.17	0.19
		(0.02)	(0.01)
		n=915	n=915

Weighted statistics are estimated using YouGov survey weights. Standard errors are in parentheses.

FOOTNOTES

1. Google, "The Four Rs of Responsibility, Part 2: Raising Authoritative Content and Reducing Borderline Content and Harmful Misinformation," December 3, 2019, https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/.
2. We drew our lists of alternative (e.g., "the Alternative Influence Network," "Intellectual Dark Web," etc.) and extremist (e.g., white supremacist, "alt-right," "white identitarian," etc.) channels from previous research by scholars and subject matter experts. Details are provided on page 7 below.
3. Racial resentment is measured via a four-question scale (Kinder and Sanders 1996).
4. Google, "The Four Rs of Responsibility."
5. Andrew Perrin and Monica Anderson, "Share of U.S. Adults Using Social Media, Including Facebook, Is Mostly Unchanged since 2018," Pew Research Center (Pew Research Center, April 10, 2019), https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/.
6. Paul Covington, Jay Adams, and Emre Sargin, "Deep Neural Networks for YouTube Recommendations," Proceedings of the 10th ACM Conference on Recommender Systems – RecSys '16, 2016.
7. Donald Horton and R. Richard Wohl, "Mass Communication and Para-Social Interaction," Psychiatry 19, no. 3 (August 1956): 215–29.
8. Arienne Ferchaud et al., "Parasocial Attributes and YouTube Personalities: Exploring Content Trends Across the Most Subscribed YouTube Channels," Computers in Human Behavior 80 (March 2018): 88–96.
9. Becca Lewis, "Alternative Influence," Data & Society, September 18, 2018, https://datasociety.net/library/alternative-influence.
10. Joan E. Solsman, "YouTube's AI Is the puppet master over most of what you watch," CNET (CNET, January 10, 2018), https://www.cnet.com/news/youtube-ces-2018-neal-mohan/.
11. Marc Faddoul, Guillaume Chaslot, and Hany Farid, "A longitudinal analysis of YouTube's promotion of conspiracy videos," 2020, https://arxiv.org/pdf/2003.03318.pdf.
12. Zeynep Tufekci, "Opinion | YouTube, the Great Radicalizer," The New York Times, March 10, 2018, https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html.
13. Kevin Roose, "The Making of a YouTube Radical," The New York Times, June 8, 2019, https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html.
14. Jack Nicas, "How YouTube Drives People to the Internet's Darkest Corners," WSJ (Wall Street Journal, February 7, 2018), https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478.
15. Roose, "The Making."
16. Ribeiro, Manoel Horta et al., "Auditing Radicalization Pathways on YouTube," 2019, https://arxiv.org/abs/1908.08313.
17. Mark Ledwich and Anna Zaitsev, "Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization," 2019, https://arxiv.org/abs/1912.11211.
18. Kevin Munger and Joseph Phillips, "Right-Wing YouTube: A Supply and Demand Perspective," The International Journal of Press/Politics, October 21, 2020. 3.
19. Google, "The Four Rs of Responsibility."
20. Buntain, Cody et al., "YouTube Recommendations and Effects on Sharing Across Online Social Platforms," ArXiv:2003.00970 [Cs], July 20, 2020, https://arxiv.org/abs/2003.00970.
21. Faddoul, Chaslot, and Farid, "A longitudinal analysis of YouTube's promotion of conspiracy videos."
22. Eslam Hussein, Prerna Juneja, and Tanushree Mitra, "Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube," Proceedings of the ACM on Human-Computer Interaction 4, no. CSCW1 (May 28, 2020): 1–27.

FOOTNOTES

23. Kostantinos Papadamou, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, Michael Sirivianos, "'It is just a flu': Assessing the Effect of Watch History on YouTube's Pseudoscientific Video Recommendations," 2020, https://arxiv.org/pdf/2010.11638.pdf.
24. Homa Hosseinmardi, Amir Ghasemian, Aaron Clauset, David M. Rothschild, Markus Mobius, and Duncan J. Watts, "Evaluating the scale, growth, and origins of right-wing echo chambers on YouTube," 2020, https://arxiv.org/pdf/2011.12843.pdf.
25. Measuring YouTube exposure data on mobile devices is not currently technically feasible but is an important goal for future research.
26. Lewis, "Alternative Influence."
27. Ribeiro, "Auditing Radicalization Pathways."
28. Ledwich, "Algorithmic Extremism."
29. Any channels that met the criteria for both the alternative and extremist/white supremacist channel lists were included only in the latter.
30. Christopher Charles, "(Main)Streaming Hate: Analyzing White Supremacist Content and Framing Devices on YouTube" (2020), https://stars.library.ucf.edu/etd2020/27/.
31. Ribeiro, "Auditing Radicalization Pathways."
32. Ledwich, "Algorithmic Extremism."
33. Aaron Sankin, "YouTube Said It Was Getting Serious About Hate Speech. Why Is It Still Full of Extremists?," Gizmodo, July 25, 2019, https://gizmodo.com/youtube-said-it-was-getting-serious-about-hate-speech-1836596239.
34. While many of these channels were purged in June 2019 and June 2020 (see https://www.theverge.com/2019/6/5/18652576/youtube-supremacist-content-ban-borderline-extremist-terms-of-service https://variety.com/2020/digital/news/youtube-bans-stefan-molyneux-david-duke-richard-spencer-hate-speech-1234694079/), we cannot definitely conclude that participants did not see a video from one of these channels. Since we cannot fetch channel information for deleted videos or videos from deleted channels, it is plausible that participants watched a video from one of these channels before it was removed and before we were able to link the video to any channel.
35. See, for example, Andrew M. Guess, Brendan Nyhan, and Jason Reifler, "Exposure to Untrustworthy Websites in the 2016 US Election," Nature Human Behaviour, March 2, 2020.
36. Dartmouth CPHS STUDY00032001, Northeastern IRB #20-03-04, Princeton IRB #12442, University of Exeter Social Sciences and International Studies Ethics Committee #201920-111.
37. We measure these views using participants' prior responses to a standard four-question measure of racial resentment (Kinder and Sanders 1996), two questions from a scale measuring perceptions of racism (DeSante and Smith 2020), and two questions measuring perceptions of sexism (Glick and Fiske 1995). Respondents were eligible for inclusion in the oversample if they scored in the top decile of 2018 Cooperative Congressional Election Study participants in an IRT model of these responses.
38. These statistics and all analyses below include survey weights provided by YouGov.
39. We create a racial resentment scale by combining respondents' answers to the standard four-question measure of racial resentment by Kinder and Sander (1996). Attitudes towards Jews are captured with a feeling thermometer.
40. Participants may have cleared their browsing history, or permanently disabled collection of browsing history, prior to the installation of our browser extension.
41. Google, "The Four Rs of Responsibility."
42. E.g. Munger, "Right-Wing YouTube."
43. Google, "The Four Rs of Responsibility."

SUPPORT

Author Acknowledgments

We are grateful to the Anti-Defamation League, Russell Sage Foundation, Carnegie Corporation of New York, and the National Science Foundation for financial support and to Samantha Luks at YouGov for outstanding survey administration assistance. We also would like to thank Virgílio A. F. Almeida, Christopher Charles, Mark Ledwich, Becca Lewis, Wagner Meira, Raphael Ottoni, Manoel Horta Ribeiro, Aaron Sankin, Robert West, and Anna Zaitsev for sharing their data with us or making it publicly available. All conclusions and errors are our own.

This research utilized 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 NSF.

This work is made possible in part by the generous support of:

Anonymous
Anonymous
The Robert Belfer Family
Dr. Georgette Bennett
Catena Foundation
Craig Newmark Philanthropies
The David Tepper Charitable Foundation Inc.
The Grove Foundation
Joyce and Irving Goldman Family Foundation
Horace W. Goldsmith Foundation
Walter & Elise Haas Fund
Luminate
One8 Foundation
John Pritzker Family Fund
Qatalyst Partners
Quadrivium Foundation
Righteous Persons Foundation
Riot Games
Amy and Robert Stavis
Zegar Family Foundation

The Belfer Fellowship

The Belfer Fellowship was established by the Robert Belfer Family to support innovative research and thought-leadership on combating online hate and harassment for all. Fellows are drawn from the technologist community, academia, and public policy to push innovation, research and knowledge development around the online hate ecosystem. ADL and the Center for Technology and Society thank the Robert Belfer Family for their dedication to our work, and their leadership in establishing the Fellows program.

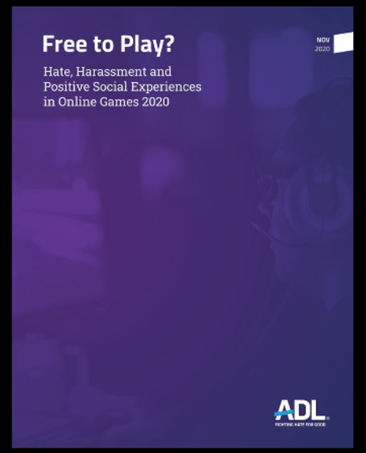
TAKE ACTION

Partner with ADL to fight hate in your community and beyond.

- Sign up at adl.org for our email newsletters to stay informed about events in our world and ADL's response.
- Report hate crimes and bias-related incidents in your area to your regional ADL office.
- Engage in respectful dialogue to build understanding among people with different views.
- Get involved with ADL in your region.

FEATURED RESOURCES

FROM THE ADL CENTER FOR TECHNOLOGY AND SOCIETY



Free to Play?
Hate, Harassment and Positive Social Experiences in Online Games 2020

www.adl.org/online-hate-2020

Online Hate and Harassment
The American Experience 2020

www.adl.org/online-hate-2020

The Trolls are Organized and Everyone's a Target
The Effects of Online Hate and Harassment

www.adl.org/trollsharassment

adl.org

Anti-Defamation League

@ADL

@adl_national

ADL® | CENTER FOR TECHNOLOGY & SOCIETY