Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users

August 23, 2022

Abstract

To what extent does the YouTube recommendation algorithm push users into echo chambers, ideologically biased content, or rabbit holes? Despite growing popular concern, recent work suggests that the recommendation algorithm is not pushing users into these echo chambers. However, existing research relies heavily on the use of anonymous data collection that does not account for the personalized nature of the recommendation algorithm. We asked a sample of real users to install a browser extension that downloaded the list of videos they were recommended. We instructed these users to start on an assigned video and then click through 20 sets of recommendations, capturing what they were being shown in real time as they used the platform logged into their real accounts. Using a novel method to estimate the ideology of a YouTube video, we demonstrate that the YouTube recommendation algorithm does, in fact, push real users into mild ideological echo chambers where, by the end of the data collection task, liberals and conservatives received different distributions of recommendations from each other, though this difference is small. While we find evidence that this difference increases the longer the user followed the recommendation algorithm, we do not find evidence that many go down 'rabbit holes' that lead them to ideologically extreme content. Finally, we find that YouTube pushes all users, regardless of ideology, towards moderately conservative and an increasingly narrow range of ideological content the longer they follow YouTube's recommendations. Our analysis provides an ecologically valid estimate of the degree to which the YouTube recommendation algorithm influences the ideological content that users consume.

Keywords: YouTube, Recommendation Algorithm, Echo Chambers, Theory Testing

Introduction

By many measures, mass polarization is on the rise in the United States (Finkel et al. 2020). Americans are more willing to condone violence (Kalmoe and Mason 2022), less open to social relationship that cut across party or ideological lines (Fiorina and Abrams 2008, Hetherington 2009, Lelkes 2016), and more prone to partisan motivated reasoning across a number of dimensions (Bolsen, Druckman and Cook 2014, Bisgaard 2015, Khanna and Sood 2018). In the first two years of the 2020s alone, the United States witnessed partisanship undermine efforts to combat a public health crisis and threaten the peaceful transition of power that has characterized the country since its birth.

While there are many explanations for the growth of mass polarization in recent years, a prominent concern emphasizes the effects of a rapidly evolving information environment in which ideological outlets have proliferated (Nicas 2018, Schroeder 2019). The conceptual concern is that, by supplying the public with a menu of ideologically narrow outlets, individuals can exist in ideological "echo chambers" in which they rarely are confronted with alternative perspectives. Empirical evidence of user preference for homophilous networks of such echo chambers is plentiful (Bakshy, Messing and Adamic 2015, Guess 2021).

Less well-understood is the degree to which the hubs of online communities – online social networks such as Facebook, Twitter, YouTube, and Reddit – are to blame for the segregation of the public into ideological echo chambers. On the one hand, most of the empirical evidence of echo chambers finds that they are primarily a reflection user behavior (Ribeiro et al. 2020, Bakshy, Messing and Adamic 2015, Chen et al. 2021). On the other hand, mainstream media argues that these platforms – and specifically the recommendation algorithms that use artificial intelligence to suggest content to users – are instrumental in pushing people into echo chambers (Nicas 2018, Weill 2018, Roose 2019, Schroeder 2019) and down ideologically extreme rabbit holes (Tufekci 2018).

Part of the challenge in reconciling this debate stems from data limitations. Existing academic research that finds no evidence of a recommendation algorithm effect typically relies on either user watch histories or some type of anonymized data scraping method, both of which make a careful analysis of platform-specific effects hard to measure. User watch histories cannot untangle platform-specific features like recommendation algorithms from user behavior, since all that is recorded is the final user decision which is endogenous to both individual behavior and platform features (Hosseinmardi et al. 2020, Chen et al. 2021). Datasets assembled via anonymous scraping methods – i.e., relying on APIs or using "headless" browsers to scrape platforms – disconnect the sophisticated recommendation algorithms from the information on which they rely to operate, and are therefore of questionable construct validity (Ledwich 2020, Ribeiro et al. 2020).

Understanding the degree to which platforms versus individuals are responsible for echo chambers is of both practical and theoretical importance. From a practical perspective, determining whether the prevalence of online echo chambers is primarily the result of individual behaviors versus platform-level features is an essential first step toward reducing their prevalence. From a theoretical perspective, understanding how utility-maximizing individuals interact with a profit-maximizing institution (i.e., the social media platform) provides a useful framework to understand echo chambers as a scientific phenomenon of interest.

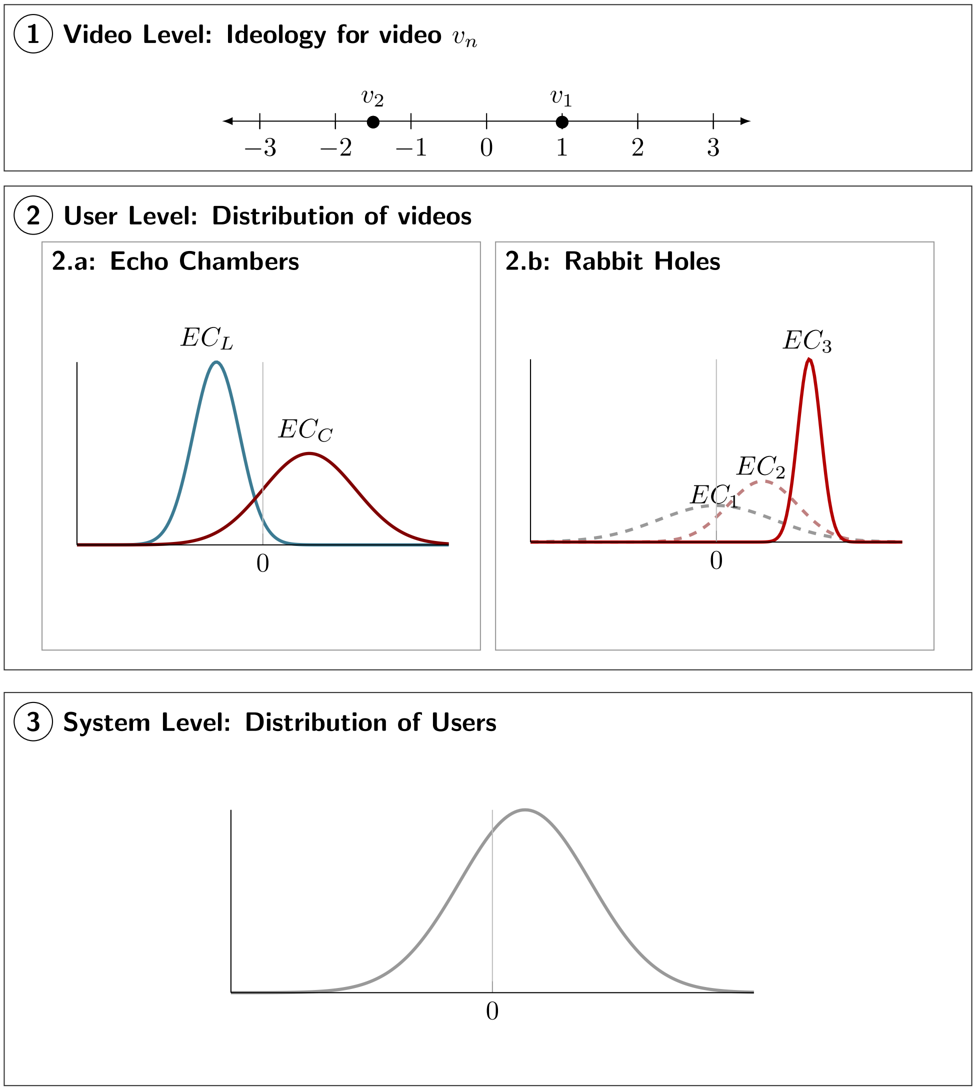
In this paper, we define a set of three theoretically important concepts that are at the core of the debate: ideological echo chambers, extremist rabbit holes, and platform-wide ideological bias. We extend well-known models of utility-maximizing behavior to define each of these concepts, and link these formal definitions with their observable implications. We then take these concepts to the data, fielding a survey of U.S.-based YouTube users in the fall of 2020 in which we experimentally manipulate aspects of their behavior to overcome the limitations described above with existing empirical work. These data provide us with ecologically valid measures of how YouTube's recommendation algorithm suggests real content to real users, while holding constant the behaviors of the users that conflate platform-specific effects with individual behaviors. We find only limited evidence of YouTube's recommendation algorithm pushing users into ideological echo chambers or extremist rabbit holes in the fall of 2020. We find stronger evidence of a platform-wide bias toward more conservative content, although this algorithmic nudge is toward a moderately conservative space, not the extremes that are the concern of journalistic investigations.

Our paper makes three contributions to the literature. First, it defines three distinct concepts of online information environments and links them with familiar spatial models of utility-maximizing individuals operating within profit-maximizing institutions. Second, it gathers and analyzes a novel dataset that overcomes the limitations associated with existing research to reconcile the debate over the role of platform-specific features in promoting echo chambers online. Third, it addresses public concerns with recommendation algorithms, finding little evidence to support the claims made in the popular press that the YouTube recommendation system systematically leads the average user to extremist content. To be very clear, this is not to suggest that YouTube is not a repository of extremist content that interested users can find through search functions, but rather that solely focusing on the recommendation algorithm may be missing the primary avenues by which individuals encounter extreme content on YouTube.

1 Echo Chambers, Rabbit Holes, and Platform-wide Ideological Biases

The broad conceptual concern with a fractured information environment can be divided into three dimensions: ideological echo chambers, extremist rabbit holes, and platform-wide ideological bias. Each of these concepts has been raised as a concern in the popular press when discussing websites such as Facebook, Twitter, and YouTube, although most accounts fail to define or differentiate the terms. Often "echo chambers" and "rabbit holes" are used interchangeably, and platform-wide biases are realized by recommendation algorithms that push users into these spaces.

To provide a road map of what follows, we structure our definitions hierarchically as depicted in Figure 1, starting by defining ideology as a continuous single dimension, in line with a rich political science literature (Poole and Rosenthal 1985, Barberá 2015, Eady et al. 2019). Each piece of content (i.e., a video on YouTube) has its own ideology, which can be placed on this single dimensional left-right spectrum. An individual user is exposed to multiple pieces of content at a given moment in time, producing a individual-specific distribution that may be more or less of an *echo chamber* if it is tightly centered around an individual's particular ideological position. Alternatively, over the course of spending time on a given platform, a user may may be pushed towards more and more extreme content, essentially falling into an extremist *rabbit hole*. Finally, *platform-wide ideological biases* occur when, at a system level, users are pushed towards videos that are systematically in one ideological direction (i.e., there could be a left-wing ideological bias or a right-wing ideological bias on a platform). In the following three subsections, we define each of these concepts more precisely.



1 Video Level: Ideology for video v_n

2 User Level: Distribution of videos

2.a: Echo Chambers

2.b: Rabbit Holes

3 System Level: Distribution of Users

Figure 1: Hierarchy of concepts: By arraying videos on a left-right spectrum of ideology (panel 1), we can characterize concepts as distributions of videos. Echo chambers are user-specific at a single point in time, and can be homogeneous and liberal (blue distribution, panel 2.a) or more ideologically diverse and conservative (red distribution, panel 2.a). Rabbit holes are dynamic sequences of echo chambers where individuals start on a diverse moderate set of videos (dashed gray line, panel 2.b), and gradually move toward a narrow and ideologically biased set of videos (solid red line, panel 2.b). System-wide biases aggregate over all users in either a liberal or (as shown) conservative direction (panel 3).

1.1 Echo Chambers

For the purposes of this paper, we define an "ideological echo chamber" as a distribution of videos for a given user that is ideologically homogeneous and centered on the individual's own ideology. Empirically, these dimensions are calculated as the average ideology of the set of videos a users is exposed to at a given moment in time (capturing bias) and the variance of these same videos (capturing homogeneity). This concept is defined at the level of an individual user, and captures the experience of Republicans only watching Fox News, or liberals only reading Democracy Now!. These users are in an "echo chamber" because they are only exposed to information and perspectives that are consistent with their ideology: their information environment echos their worldview back to them.

Despite being defined at the level of an individual, ideological echo chambers carry concerning implications for both the individual and for the society. Crucially, echo chambers are only of interest if there is more than one, allowing the public to segregate themselves into different information environments and undermine the promises of deliberative democracy.[1] Multiple echo chambers populated by distinct subsets of the population may lead to mass polarization if there is no common ground on which for the different groups to agree.[2]

[1] This is not to suggest that a single society-wide echo chamber is not of conceptual interest. For the purposes of this paper however, we are interested in the fractured information environments that connect with growing mass polarization. A single echo chamber would be unrelated to mass polarization.

[2] This notion of common ground underscores the importance of defining echo chambers as continuous along the dimensions of bias and homogeneity. The mean ideological placement (hereafter referred to as "ideological bias") combined with the homogeneity, of an ideological echo chamber, literally defines the size of the common ideological ground available to the public, represented visually as the overlap between the blue and red distributions in the echo chambers . Homogeneous but only slightly biased echo chambers lead to a public who hear limited cross cutting information, but are not too dissimilar from each other on average. Diverse but extremely biased echo chambers lead to a public who are exposed to more cross cutting information but whose views are more distant from each other on average. Extreme bias combined with high homogeneity leaves little common ground whatsoever.

1.2 Rabbit Holes

While an echo chamber is a static concept, a "rabbit hole" is dynamic and captures the process by which a user starts in a rich information environment and winds up in an ideologically extreme echo chamber. For example, a "conservative extremist rabbit hole" is a specific type of process wherein a user starts on content about Donald Trump, and ends on content produced by Holocaust deniers and white supremacists (Tufekci 2018). These extremist rabbit holes compound the normative concerns of ideological echo chambers, creating a public who not only hears different information, but hears only the most extreme versions of this information. Conceptually, we define an extremist rabbit hole as a sequence of ideological echo chambers whose bias becomes more extreme, and whose homogeneity increases, at each step in a traversal of YouTube recommended videos.

1.3 System-Wide Ideological Bias

Finally, we define a system-wide ideological bias (or "system-wide biases" for short) as the ideological bias in the recommendations of the majority of users. More specifically, ideological bias is when users are pushed towards content that is ideologically biased relative to ideologically moderate content. Here we treat bias as relative to the center of the ideology scale we have employed to measure videos. Substantively, our scale is centered around r/neutral_news, a subreddit dedicated to neutral conversations around news and current events. On YouTube, this is approximately equivalent to C-SPAN. Empirically, this aggregates over users to calculate the average ideological content that is recommended to users on the platform. A system-wide bias can coexist with echo chambers and rabbit holes centered on different average ideologies. For example, while Twitter may be a relatively liberal platform, there exist many conservative echo chambers among its users (Barberá et al. 2015). Conversely, while the alarm has been raised that YouTube has a system-wide ideological biases, there may still be users who only experience liberal (or conservative) echo chambers while on the platform (Barrett and Sims 2021).

Using these definitions, we assess the possibility that YouTube's recommendation engine contributes to echo chambers, rabbit holes, and ideological bias amongst our study participants.

Before doing so, we first describe our research design and methodology for estimating the ideology of a YouTube video.

2 Data and Methods

We are fundamentally interested in testing each of the three possibilities from our theoretical framework regarding the possible impacts of YouTube's recommendation algorithm: that it leads to ideological echo chambers, that it leads to extremist rabbit holes, and that it produces a system-wide conservative bias. To assess these possibilities, we fielded a novel survey of YouTube users who navigated the platform in the fall of 2020 according to a set of randomly assigned rules, and allowed us to record the recommendations they were shown while doing so. We then estimated the ideology of each of these recommendations, providing us with the empirical distribution of the ideological content recommended to each user at each moment in time. We summarize the method for estimating a YouTube video's ideology first, before turning to a description of the survey task and how we translated ideology scores for several hundred thousand videos into measures that capture our three quantities of interest: ideological echo chambers, extremist rabbit holes, and system-wide bias.

2.1 Ideology Estimation

To determine whether recommendations systematically lead users into ideological echo chambers, we estimate the ideology of a YouTube video using a procedure described in detail in Lai et al. (2022). This approach builds on solutions for estimating ideology in other contexts such as Twitter, the Supreme Court, and Congress by exploiting observed behaviors to estimate a unidimensional measure of ideology as a latent trait (Poole and Rosenthal 1985, Barberá 2015, Eady et al. 2019). Specifically, we use the observed behavior of sharing YouTube videos in the domain of ideological subreddits to calculate each video's ideology that appears on Reddit. This set of more than 50,000 videos with an ideology score are then used as training data for a natural language classifier, which is then used to predict the ideology of any video on YouTube. Below, we summarize the broad contours of the approach and direct the interested reader to Lai et al. (2022) for a more detailed description.

Our methodological approach takes advantage of the availability of Reddit data that is already grouped by ideology. More specifically, the method utilizes Reddit data from 1,230 politically-oriented subreddits, meaning Reddit communities oriented around a particular topic, interest group, or political orientation (e.g. r/Conservative and r/liberal, subreddits dedicated to discussing conservative positions on political topics and liberal positions on political topics, respectively). We collect all submissions, a type of post on the platform that contains a link, from each of the subreddits from December 31, 2011 through June 21, 2021. The core assumption of this method is that Reddit users post YouTube videos in subreddits with which the videos are ideologically aligned. For example, a hypothetical Fox News video would be more likely to show up in a subreddit like r/Conservative than a subreddit like r/liberal. In addition, Reddit allows users to "up-vote" or "down-vote" pieces of content, where up-votes can be considered endorsements of the content and down-votes are the opposite. The score of a post takes these up-votes and down-votes into account and therefore captures the extent to which a given post aligns with that subreddit.

These posts are then filtered to isolate those with a link to a YouTube video, resulting in 31,113,005 posts across the 1,230 subreddits. Next, posts are again filtered for posts with a positive score that appear in at least one of the 1,230 subreddits, resulting in 1,268,207 posts with links to a YouTube video. The remaining posts are then iteratively filtered for basic popularity metrics: subreddits are kept where at least five unique videos have been posted and videos are retained only if they have been posted in a minimum of three subreddits. This procedure results in 362,360 posts containing links to YouTube videos, with 62,558 unique videos posted across 886 subreddits. Finally, a subreddit-video matrix is created with videos as rows and subreddits as columns. If video v receives a score of x in subreddit s, then the corresponding matrix entry for (v, s) is $\ln(x + 1)$ — we take the natural log plus one due to the wide range of scores. If a video does not appear in a subreddit, the corresponding matrix entry is 0. To complete the first stage of video classification, a correspondence analysis-based model is then fit on the correspondence matrix of videos and subreddits in three dimensions. The first dimension of the correspondence analysis is then selected to represent the ideology of the YouTube video.[3]

Finally, to expand the videos for which one can generate ideology scores beyond those that appear on Reddit, labels are propagated using the correspondence analysis model to label videos that did not appear in the subreddits using a finetuned text-based model (Devlin et al. 2019). For each of the videos for which an ideology score could be estimated using the previously described method, video metadata was collected from the YouTube Data API, which contains the video description, title, tags, and channel title. This model is then used to predict the ideology of the videos in the recommendations dataset described in the following section.[4] A schematic of this method can be found in Figure 2.

2.2 Survey Task

Isolating the recommendation algorithm is a challenging task. On one hand, using data from the YouTube Data API or from web scraping presents a low-cost method for collecting these data at scale. However, this methodology removes a core part of the algorithm: personalization.

On the other hand, we could rely on user watch histories alone – the videos they watched in the order they watched them – but this potentially confounds the behavior of the recommendation algorithm with user preferences for content.[5] Specifically, because we can only observe what

[3] These ideology scores were validated by human coders who were asked to compare two videos and determine which was more liberal / conservative; see (Lai et al. 2022) for more detail.

[4] As described in (Lai et al. 2022), there were 52,463 videos for which metadata could be recovered metadata at the time of analysis (the rest were removed, made private, or were otherwise publicly unavailable at the time of analysis). These videos were then used as training data for a BERT (Bi-directional Encoder Representations from Transformers) model—a pre-trained transformer-based model for language understanding—with a regression head (Devlin et al. 2019). The input features are the concatenation of the available text features from the YouTube video metadata, and the target outputs are the ideology scores derived from the unsupervised network model. On the test set, the text-based predictions and ground-truth correspondence analysis scores have a correlation coefficient of 0.891, with $R^2 \approx 0.794$. The mean squared error is 0.171: roughly five percent of 3.329, the range of the ground truth scores, and roughly 19% of 0.907, the standard deviation of the same. Further details on human validation and model performance can be found in the footnote 6 and in (Lai et al. 2022).

[5] Watch histories can be collected for analysis from consenting users who are willing to install browser tracking programs or submit their watch histories from the YouTube “Download Your Data” feature.

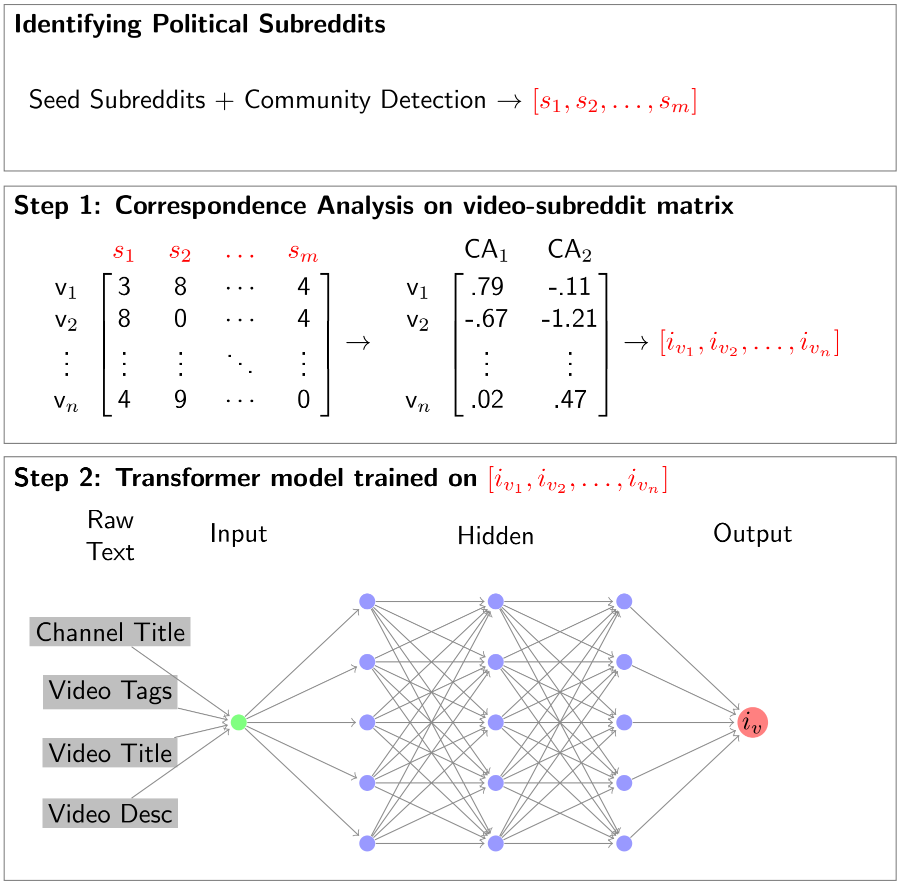

Identifying Political Subreddits
Seed Subreddits + Community Detection → $[s_1, s_2, \ldots, s_m]$
Step 1: Correspondence Analysis on video-subreddit matrix
s_1 s_2 … s_m
v$_1$ 3 8 ⋯ 4
v$_2$ 8 0 ⋯ 4
v$_n$ 4 9 ⋯ 0
→
CA$_1$ CA$_2$
v$_1$ .79 -.11
v$_2$ -.67 -1.21
v$_n$ .02 .47
→ $[i_{v_1}, i_{v_2}, \ldots, i_{v_n}]$
Step 2: Transformer model trained on $[i_{v_1}, i_{v_2}, \ldots, i_{v_n}]$
Raw Text
Input
Hidden
Output
Channel Title
Video Tags
Video Title
Video Desc
i_v

Figure 2: A schematic of the overall method for ideology estimation from Lai et al. (2022).

was actually *watched* and not the full list of what was *recommended*, we cannot be sure that any biases we document are due to the recommendation algorithm or to an individual user's decision to click on a given video. To address these concerns, we fielded a survey in late 2020 where we strictly controlled the behavior of users to collect an ecologically valid snapshot of the recommendation algorithm. We pair these data with the ideology estimation previously described to evaluate the prevalence of ideological echo chambers, rabbit holes, and ideological bias *resulting from the YouTube recommendation algorithm* amongst US users.

More specifically, from October 2, 2020 to December 7, 2020, we recruited a convenience sample of 1,063 YouTube users using Facebook ads.[6] We asked respondents to install a web browsing plug-in to record their YouTube recommendations for the duration of the task. Additionally,

[6] Our sample was recruited using Facebook ads targeting American residents aged 18 years and older. A more detailed description of the recruiting strategy and demographics is included in the Supporting Information.

they answered a brief survey after the fact regarding their demographics and usage of YouTube. Participants were compensated $5 for the task and survey. The complete survey is available in the appendix.

Each study participant was asked to complete a “traversal task”. For this task, we randomly assigned each participant a starting video from one of 25 potential starting videos (consisting of a mixture of political content across the ideological spectrum and some non-political content from music, gaming, and sports).[7] The user navigated to the video and then was randomly assigned to one of five “traversal rules”. Participants were instructed to always click on a pre-determined recommendation; that is, always click the first video, the second video, the third video, the fourth video, or the fifth video. Respondents followed their assigned rule for a total of twenty traversals, during which the browser extension passively collected the list of recommended videos presented at each traversal step (typically approximately 20 videos were collected at each step).

Once the survey was complete, we used the procedure described above to generate an estimate of the ideology of every video shown to our respondents, mapped onto a common unidimensional space Lai et al. (2022). We visualize an example of the traversal results for a given respondent in Figure 3, arraying the recommendations shown at each traversal step (x-axis) by predicted ideology (y-axis). This respondent started the task on the randomly assigned seed video j, which we outline with a thick black border and position according to its predicted ideology of approximately -1 on the y-axis at traversal step 0 on the x-axis. As they watched this video, they were recommended approximately 20 videos, which we depict as rectangles of varying size at traversal step 1. Videos that appear higher in the recommendation list receive a larger rectangle, while videos lower in the list receive smaller rectangles. This particular respondent was randomly assigned to always click on the third video in the list of recommendations, which we highlight with a black border and line linking the current video with the subsequent. We construct a respondent-by-recommendation dataset where for a given user, for whom we know demographics and general YouTube habits, we

[7] That is, we randomized which of the 25 potential seed videos each participant was assigned, but the list of potential seed videos was not itself randomly selected. The list of videos was selected to include fifteen political videos across the ideological spectrum and nine videos from nonpolitical categories. For a list of the seed videos, see the SI.

have a 20x20 set of ecologically valid recommendations like the one outlined in Figure 3.

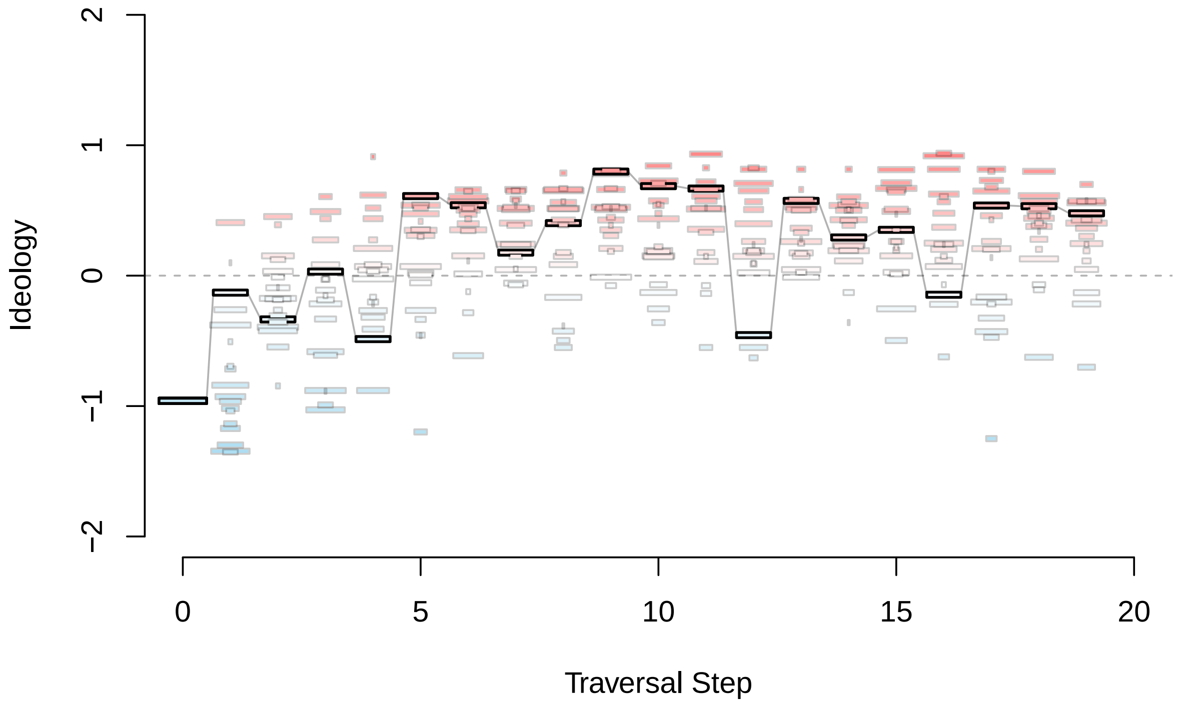


Figure 3: Example of an empirical traversal: On the x-axis we show the traversal step, and on the y-axis the estimated ideology of the video. Positive values indicate that the video is more conservative while negative values indicate that the video is more liberal. Videos outlined in black are those that the respondent clicked on, linking each set of recommendations across traversal steps. The respondent starts on a center-left video and randomly selects the next video. We show the distribution of ideology of the recommendations where each recommendation is sized by its rank in the list of recommendations. Videos that appear higher in the recommendations are sized larger.

2.3 Evaluating Recommendations

Recall from above that we are interested in three quantities of interest: ideological echo chambers, extremist rabbit holes, and system-wide bias. To convert our rich respondent-by-recommendation data into a format that will allow us to empirical measure these three concepts, we can use the empirical traversal in Figure 3 as a motivating example, which starts on a liberal seed video. We can see that the recommendations to this respondent are widely distributed across the ideological spectrum, starting in a more liberal position for the first few traversal steps before shifting toward a reasonably diverse set of recommendations centered around moderate content. Substantively, this particular user's experience is not consistent with ideological echo chambers at any given step with the exception of the first set of recommendations (step 1), nor is there evidence of the respondent being pushed down an extremist rabbit hole. If anything, these data are potentially reassuring evidence of an anti echo-chamber nudge on the part of the recommendation algorithm.

Conversely, in Figure 4 we show an experience from a different respondent. This respondent starts on a moderate video, which has a wide distribution of recommendations. After the second step, the respondent's recommendations become very conservative and very narrow. They remain this way for the duration of the traversal.[8]

The contrast between the two example respondents highlights two of our three quantities of interest: ideological echo chambers and extremist rabbit holes. The first respondent was recommended predominantly liberal content at their first video, although these recommendations were relatively diverse, covering a range between less than -1.2 and greater than 0. Similarly, the second respondent's first set of recommendations were predominantly conservative but similarly diverse. The distributions of recommendations for both respondents at their initial step are consistent with mild ideological echo chambers: the average ideology was biased toward the respondent's views, but the variance was relatively large indicating a diversity of recommendations.

However, the ensuing traversal steps reveal a divergence in the recommendations shown to both respondents at ensuing steps. For the first respondent, each subsequent video clicked was associated with a distribution of recommendations that was equally or more diverse, and with an average that trended toward zero – distributions incompatible with our definition of an ideological echo chamber, and a trajectory inconsistent with our definition of an extremist rabbit hole. Conversely, the second respondent spends most of their time in ideological echo chambers,

[8] This particular respondent was a conservative Republican white woman whose recommendations largely consist of Fox News, press briefings from the Trump White House, and a smattering of conservative pundits. The recurring recommendation that scores left of center is also from the White House YouTube channel, and is coverage of Ivanka Trump's pledge to American workers. This outlier underscores the value of our method's ability to generate video-level ideology scores.

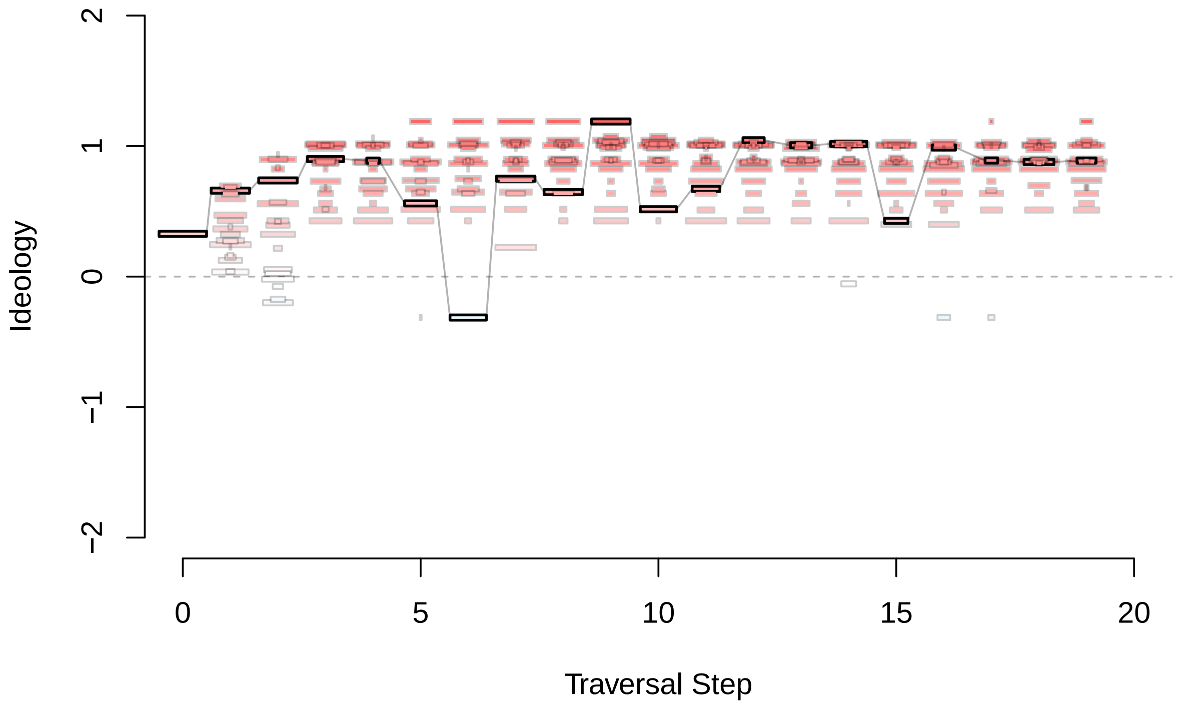


Figure 4: Example of an empirical traversal: The respondent starts on a center right video and randomly selects the next video. We show the distribution of ideology of the recommendations where each recommendation is sized by its rank in the list of recommendations. Videos that appear higher in the recommendations are sized larger.

characterized by strongly conservative content on average (mean ideology) combined with a very narrow range of recommendations to choose from (variance) at each traversal step after the second. Furthermore, this transition from content that was moderate and diverse to content that was extreme and homogeneous happened very quickly, consistent with an extremist rabbit hole.

Based on this description, we operationalize a two dimensional measure of an echo chamber based on the mean and the variance of the recommendations shown at each step, which we then use as our main dependent variables in the regression analysis. Formally, let $y_{u,j,k,t}$ denote either the mean ideology of a set of recommendations or the variance of a set of recommendations shown to respondent u at traversal step t who was randomly assigned to seed video j and traversal rule k. (For simplicity, we drop the j and k subscripts which are associated with respondent u.) Mean ideology is calculated as weighted average of each recommended video v_i, weighted by the inverse of its recommendation rank i, or: $y_{u,t} = \sum_i v_{i,u,t} * \frac{1}{i}$. We choose a weighted measure of mean ideology to reflect the fact that videos which appear more toward the top of the recommendation list are more visible to the viewer. Similarly, we calculate the weighted variance as $y_{u,t} = \frac{\sum_i \frac{1}{i} * (v_{i,u,t} - \bar{v}_{u,t})^2}{|v_{u,t}|-1}$ where $|v_{u,t}|$ indicates the cardinality of $v_{u,t}$, or the number of recommendations shown at a given step.

Table 1 summarizes these measures for clarity:

Concept	**Measure**	**Formalization**	**Interpretation**
Ideological Bias	Average ideology	$\bar{y}_{u,t} = \sum_i v_{i,u,t} * \frac{1}{i}$	$\lvert\bar{y}\rvert > 0$: more biased
Ideological Diversity	Variance	$\mathrm{Var}(y_{u,t}) = \frac{\sum_i \frac{1}{i} * (v_{i,u,t} - \bar{v}_{u,t})^2}{\lvert v_{u,t}\rvert - 1}$	$\mathrm{Var}(y) > 0$: more diverse
Echo Chamber	Bias & Diversity		$\lvert\bar{y}\rvert \uparrow$ & $\mathrm{Var}(y) \downarrow$

Table 1: Mapping between concepts of interest and empirical measures. Each measure is calculated at the user-traversal step unit, aggregating over all recommendations suggested by the algorithm at each step.

Our main approach to addressing our three research questions relies on a combination of descriptive statistics and linear regression analysis. One implication of echo chambers is that conservatives would be shown more conservative recommendations than liberals on average. To investigate this implication, we predict the ideology of recommendations as a function of the user u's self-reported ideology (ideo_u) on a scale ranging from 1 (most liberal) to 7 (most conservative).[9] Formally,

$$y_{u,j,k} = \alpha_j + \gamma_k + \beta_1 \text{ideo}_u + \varepsilon_{u,j,k} \tag{1}$$

where α_j represent fixed effects for the seed video and γ_k are fixed effects for the traversal rule. If conservatives are shown more conservative content than liberals, we would expect the β_1 coefficient to be significant and positive.

To investigate the second research question pertaining to the existence of rabbit holes, we remind the reader that rabbit holes refers to the process by which users arrive at echo chambers,

[9] As a robustness check, we treat user ideology as a categorical variable instead of the continuous self-placement on a 1 to 7 scale. We do so by assigning users to be either liberals (self-placement less than 4), conservatives (self-placement greater than 4), or true moderates (self-placement in the middle of the scale). $y_{u,j,k} = \alpha_j + \gamma_k + \beta_1 \text{Mod}_u + \beta_2 \text{Cons}_u \varepsilon_{u,j,k}$.

requiring us to incorporate the amount of time a user spends following the recommendations into our analysis. Specifically, we predict the ideology of recommended videos as a function of both user ideology and the amount of time they have spent following our randomly assigned traversal rule. Formally:

$$y_{u,j,k,t} = \alpha_j + \gamma_k + \beta_1 \mathrm{t}_{u,j,k} + \beta_2 \mathrm{Mod}_u + \beta_3 \mathrm{Cons}_u + \beta_4 \mathrm{t}_{u,j,k} * \mathrm{Mod}_u + \beta_5 \mathrm{t}_{u,j,k} * \mathrm{Cons}_u + \varepsilon_{u,j,k,t} \quad (2)$$

where $\mathrm{t}_{u,j,k}$ is the traversal step for user u who started on seed video j and followed traversal rule k. This specification allows us to test not only if users who spend more time following YouTube's recommendations are pushed further apart from each other, but also whether this divergence is due to liberals being recommended more liberal content while conservatives are recommended more conservative content.[10]

Recall from Figure 1, panel 2.b, that rabbit holes are defined along two dimensions: average ideology and the variance (or "diversity") of the recommended videos. To test this second dimension, we re-run Equation 2, replacing the average ideology of the 20 recommendations suggested at each step with the variance of the 20 recommendations. Here, we are principally interested in β_1, which indicates whether recommendations get less diverse the more time the user spends following the algorithm (i.e., $\beta_1 < 0$). However, we also examine the interaction terms β_4 and β_5 to test if the strength of the rabbit hole push is larger for one ideological group than another, although we have no theoretical reason to suspect so.

The specification represented by Equation 2 also speaks to the third research question about system-wide ideological bias. Namely, β_1 captures the overall average push of the algorithm after controlling for user ideology, as well as seed video and traversal rule random assignment.

[10] To examine the robustness of this result, we also re-estimate subsetting our data to each ideological group of users and predicting average recommendation ideology as a function of a cubic polynomial measure of the traversal step: $y_{u,j,k} = \alpha_j + \gamma_k + \beta_1 \mathrm{t}_{u,j,k} + \beta_2 \mathrm{t}^2_{u,j,k} + \beta_3 \mathrm{t}^3_{u,j,k} + \varepsilon_{u,j,k} \; \forall u \in \{\text{lib,mod,cons}\}$. These results are included in the Supporting Information.

3 Results

In the following section, we assess the prevalence of echo chambers, rabbit holes, and system-wide ideological bias on YouTube.

3.1 RQ1: Echo Chambers

We start by plotting the average ideology of recommendations shown to users, broken out by self-reported ideology, in Figure 5. Based on this simple descriptive, there is very little evidence of ideological echo chambers in the recommendations shown to people who were randomly assigned to a seed video and traversal rule. While there is evidence that more conservative users are recommended more conservative content on average, the difference between the most conservative and most liberal users is very small, amounting to roughly 0.1 units on an ideological scale ranging between -1.5 and +1.5. Furthermore, the distributions are wide, indicating that all users see a lot of overlapping content – at least in terms of how recommendations are mapped onto a uni-dimensional ideological space.

From Figure 5, there is little evidence to suggest that the YouTube recommendation algorithm puts real users into ideological echo chambers. To formally test this proposition, we run the regression specified in Equation 1 and summarize the findings in Table 2. Here we do find a mildly significant positive association between the ideology of our users and the ideology of what they are recommended. As an additional robustness check, we drop respondents who self-report paradoxical ideology-partisanship pairings (i.e., extremely conservative Democrats and extremely liberal Republicans). These checks find even stronger evidence of a significant difference in the recommendations suggested to liberal and conservative users. However, even the strongest findings suggest no more than a 0.1 unit difference between the most liberal and most conservative respondents. Put simply, the ideological difference between content recommended to the most liberal and the most conservative users is statistically significant, but small.[11]

[11] We run also this regression instead with the party ID of the respondent and find no significant difference between Republicans and Democrats. These findings are summarized in the SI.

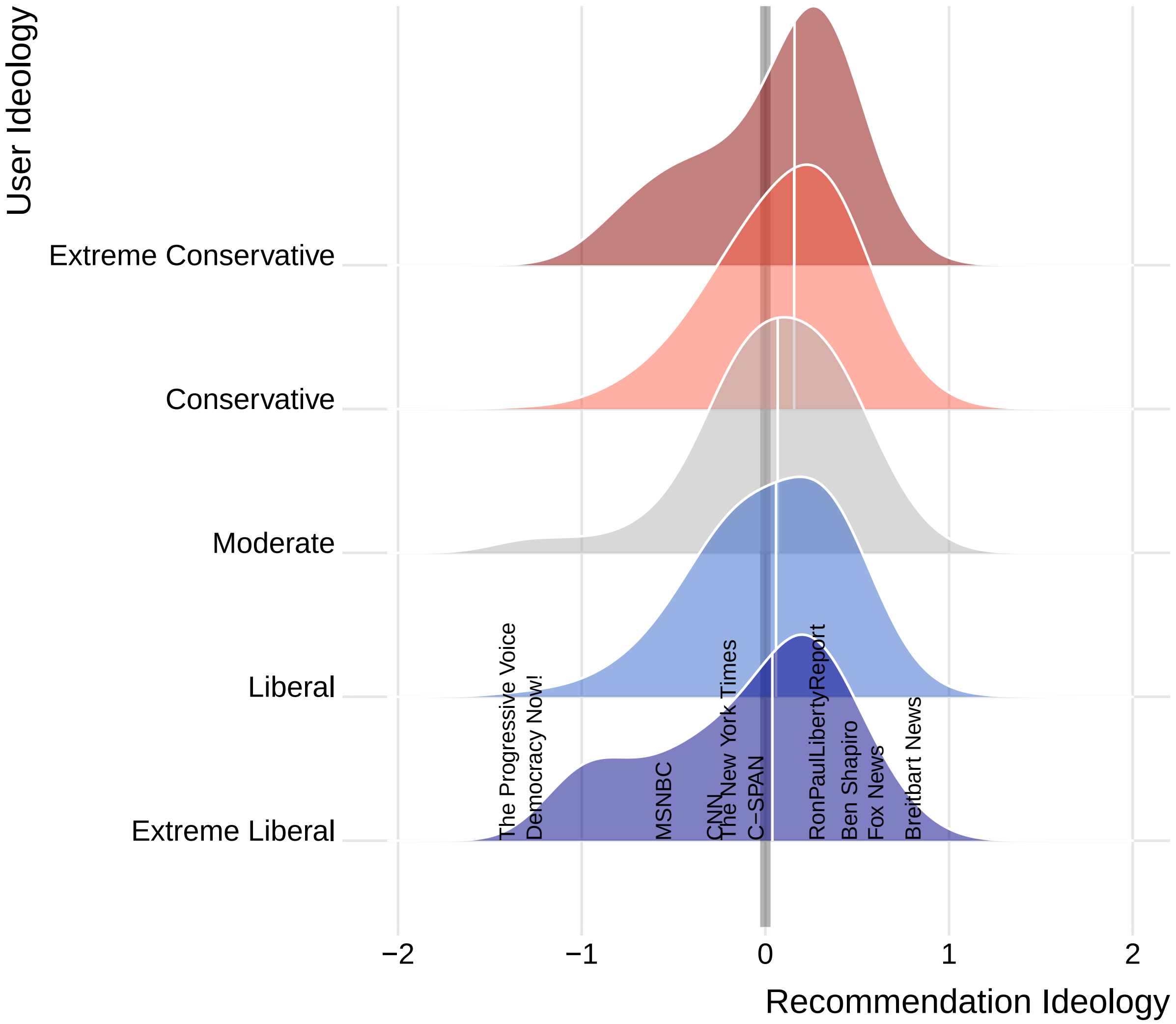


Figure 5: Distribution of ideology of all recommendations (x-axis) shown to users by self-reported ideology (y-axis). Average ideology of specific YouTube channels displayed along bottom for reference.

3.2 RQ2: Rabbit Holes

To investigate our second research question, we begin with descriptive evidence, plotting the average ideology of recommendations shown to users as they proceed further into the traversal task. Figure 6 plots these distributions, aggregating to liberals (blue), moderates (grey), and conservatives (red), and over every five traversal steps (y-axis). As illustrated, all groups move slightly to the right of center the further they follow the recommendation algorithm. Again, there is little descriptive evidence of a substantial narrowing of the diversity of the content recommended.

Table 2: Average ideology of recommendations

	Average Recommendation Ideology			
User Ideology	(1)	(2)	(3)	(4)
Continuous Ideology	0.0148*	0.0234**		
	(0.0078)	(0.0092)		
Moderate			0.0647**	0.0633**
			(0.0290)	(0.0297)
Conservative			0.0530**	0.0737***
			(0.0228)	(0.0238)
Fixed-effects				
Seed Video	Yes	Yes	Yes	Yes
Traversal Rule	Yes	Yes	Yes	Yes
Sample				
Drop incongruous partisans	No	Yes	No	Yes
Fit statistics				
Observations	13,646	12,384	13,646	12,384
R^2	0.12742	0.13115	0.12972	0.13283
Within R^2	0.00435	0.00899	0.00697	0.01091

Clustered (respondent_id) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

We estimate equation 2 to formally test the significance of these patterns, again finding statistically significant associations between self-reported ideology and average recommendation ideology that grow more pronounced the longer the users spend clicking on recommended videos. We report the coefficients in Table 3. To facilitate interpretation of the interaction terms, we plot these results as marginal effects in Figure 7, examining the marginal effects of the continuous version of self-reported ideology on the average ideology of recommended videos at each traversal step in the left panel and the binned version of the same in the right panel. In both cases, we find statistically significant evidence that the gap between recommendations shown to liberals and conservatives grows wider over time, with conservatives being shown significantly more conservative content than liberals after following the recommendation algorithm for about 10 steps. Again, however, we emphasize that these differences are small, amounting to a difference of at most 0.1 units on our -1.5 to +1.5 unit ideology scale. (We note that the coefficient on the uninteracted traversal step variable is significant and positive in column 1, indicating that even liberal respondents are

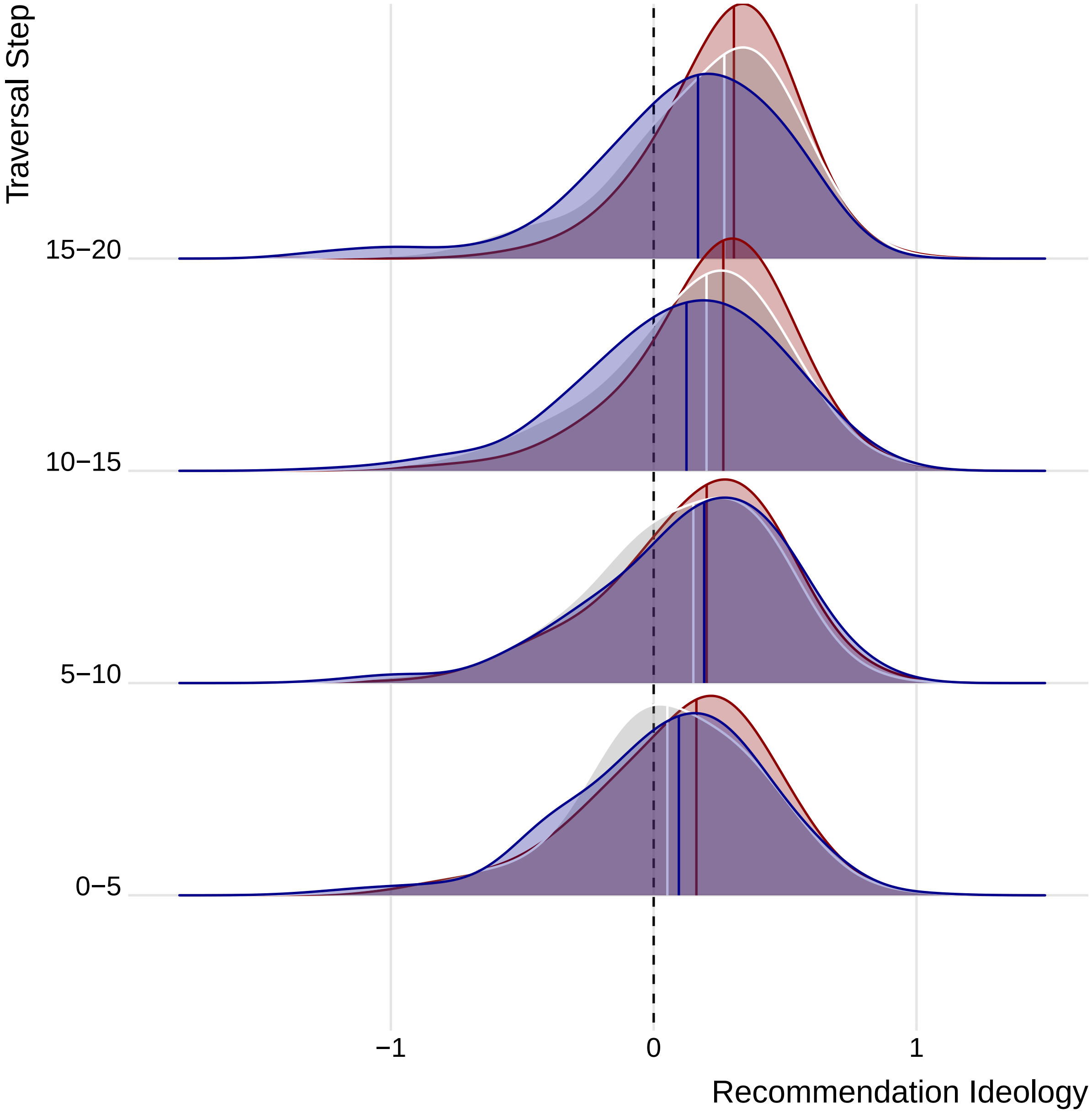


Figure 6: Distribution of ideology of all recommendations (x-axis) shown to users by self-reported ideology (liberals in blue, moderates in grey, conservatives in red), by how deep into the traversal task they are (y-axis)

recommended more conservative videos the longer they spend on the platform. We discuss this result in more detail below in reference to our third research question.)

Table 3: Average and Variance predicted by User Ideology and Traversal Step

	Average Ideology		Ideological Variance	
	(1)	(2)	(3)	(4)
Moderate (ref Lib)	0.0537		-0.0058	
	(0.0389)		(0.0179)	
Conservative (ref Lib)	-0.0041		-0.0002	
	(0.0304)		(0.0123)	
Continuous Ideo		-0.0071		0.0067*
		(0.0099)		(0.0038)
Traversal Step	0.0071***	0.0011	-0.0046***	-0.0018
	(0.0015)	(0.0026)	(0.0006)	(0.0011)
Moderate × Step	0.0014		-0.00006	
	(0.0028)		(0.0013)	
Conservative × Step	0.0055**		-0.0015*	
	(0.0022)		(0.0009)	
Continuous × Step		0.0021***		-0.0009***
		(0.0006)		(0.0003)
Fixed-effects				
root_video	Yes	Yes	Yes	Yes
travRule	Yes	Yes	Yes	Yes
Fit statistics				
Observations	13,646	13,646	10,710	10,710
R^2	0.16	0.16	0.12	0.12
Within R^2	0.04	0.04	0.05	0.05

Clustered (respondent_id) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

In terms of whether these echo chambers grow increasingly narrow as well as increasingly separated, we re-run the preceding specification replacing the average ideology of recommendations with the average variance of recommendations to a given user as the outcome variable. In table 3 we see that all coefficients for traversal step are estimated to be negative in columns 3 and 4, suggesting that ideological diversity decreases over time as users spend more time clicking on recommended content. Like the previous analysis, the magnitude of this change is very small.

As a final descriptive summary, we define a radical rabbit hole as a set of recommendations that is more ideologically extreme than either positive or negative 0.4, and is more ideologically homogeneous than a variance of 0.15 over the final set of five traversals. Substantively, these cut-offs are approximately equivalent to being to the right of Ben Shapiro's channel and to the left of

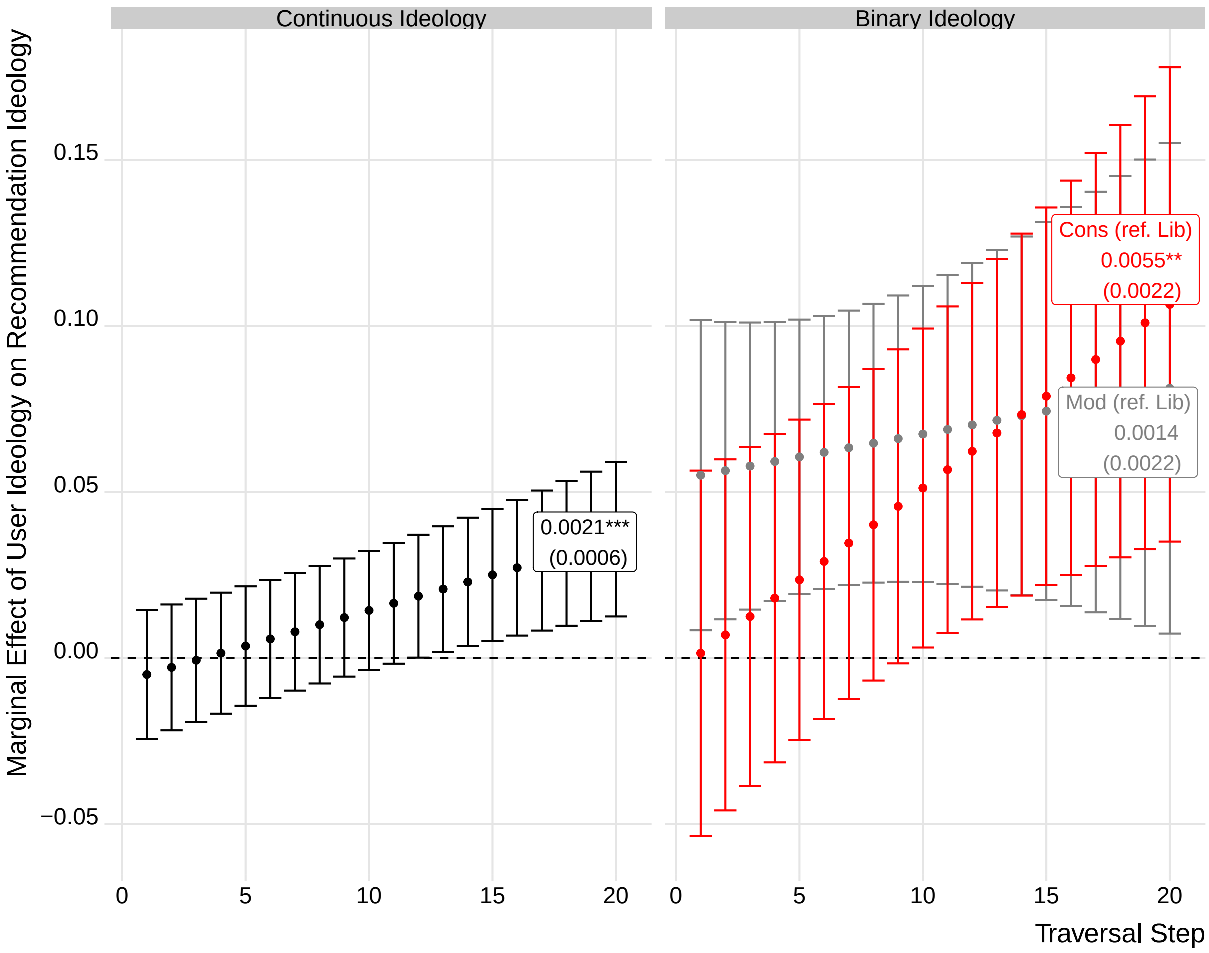


Figure 7: Marginal effects (y-axes) of user ideology on recommendation ideology, across different durations of the traversal task (x-axes). Left panel summarizes regression results using continuous measure of self-report ideology. Right panel summarizes results using binned categorical version of self-reported ideology.

MSNBC. We then count the number of users whose experience following the algorithm exceeds these thresholds, concluding that 14 out of our 527 respondents who followed the recommendations at random fell into this definition of a radical rabbit hole. Of these 14 users, 11 arrived at a conservative filter bubble and 3 arrived at a liberal filter bubble. Importantly, of those that arrived at the conservative filter bubble, only five of the fourteen were self-reported conservatives, underscoring the lack of evidence that these narrow bands of ideological recommendations reflect the concept of an echo chamber of ideologically congruent content that reinforces a user's prior beliefs. While we do not find that users fall into so-called "rabbit holes" en masse – and thus reject the hypothesis that YouTube's recommendation algorithm leads the *average* users to ideologically extreme content

– it is still important to remember that due to the fact that YouTube is so widely used, even small percentages of users falling into rabbit holes could amount to many users having this experience.

3.3 RQ3: Ideological Bias

The preceding analysis finds systematic evidence of conservatives being recommended more conservative content than liberals, and that this divergence grows as users spend more time clicking on recommended content. Similarly, there was some evidence that conservatives arrived at more homogeneous content than liberals by the end of their traversal task, but this is not robust to the choice of user ideology measure. Furthermore, the differences, even those that are robustly significant at the 95% threshold, are substantively small. But what of the third research question pertaining to the overall bias of the recommendation algorithm?

Already, we observe a systematic rightward bias away from zero in Figures 5 and 6. In addition, the coefficient estimates on the traversal step predictor in Table 3 are significant and positive in column 1, estimating the average ideology, and significant and negative in column 3, estimating the average variance, further suggesting that the overall trend is to push users into more conservative, narrower content overall. (The patterns in columns 2 and 4 indicate that this pattern exists, but that it is driven primarily by conservative users.) Figure 8 shows the averages of both ideology and variance at each traversal step, aggregating by user ideology and seed video ideology, which we also bin into liberal, moderate, and conservative seed videos. Doing so highlights the compelling evidence of a system-wide conservative nudge that moves all users, regardless of ideology or randomly assigned seed video, toward more conservative content. However, we emphasize that this nudge is small in magnitude, constituting a shift from roughly C-SPAN to roughly RonPaulLibertyReport. Importantly, these patterns obtain even among users who were randomly assigned to start on liberal videos, suggesting that the recommendation algorithm's conservative bias supercedes the influence of whichever video a user happens to be on (the context) as well as the user's ideology (the personalization). While users who started on a seed video began their traversal task on more liberal content than those who started on a conservative seed, these differences dissipated over the course of clicking on subsequent recommendations.

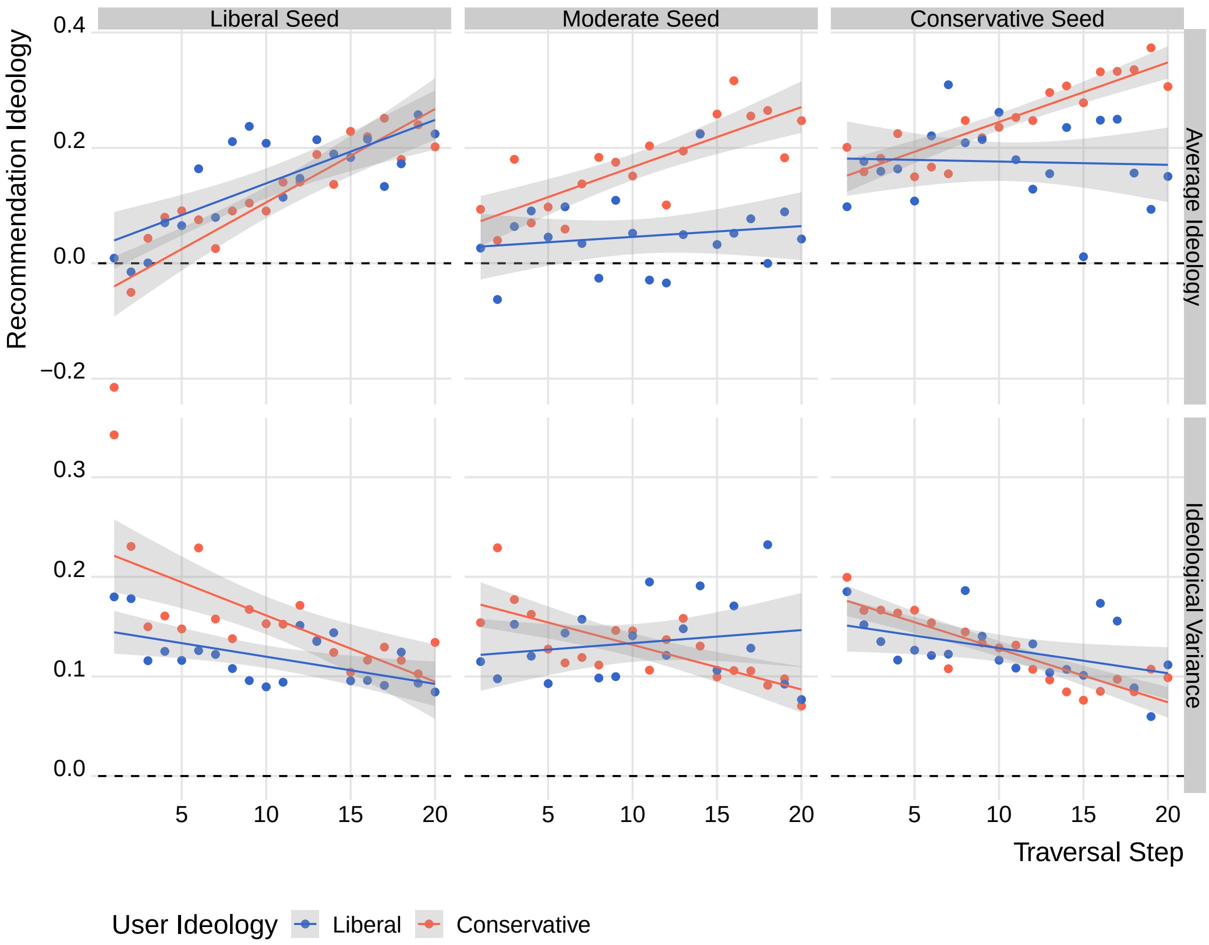


Figure 8: Top row: average ideology of recommendations shown to conservative (red) and liberal (blue) users at each traversal step (x-axes), separated out by whether the user was randomly assigned to start on a liberal seed video (left), moderate seed video (center) or conservative seed video (right). Bottom row: variance of recommendations shown to same groups, disaggregated by same seeds.

4 Discussion

These findings present a first look at the ideological distribution of recommendations to real users on YouTube. We evaluate the prevalence of echo chambers, rabbit holes, and system-wide ideological bias. Previous research has relied on automated recommendation collection strategies, which do not account for user personalization, or observational web browsing data, with which we cannot disentangle user preferences from the recommendations that YouTube supplies. By asking real users to navigate YouTube using their real accounts, we find that there is only mild evidence of echo chambers on YouTube. While conservatives see content that is more conservative than liberals, the magnitude of this difference is small, and significance is not robust to alternative specifications such as using party identification as the explanatory variable rather than ideology. Additionally, we find that this difference between liberals and conservatives increases as users follow YouTube's recommendations. After approximately ten traversal steps, conservatives see significantly more conservative content than liberals, but these differences are small, amounting to a difference of at most 0.1 units on a -1.5 to +1.5 ideology scale. We find evidence that the ideology of videos narrows over time, but this does not differ systematically by user ideology. Finally, we find that despite the mild differences between the experiences of conservatives and liberals on the platform, all users regardless of ideology receive more conservative and less ideologically diverse recommendations over time.

Our results speak to three hypotheses about YouTube's recommendation algorithm: that YouTube drives users into echo chambers, rabbit holes, or towards ideologically biased content.[12] Our empirical results are consistent with the theoretical intuition of ideological bias, described in the SI, suggesting that more conservative videos score higher on an unobserved "valence" dimension. Substantively, this might be due to conservative content producers being more adept at crafting attractive videos or video content such as titles, thumbnails, or descriptions that are not captured in the measure of video ideology we apply. Alternatively, these findings are also consistent with the idea that YouTube is choosing videos at random from a library that leans conservative (i.e., it may be that the majority of content available on YouTube leans conservative).

However, these results should be interpreted with caution, particularly with respect to the differential results for liberals and conservatives. To start, we recruit from a convenience sample online, and individuals who are willing to share their data with researchers may fundamentally differ from the general population in ways that we cannot observe. Although our data allow us to isolate the role played by the recommendation algorithm, we are unable to peer inside the black box. Without this clarity, we can't determine whether the algorithm operates more forcefully for conservatives because they are more demanding of ideologically congruent content than liberals, or for some other reason. For example, if conservatives more consistently click on conservative videos than liberals click on liberal videos, an algorithm trained to provide users with videos they would

[12] See SI for a formalization of this logic.

most likely want to watch will naturally better serve the republicans. Conversely, if conservative content is simply more abundant on the platform, the mild conservative bias across all traversals we observe might reflect the underlying distribution of the available supply of content.

Nevertheless, our findings indicate that YouTube's recommendation algorithm is not pushing large proportions of users into highly isolated information environments in which liberals and conservatives see little overlapping content. Yet we also find that this content becomes somewhat more conservative – and that the ideological diversity of these recommendations narrows – the longer users follow the recommendations suggested by the algorithm.

References

Bakshy, Eytan, Solomon Messing and Lada A Adamic. 2015. “Exposure to ideologically diverse news and opinion on Facebook.” *Science* 348(6239):1130–1132.

Barberá, Pablo. 2015. “Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data.” *Political Analysis* 23(1):76–91.

Barberá, Pablo, John T Jost, Jonathan Nagler, Joshua A Tucker and Richard Bonneau. 2015. “Tweeting from left to right: Is online political communication more than an echo chamber?” *Psychological science* 26(10):1531–1542.

Barrett, Paul M and Brandi Sims, J Grant. 2021. False Accusation: The Unfounded Claim that Social Media Companies Censor Conservatives. Technical report NYU Stern, Center for Business and Human Rights.

Bisgaard, Martin. 2015. “Bias will find a way: Economic perceptions, attributions of blame, and partisan-motivated reasoning during crisis.” *The Journal of Politics* 77(3):849–860.

Bolsen, Toby, James N Druckman and Fay Lomax Cook. 2014. “The influence of partisan motivated reasoning on public opinion.” *Political Behavior* 36(2):235–262.

Chen, Annie Y., Brendan Nyhan, Reifler Jason, Ronald E. Robertson and Wilson. Christo. 2021. Exposure to Alternative Extremist Content on YouTube. Technical report Anti-Defamation League.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2019. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.”.

Eady, Gregory, Richard Bonneau, Jonathan Nagler and Joshua Tucker. 2019. “Partisan News on Social Media: Mapping the Ideology of News Media Content, Citizens, and Politicians.” *Working Paper* .

Finkel, Eli J, Christopher A Bail, Mina Cikara, Peter H Ditto, Shanto Iyengar, Samara Klar, Lilliana Mason, Mary C McGrath, Brendan Nyhan, David G Rand et al. 2020. “Political sectarianism in America.” *Science* 370(6516):533–536.

Fiorina, Morris P and Samuel J Abrams. 2008. "Political polarization in the American public." *Annu. Rev. Polit. Sci.* 11:563–588.

Guess, Andrew M. 2021. "(Almost) Everything in Moderation: New Evidence on Americans' Online Media Diets.".

Hetherington, Marc J. 2009. "Putting polarization in perspective." *British Journal of Political Science* 39(2):413–448.

Hosseinmardi, Homa, Amir Ghasemian, Aaron Clauset, David M Rothschild, Markus Mobius and Duncan J Watts. 2020. "Evaluating the scale, growth, and origins of right-wing echo chambers on YouTube." *arXiv preprint arXiv:2011.12843* .

Kalmoe, Nathan P and Lilliana Mason. 2022. *Radical American Partisanship: Mapping Violent Hostility, Its Causes, and the Consequences for Democracy.* University of Chicago Press.

Khanna, Kabir and Gaurav Sood. 2018. "Motivated responding in studies of factual learning." *Political Behavior* 40(1):79–101.

Lai, Angela, Megan A. Brown, James Bisbee, Richard Bonneau, Joshua A. Tucker and Jonathan Nagler. 2022. "Estimating the Ideology of Political YouTube Videos." To appear.

Ledwich, Mark, Zaitsev Anna. 2020. "Algorithmic extremism: Examining YouTube's rabbit hole of radicalization." *First Monday* 25(3).

Lelkes, Yphtach. 2016. "Mass polarization: Manifestations and measurements." *Public Opinion Quarterly* 80(S1):392–410.

Nicas, Jack. 2018. "How YouTube Drives People to the Internet's Darkest Corners.".
URL: *https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478*

Poole, Keith T and Howard Rosenthal. 1985. "A spatial model for legislative roll call analysis." *American Journal of Political Science* pp. 357–384.

Ribeiro, Manoel H., Raphael Ottoni, Robert West, Virgílio A.F. Almeida and Wagner Meira Jr. 2020. Auditing Radicalization Pathways on YouTube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* pp. 131–141.

Roose, Kevin. 2019. "The Making of a YouTube Radical." *The New York Times (June 2019). https://www. nytimes. com/interactive/2019/06/08/technology/youtuberadical. html, https://www. nytimes. com/interactive/2019/06/08/technology/youtube-radical. html* .

Schroeder, Joanna. 2019. "Racists Are Recruiting. Watch Your White Sons.".
URL: *https://www.nytimes.com/2019/10/12/opinion/sunday/white-supremacist-recruitment.html*

Tufekci, Zeynep. 2018. "YouTube, the great radicalizer." *The New York Times* 12:15.

Weill, Kelly. 2018. "How YouTube built a radicalization machine for the far-right.".
URL: *https://www.thedailybeast.com/how-youtube-pulled-these-men-down-a-vortex-of-far-right-hate*