Evaluating echo chambers, rabbit holes, and radicalization pathways on YouTube

Megan A. Brown,[1‡] James Bisbee,[2] Angela Lai,[3]
Richard Bonneau,[4] Jonathan Nagler,[5,6] Joshua A. Tucker[5,6]

[1]School of Information, University of Michigan
[2]Department of Political Science, Vanderbilt University
[3]Firsthand AI
[4]Genentech / Roche
[5]Center for Social Media and Politics, New York University
[6]Politics Department, New York University

[‡]To whom correspondence should be addressed: mgnbrown@umich.edu

Abstract

To what extent does the YouTube recommendation algorithm push users into echo chambers, rabbit holes, or radicalization pathways? Using a novel method to estimate the ideology of YouTube videos and an original audit design, we demonstrate that YouTube users are in mild ideological echo chambers, but this is primarily driven by user behavior. We also demonstrate that the recommendation algorithm prioritizes content that is similar to what the user is currently watching, producing content rabbit holes. However, we do not find evidence of radicalization pathways where users are driven into increasingly extreme content rabbit holes. Instead, 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.

Keywords: YouTube, Recommendation Algorithm, Echo Chambers, Radicalization, Selective Exposure, Information Foraging

Introduction

By many measures, mass polarization is on the rise in the United States (Finkel et al., 2020). While there are many explanations for the growth of mass polarization in recent years, a prominent concern emphasizes the effects of a rapidly evolving digital 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. The process by which users arrive in these information environments is also of normative concern, with the concept of a "rabbit hole" used to describe an unexpected and iterative process through which users might arrive on content in which they had not initially been interested (Sutton and Douglas, 2022). Finally, the combination of these phenomena might produce a "radicalization pathway," in which each subsequent piece of content consumed pushes users into more ideologically extreme directions, while at the same time stifling exposure to contradictory information to this more extreme concept (i.e., that radicalization pathway ends up at a place that is itself on echo chamber).

Empirical evidence of user preference for homophilous networks of such echo chambers is plentiful (Bakshy, Messing and Adamic, 2015) and consistent with a well-developed literature on selective exposure dating back to the 1940s (Sears and Freedman, 1967). However, research on the prevalence of echo chambers is mixed, largely showing only modest evidence of their prevalence (Guess, 2021; Ross Arguedas et al., 2022; Barberá et al., 2015). 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, much of the empirical evidence of echo chambers finds that they are primarily a reflection of user behavior (Ribeiro et al., 2020; Bakshy, Messing and Adamic, 2015; Chen et al., 2021) rather than platform features. On the other hand, mainstream media argues that these platforms – and specifically the algorithms

that use recommendation systems to suggest content to users – are instrumental in pushing people into echo chambers (Nicas, 2018; Weill, 2018; Roose, 2019; Schroeder, 2019) and may even radicalize some (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., 2021; Chen et al., 2021). Datasets assembled via anonymous scraping methods – i.e., relying on Application Programming Interfaces (APIs) or using "headless" browsers to scrape platforms – disconnect the sophisticated recommendation algorithms from the information on which they rely to operate – prior user behavior – and are therefore of questionable construct validity (Ledwich, 2020; Ribeiro et al., 2020).

In this paper, we define a set of three theoretically important concepts that are at the core of the debate on YouTube: ideological echo chambers, where recommendations are ideologically homogeneous and consistent with the user's ideology; content rabbit holes, where users are drawn into increasingly niche recommendations; and radicalization pathways, where ideological echo chambers and rabbit holes together generate pathways by which users encounter increasingly extreme content. 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 in a survey of U.S.-based YouTube users we fielded in the fall of 2020 in which we experimentally manipulated aspects of real users' experiences on YouTube to overcome the previously described limitations with existing empirical work and separate the influence of individual user choices from platform-

developed recommendation algorithms on downstream information environments. These data provide us with ecologically valid measures of how YouTube's recommendation algorithm suggests 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 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 most public commentary. However, our most consistent and striking finding is for content rabbit holes, suggesting that the algorithm's primary goal is to keep users watching content similar to that in which they are currently interested.[1]

Our paper makes several contributions to the literature. First, we define three distinct concepts of online information environments and links them with familiar spatial models of utility-maximizing individuals operating within profit-maximizing institutions. The definitions and their underlying theories contribute a coherent framework for understanding not just YouTube's recommendation algorithm *per se*, but any online recommendation system which are – in 2025 – ubiquitous across all types of platforms. Second, we develop an audit method that can be extended to other recommendation systems of interest to political communication scholars such as TikTok's For You Page, Facebook's News Feed, and X (formerly Twitter)'s Home Timeline. Third, we gather and analyze 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. Fourth, we engage with public and scholarly concerns with recommendation algorithms, finding little evidence to support the claims made in the popular press that the YouTube recommendation system radicalizes its users.

[1]Note that this is not intended as a normative statement about *why* the algorithm seems to encourage content echo chambers; it may of course be the case that giving users more of the same is a simple heuristic for an engagement maximizing algorithm to follow.

1 Echo Chambers, Rabbit Holes, and Radicalization

Radicalization is a process by which users may, by consuming online content, develop more extremist attitudes. Observational evidence of radicalization is causally confounded because a given user could already be radicalized and seek out more extreme content, and the counterfactual for if a user would not have been radicalized if it weren't for the presence of such online content is unobservable. In this study, we test whether the latter pathway is open by isolating the independent effect of YouTube's recommendation algorithm on the content to which users are exposed.

We divide the concept of online radicalization into two constitutive parts: echo chambers and rabbit holes. To provide a road map of what follows, we structure our definitions hierarchically as illustrated 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). At the simplest level, each piece of content (e.g., a video on YouTube) has its own ideology, which can be placed on this single dimensional left-right spectrum (panel 1 in Figure 1). The static distribution of these pieces of content at the level of an individual user captures our definition of echo chambers (panel 2.a); while the dynamic process by which these distributions evolve captures our definition of rabbit holes (panel 2.b). Each of these phenomena on their own are not necessarily problematic for information environments, but when they are combined they can produce a radicalization pathway (panel 2.c). In the following three subsections, we define each of these concepts more precisely.[2]

[2]We elaborate on the underlying intuition and theory driving these conceptualizations in the Supporting Materials (section 1).

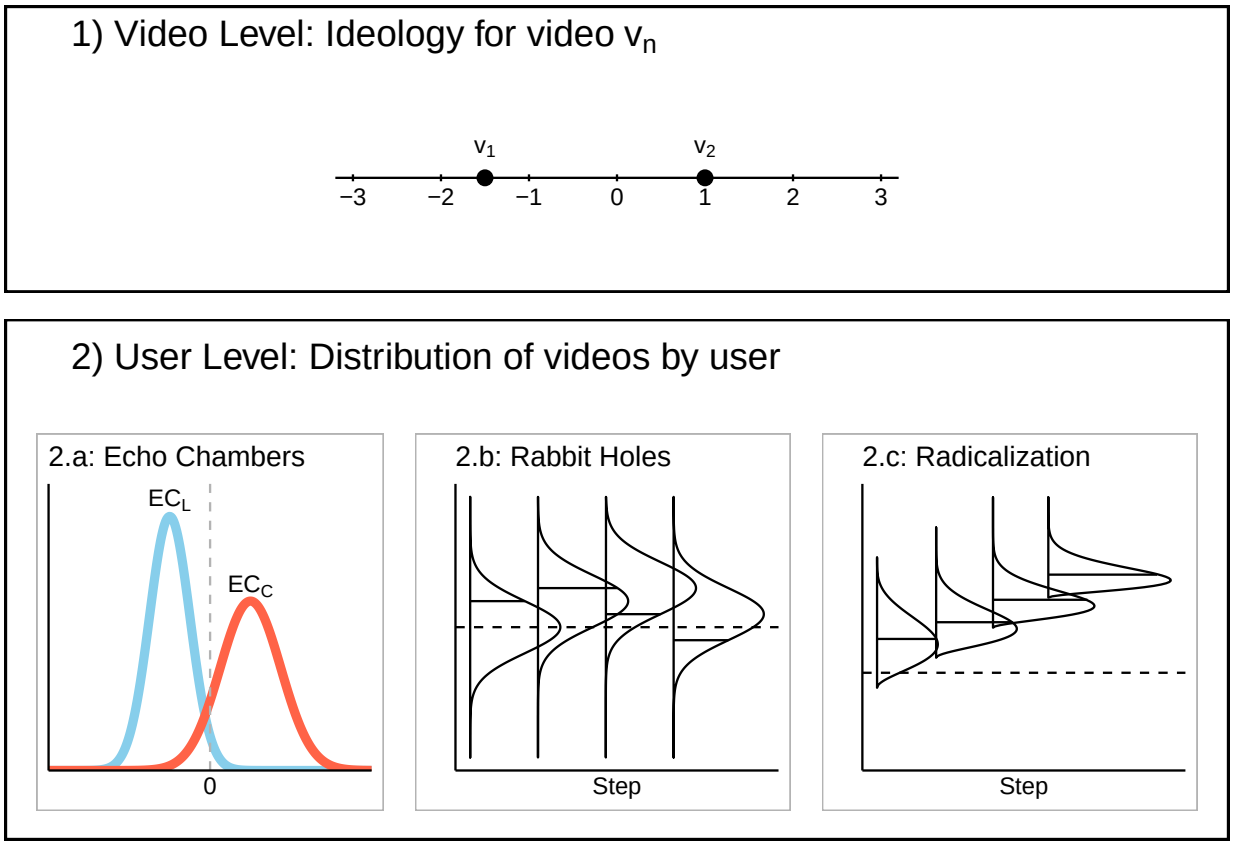

Figure 1: By arraying videos on a left-right spectrum of ideology (panel 1), we can characterize concepts as distributions of videos (panel 2). Echo chambers are user-specific at a single point in time, and are a function of the variance of the distribution, where a more extreme echo chamber is characterized by a tighter distribution (as illustrated by the blue distribution, panel 2.a) while a less extreme echo chamber exhibits greater heterogeneity (as illustrated by the red distribution, panel 2.a). Rabbit holes are dynamic sequences of video distributions where individuals' recommendations are a function of the video currently being watched (as illustrated by the sequence of distributions in panel 2.b, where each subsequent distribution is centered on the video selected from the preceding distribution, indicated with a horizontal black bar). Radicalization occurs when the bias of echo chambers is combined with the serial correlation of rabbit holes. This produces a biased sequence of recommendations where individuals start on a diverse moderate set of videos and are sequentially recommended more ideologically extreme and narrow content as they spend more time on the platform (moving left-to-right across the x-axis in panel 2.c).

1.1 Echo Chambers

Although the term "echo chamber" is widely used, we build on a definition provided in Ross Arguedas et al. (2022) who define an echo chamber as "a bounded, enclosed media

space that has the potential to both magnify the messages delivered within it and insulate them from rebuttal" (pg. 4). For the purposes of this paper, we operationalize 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.[3] As depicted in Panel 2.a of Figure 1, the liberal user L is in a narrower echo chamber than the conservative user C. Importantly, we conceptualize of an echo chamber as a *static* concept, whose prevalence is understood by measuring the degree of separation between two users' information environments. This definition provides an empirically tractable measure of echo chambers that can be captured by simply characterizing the distribution of a user's recommendations: their mean and variance describes the theoretical quantity of interest.

1.2 Rabbit Holes

There is less scholarly work that clearly defines a "rabbit hole", despite the widespread colloquial use of the term to refer to online patterns of consumption. We build on the definition provided in Sutton and Douglas (2022) which emphasizes the initial "incidental" experience of being drawn into to a niche topic, which then becomes a "non-linear descent." While an echo chamber is a static concept, a rabbit hole is dynamic and captures the iterative, associative process by which a subsequent piece of content is related to the previous but covers a different dimension of a topic, or goes into more detail on some aspect. On its own, a rabbit hole needn't (necessarily) draw a user toward one political perspective or another. For example, a rabbit hole could be a user diving into a variety of political perspectives on one specific issue (e.g. immigration). However, it does require that the sequence of information be serially correlated, capturing the iterative and associative qualities defined here. Depending

[3]We rely on the unidimensional space of a left-right ideological spectrum, but argue that the intuition can be extended to multidimensional concepts. For a deeper discussion, please see Supporting Information Section 1.

on the topic, rabbit holes may or may not grow more ideologically homogeneous.

1.3 Radicalization

It is the combination of these two concepts that produces a radicalization pathway. On its own, a rabbit hole might start from a video about a popular video game (ideologically neutral), then lead to a video about the game's creators who emphasized the importance of building a story around a strong female protagonist (ideologically progressive), which then leads to a video about social justice warriors (ideologically progressive), which then leads to a video about the backlash against socially conscious game design in the form of #Gamergate (ideologically conservative), which finally leads to a video by a prominent opponent of this movement: Anita Sarkeesian (ideologically progressive). This example underscores that our definition of a rabbit hole can move across an ideological spectrum of content producing an ideologically diverse set of perspectives, as long as each step is a random draw from the previous distribution.

However, if this process was combined with the definition of a static echo chamber at each step, we might instead observe a radicalization pathway. For example, a user might start on content about Donald Trump, and end on content produced by Holocaust deniers and white supremacists as each subsequent step in their rabbit hole moves in an increasingly conservative, and increasingly narrow, direction (Tufekci, 2018). These radicalization pathways 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 a radicalization pathway 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.[4]

[4]The expectation of an increasingly extreme rabbit hole needn't obtain if the echo chamber component is symmetrically distributed around the user's ideal point. However, as we discuss at greater length in the

2 YouTube's Algorithm and the Supply of Content

Understanding how a recommendation algorithm works requires an antecedent understanding of how users behave. In our Supporting Information, section 1 we build on existing models of selective exposure and information foraging to provide a rich theoretical foundation for how a user might self-select into either an echo chamber or a rabbit hole (or both). Here, we start from the assumption that a user has 1) a relatively time-invariant appetite for a specific type of content and 2) a time-varying interest in a specific type of content. To give an example of the former, a Republican YouTube user might be expected to have a general preference for conservative content, consistent with a rich and well-developed literature on selective exposure dating back to Lazarsfeld, Berelson and Gaudet (1968). To give an example of the latter, a user with a newborn baby might have a specific interest in content about the risks of vaccines and consume a sequence of content about this topic, consistent with more recent communication theories on information foraging (Pirolli and Card, 1999).

These aspects of user demand map on to our definitions of echo chambers and rabbit holes as described above. Specifically, if the time-invariant component dominates, Democrats and Republicans will – on average – consume ideologically congruent content with little overlap, yielding echo chambers. Conversely, if the time-varying preferences dominate, a user will consume a sequence of highly correlated content in a given period, yielding a rabbit hole. Importantly, there is no need for an algorithmic nudge to produce these phenomena. The natural characteristics of a user's demand can yield either echo chambers, or rabbit holes, or both.

SI Section 1, a directional skew can yield preferences for more ideologically extreme content as long as it is in the direction of the individual's "side" of the ideological spectrum. See Goldenberg et al. (2023) for a longer discussion of this phenomenon, which they term "acrophily."

2.1 Separating Demand and Supply

The preceding discussion puts structure on how a user might naturally experience either echo chambers or rabbit holes online without any role being played by recommendation algorithms. As such, we can think of them as demand-driven explanations for why users might wind up in ideological echo chambers, content rabbit holes, or radicalization pathways due solely to their own choices. But what of the supply side?

We start from the assumption that YouTube (and other social media platforms) are profit-maximizers whose main objective is to sell their users to advertisers. To do so, they need their users to spend more time on their platforms. In the context of YouTube, this directly translates to time spent watching videos. Thus any recommendation algorithm should primarily be designed to increase watch time.

Indeed, maximizing watch-time was explicitly the objective function of YouTube's recommendation algorithm, at least as was reported as of 2016 (Covington, Adams and Sargin, 2016). According to the engineers who worked on the recommendation algorithm, it used two embedding representations of users U (i.e., their watch histories, demographics, and socioeconomic characteristics, all of which are either explicitly self-reported from across Google's platforms and its partners, or are inferred from user behavior) and context C (i.e., the time of day, day of week, most recent search terms, current video) to first curate a short list of a few hundred candidate videos which were then sorted in order of which would maximize user watch time.[5]

Mapping these parameters into the theory above, we posit that parts of the U parameter – notably the user's partisanship and ideology – capture the theoretical quantity of echo chambers. An ideological echo chamber is one in which the partisanship of a respondent

[5]Scholarly work suggests that the detail with which a platform can describe its users even without self-reported data is rich (Kosinski, Stillwell and Graepel, 2013).

is prognostic of the ideology of the recommendations the algorithm suggests. Conversely, parts of the C parameter – notably the videos a user has recently watched – capture the theoretical quantity of rabbit holes. Here we are interested in the strength of the serial correlation between the "current" video (i.e., the video currently being played) and the distribution of recommendations that appear next to it. Thus our first two hypotheses can be operationalized as follows:

> **H1 - Echo Chambers:** The distribution of the ideology of recommendations suggested to Republicans should be more conservative on average, and separated from, the distribution of recommendation suggested to Democrats.

> **H2 - Rabbit Holes:** The ideology of recommendations should be positively correlated with the ideology of the video currently being played, along with recently watched videos.

A radicalization pathway can manifest if these two phenomena interact, meaning that the ideology of a set of recommendations is correlated with the current video, and is biased toward the user's partisanship. The underlying logic is that each time a user clicks on a YouTube video, they are shown a set of recommendations that are drawn from a distribution that mixes over the ideology of the video they click (C in the notation above), as well as their partisanship (U in the notation above). Repeated iterations of this process should naturally move Republicans in a more conservative direction and Democrats in a more liberal direction. The extent to which these pathways become increasingly extreme might either be a function of the user's underlying ideology, or a reflection of skewed utility functions.

> **H3: Radicalization Pathways** If users are in ideological echo chambers, and YouTube recommends content that is ideologically similar to the current content, the recommendation system should generate radicalization pathways.

For a deeper discussion of extensions of the spatial model that might give rise to these pathways, please see Section 1 in the Supporting Information.

3 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 creates ideological echo chambers and rabbit holes; and that these phenomena interact to produce radicalization pathways. 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 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 an empirical distribution of the ideological content recommended to each user at each step in their traversal of YouTube's recommendation pathway. 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, rabbit holes, and radicalization pathways.

3.1 Ideology Estimation

We use the ideology scores estimated in the supporting materials to Lai et al. (2024). However, privacy protocols prevent the linking of these raw scores back to our data, as the recommendation video IDs have been anonymized in our replication materials. This method builds on previous literature for estimating ideology in other contexts such as social media (Barberá et al., 2015; Barberá, 2015; Bond and Messing, 2015), legislators (Poole and Rosen-

thal, 1985; Bonica, 2014), and the Supreme Court (Abi-Hassan et al., 2023). This two-stage method starts by creating a correspondence matrix of YouTube videos shared on the platform Reddit (obtained from Baumgartner et al. 2020), where the rows are a YouTube video, the columns are the subreddit the YouTube video was posted on, and the cells contain the logged net upvotes plus 1.

The first stage of the method uses correspondence analysis to extract the "ideology" of a YouTube video as the first dimension of the dimension-reduction result.[6] These YouTube videos that have been shared on Reddit and classified for ideology are then used as training data to train a BERT (Bi-directional Encoder Representations from Transformers) (Kenton and Toutanova, 2019) model to predict a video's ideology score using only the text metadata of the video (i.e, the title, tags, description, and channel title). The trained BERT model is then used to predict the ideology scores for all videos seen by our survey respondents.

The core assumption of this approach – and one that is shared across all similar latent variable approaches to estimating ideology – is that videos are shared on subreddits according to a "homophily" principle: a liberal video would not be shared on a conservative subreddit, for example. This assumption is easier to support in the context of Reddit, where the affordance of "upvoting" or "downvoting" posts operates as a homophily-enforcement mechanism. Even if a liberal video were to be shared on a conservative subreddit, the subreddit's users would presumably downvote it. These net-upvote scores that populate the cells of the correspondence matrix thus provide firmer ground for the necessary homophily assumption. Of course, there remains the possibility that some liberal videos might be shared sardonically on conservative subreddits and accrue a positive net-upvote score, despite not reflecting the underlying ideology of the subreddit. In line with Lai et al. (2024), we assume that these deviations from the homophily assumption are sufficiently rare so as not

[6]Because we restrict the domains of subreddits to those related to U.S. politics, we assume that the dimension of maximum variance is a left-right ideology typically associated with American politics.

to undermine the construct validity of the ideology measure.

This method has been validated extensively, yielding both reassuring out-of-sample performance for the BERT classifier, as well as face validity as measured by (1) comparing the ideology estimates of YouTube channels to existing sources, (2) demonstrating that user watch histories reflect their self-reported partisanship and ideology, and (3) via human annotation tasks. We include additional tests – based on the approaches described in Lai et al. (2024) – in the Supporting Information, Section 8. For the full details of this method, please refer to Lai et al. (2024).

3.2 Audit Design

Isolating the independent effect of personalized recommendation systems is a challenging task. Our method for auditing recommendations builds upon prior research that develops methods for auditing algorithmic systems for bias in areas such as job recruitment, mortgages, loans, online ads, and credit card financing (Cain, 1996; Datta, Tschantz and Datta, 2014; Sweeney, 2013; Sandvig et al., 2014). In the online space, auditing has yielded important findings in the study of political bias in what is recommended to users online, including in Google searches, Twitter searches, Twitter's algorithmic timeline, and more (Robertson, Lazer and Wilson, 2018; Hannak et al., 2013; Kliman-Silver et al., 2015; Kulshrestha et al., 2017; Huszár et al., 2022). We expand on these studies by providing a method for auditing YouTube's recommendation system while accounting for the personalized nature of YouTube's recommendation system.

Online audits, and audits of social media platforms in particular, often take the form of sock-puppet audits—that is, audits that use artificial accounts to simulate user behavior and log the results.[7] Sock-puppet audits have previously been used in research on YouTube (e.g.

[7]Notable exceptions to this type of audit include direct experiments where one set of users is placed

see Ledwich 2020). However, these suffer from challenges with ecological validity (Metaxa et al., 2021). Namely, sock-puppets do not have user history, meaning they lack the personalized results that users may see when browsing YouTube. Or in the language of YouTube's own recommendation algorithm, these studies effectively leave out the U parameter in the $P(w_t = i \mid U, C)$.[8]

Alternatively, researchers could analyze the watch histories of users— the videos the users chose to watch in the order they watched them—by tracking users as they watch YouTube in the course of normal browsing behavior. Watch histories can be collected for analysis from consenting users who are willing to install browser tracking programs, or those who choose to submit their watch histories from the YouTube "Download Your Data" feature. However, this approach runs the risk of confounding the behavior of the recommendation algorithm with user preferences for content, or, to put more succinctly, with user choice. Specifically, because we can only observe what was actually *watched* and not the choice process of how that video was chosen out of the multitude of videos *recommended* by YouTube, 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. Put differently, such an approach risks confounding the supply side of interest with the demand side of user behavior.

Our solution is instead to enroll real users in an audit study where they use YouTube in a prescribed manner to navigate realistic recommendation data. We first prescribe which video a user starts on, which is randomized among twenty-four videos balanced on ideology

into a "default" timeline (e.g. reverse chronological), and the treatment group is placed into algorithmic content feeds. For example, Guess et al. (2023) and Huszár et al. (2022) take this approach for Facebook and Twitter, respectively. However, these are typically done in industry-academic partnerships since external researchers cannot usually manipulate platform algorithms in experimental treatments.

[8]More recent innovations rely on the use of cookies to simulate real users Haroon et al. (2023). However, it is unclear whether or the extent to which these synthetic users accurately reflect the richness of a real human-owned YouTube account.

and non-political topics.[9] This step ensures that participants, regardless of their partisanship, have an equal likelihood of starting on a video of a particular ideology, ensuring that user partisanship does not influence the ideological starting point of the audit. Second, we divide our users into two groups: those in a "preference" condition and those in an "audit" condition. In the former condition, users are instructed to click on whichever video they find most interesting. In the latter condition, we prescribe a rule for choosing among the recommended videos (e.g. always choose the second recommendation). This step ensures that the audit condition is not confounded with user choice where users self-select more ideologically congruent content in the audit condition. Yet by ensuring the users conduct the study while signed into their real accounts, we also ensure that any null results we find are not simply due to limiting the algorithm's ability to suggest the content it normally would to real users. In sum, randomization to both the seed video and traversal rule conditions is not meant to provide a causally identified estimate of these dimensions of experimental manipulation, but rather to isolate the independent influence of the recommendation algorithm.

We should note that this design is not realistic to how users behave on YouTube. We instruct users to visit videos they may not have encountered and click on videos they may not otherwise select. However, the focus of an audit is to probe the behavior of an underlying system by changing inputs (Metaxa et al., 2021)—in our case, changing users, starting videos, and recommendation selections — to better understand the behavior of the algorithm. When using real users with real histories, we strive for ecological validity since we know that user history plays a large role in what people are recommended on YouTube (Davidson et al., 2010). However, by randomizing seed videos across users and restricting their choice of videos to watch next, we claim to disentangle algorithmic behavior from

[9]YouTube classifies videos on its platform according to 15 categories which are assigned by the content creators to increase engagement. In this study, we focus primarily on the "News & Politics" and "People & Blogs" categories. A description of these categories can be found here: `https://entreresource.com/youtube-video-categories-full-list-explained-and-which-you-should-use/`.

user behavior *during* the audit. Essentially, we treat users as "sock-puppets," allowing us to draw inferences about the underlying behavior of the algorithm due to the audit design while maintaining ecological validity by using real user accounts. Our audit design aligns with best practice recommendations for socio-technical audits as proposed in Metaxa et al. (2021). We describe the specific conditions of our audit below.

3.3 Survey Task

From October 2, 2020 to December 7, 2020, we recruited a convenience sample of 1,639 YouTube users using Facebook ads.[10] Participants were required to install a web browsing plug-in to record their YouTube recommendations for the duration of the task. This browser plug-in only was active while the respondent was on a YouTube tab, and deactivated at the conclusion of the study.[11] Crucially, respondents were instructed to be logged into their YouTube accounts for the duration of the task, ensuring that the results recorded would be personalized. Additionally, they answered a brief survey after the fact regarding their demographics and usage of YouTube.[12] We collect participants' self-reported ideology and partisanship, which is used in analyses below that rely on distinguishing videos seen by conservatives from videos seen by liberals.

Participants were compensated $5 for the task and survey and an additional $5 if

[10]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 (section 2). As noted in the Supporting Information, our sample does lean more male, more educated, and younger. However, this is consistent with the population of individuals that use social media more broadly Auxier and Anderson (2021).

[11]The plugin only worked on Chrome-based computer browsers, meaning we are unable to speak to the important but understudied domain of mobile devices which comprise a growing share of how individuals experience online information environments.

[12]The complete survey is available in the appendix.

they uploaded their YouTube video watch history.[13] Each study participant u was asked to complete a "traversal task". For this task, we randomly assigned each participant a starting "seed" video j from one of 24 potential starting videos (consisting of 15 videos categorized as "News & Politics" covering the ideological spectrum, as well as 9 non-political videos categorized as music, gaming, and sports).[14] The user navigated to the video and then was randomly assigned to one of two "traversal rule" conditions k. Half of our sample was instructed to click on the recommended video they found most interesting, which we refer to as the "preference" condition. Among these respondents, we are unable to disentangle the independent influence of the recommendation algorithm from user choice. The other half of our sample was assigned to an "audit" condition in which their traversal task was randomized by design, preventing them from expressing their preferences over the recommendations they were shown. In the audit condition, respondents were assigned to one of five traversal rules: 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 t, 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, indexed by i).[15]

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., 2024). We visualize an example of the traversal results for a given respondent in Figure 2, arraying the recommendations shown at each traversal step

[13]YouTube watch histories are zipped files that users can download which contain the full history of the videos they have watched while signed into their account. Slightly less than half of our participants opted to provide these data in return for the additional $5 inducement.

[14]Our main analyses focus on the respondents assigned to the "News & Politics"-categorized seed videos. For a list of the seed videos, see the Supporting Information (section 3).

[15]Our participants largely complied with their assigned traversal rules. We provide a detailed description of compliance in Supporting Information, Section 2.2.

(x-axis) by predicted ideology (y-axis). This particular 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 0 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 fourth video in the list of recommendations, which we highlight with a black border and line linking the current video with the subsequent video. We construct a respondent-by-recommendation dataset where for a given user, for whom we know demographics and general YouTube habits, we have a 20-by-20 set of ecologically valid recommendations like the one outlined in Figure 2. Each row is therefore a recommended video i suggested to user u at traversal step t, who arrived there from seed video j and following traversal rule k. In all subsequent regressions, we cluster the standard errors at the unit at which our treatment conditions were randomized: the user.

3.4 Evaluating Recommendations

To convert our rich respondent-by-recommendation data into a format that will allow us to empirically measure these three concepts, we can use the empirical traversal in Figure 2 as a motivating example, which starts on a moderate seed video. We can see that the recommendations suggested 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, nor is there evidence of the respondent being pushed down an extremist rabbit hole. Conversely, in Figure 3 we show an experience from a different respondent.

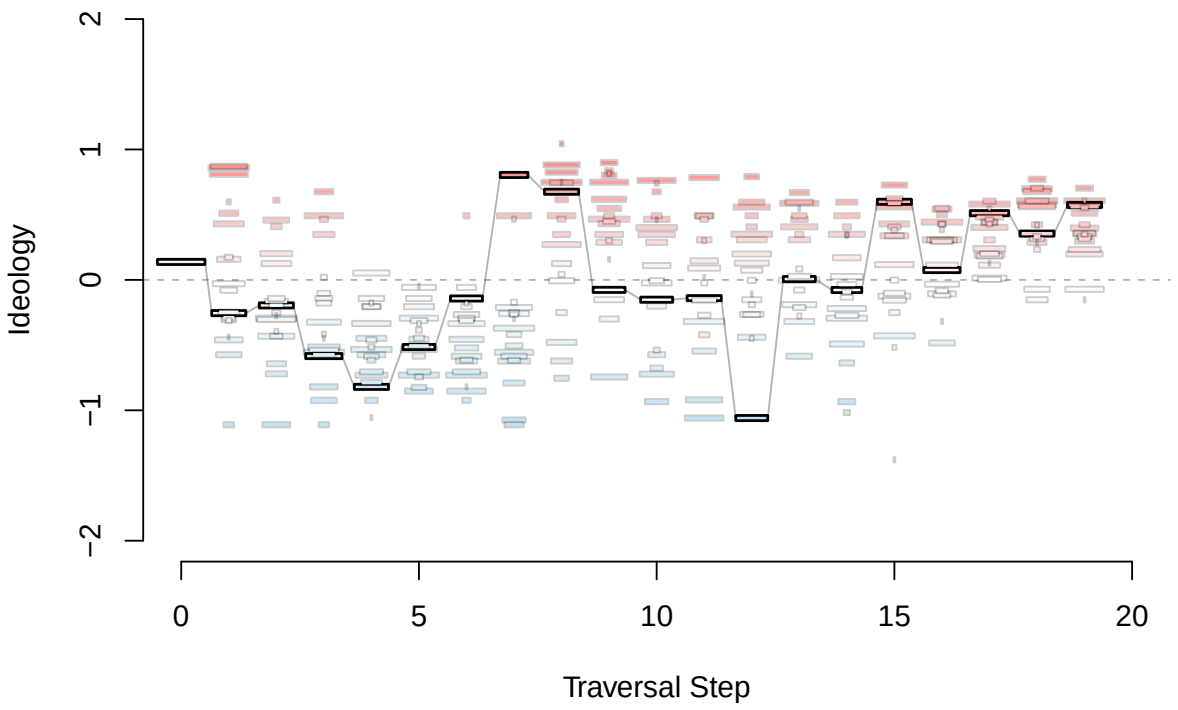

Figure 2: Example of an empirical traversal: On the x-axis we show the traversal step, and on the y-axis the estimated ideology of the video. Positive values indicate that the video is more conservative while negative values indicate that the video is more liberal. Videos outlined in black are those 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.

This respondent also starts on a moderate video, which has a somewhat wide distribution of recommendations. But after the second step, the respondent's recommendations become very conservative and very narrow. They remain this way for the duration of the traversal.[16]

The contrast between the two example respondents highlights how our theoretical quantities of interest – echo chambers and rabbit holes – appear in the data. The first re-

[16]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.

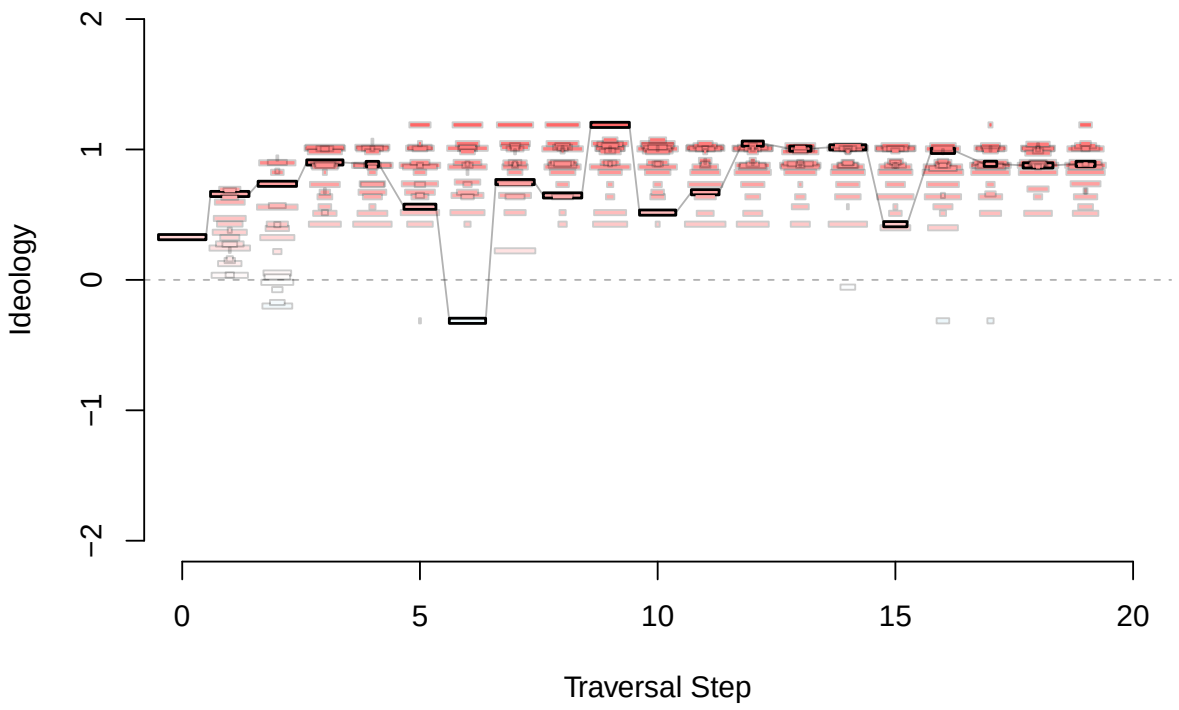

Figure 3: Example of an empirical traversal: The respondent starts on a center right video and randomly selects the next video. We show the distribution of ideology of the recommendations where each recommendation is sized by its rank in the list of recommendations. Videos that appear higher in the recommendations are sized larger.

spondent 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. 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, 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.

We comprehensively evaluate our hypotheses by aggregating over recommendations, traversals, and users. Recall from hypothesis 1 that Republicans would be shown more conservative recommendations than Democrats on average if echo chambers exist and are driven by the recommendation algorithm. To investigate this implication, we predict the ideology of recommendations as a function of the user u's self-reported partisanship, binarized into Republicans and non-Republicans (GOP_u). To investigate rabbit holes, we examine the extent to which the distribution of recommendation is a function of the "context" C, which we operationalize using both the seed video to which users were randomly assigned ($\text{seed}_{u,j}$), as well as the ideology of the current video that they are watching ($\text{current}_{t,u,j,k}$). Our full regression specification can be written as:

$$\begin{aligned} y_{i,t,u,j,k} = \alpha_j + \delta_k & \\ + \beta_1 \text{GOP}_u + \beta_2 \text{history}_u & \\ + \rho_1 \text{current}_{t,u,j,k} + \rho_2 \text{seed}_{u,j} & \\ + \lambda \mathbf{X}_u + \varepsilon_{i,t,u,j,k} & \end{aligned} \tag{1}$$

where α_j represent fixed effects for the seed video, δ_t are fixed effects for the traversal step, and $\mathbf{X}$ is a vector of controls measured at the user level, including age, gender, education, income, and race. In the results that follow, we disaggregate the full specification to focus first on the user-level predictors party_u and history_u, then isolating the context-level pre-

dictors $\text{current}_{t,u,j,k}$ and $\text{seed}_{u,j}$, before combining them in the single specification described in Equation 1. For specifications only focused on the current video ideology, we implement user-fixed effects, dropping the $\mathbf{X}$ matrix.

The β coefficients capture the user-level information which we interpret as evidence of ideological echo chambers: substantively, positive coefficients indicate that the more conservative a user is (or the more conservative is the content they choose to watch), the more conservative are the recommendations shown to them. The ρ coefficients capture the context-level information which we interpret as evidence of rabbit holes: substantively, positive coefficients indicate that when the current video is more conservative (or the randomly assigned seed video is more conservative), the recommendations are more conservative. We estimate both components of the regression separately first, before combining them to explore whether and how our conclusions change.

Across all of these tests, we compare the experiences of those in the "audit" condition to those in the "preference" condition. Recall from above that participants in the "preference" condition were randomly assigned to the same set of 24 seed videos as those in the audit condition, but that they could then express their preferences for YouTube content by clicking on whichever recommendation looked most interesting to them. Conversely, those in the "audit" condition were constrained to only follow their randomly assigned traversal rule. The contrast between these two conditions speaks to the underlying theories of interest, as well as to the broader research question motivating this study. On the one hand, we might expect to find stronger evidence of echo chambers and rabbit holes among users in the preference condition who can select whichever recommendation is most interesting to them. Under the assumption that users have a preference for ideologically congruent content, we might expect to find stronger evidence of liberal users choosing liberal recommendations, which might then exaggerate the liberal skew in subsequent sets of recommendations. On the other hand, it might be the case that most users have a preference for non-political content,

thereby choosing other types of videos (i.e., entertainment, sports, music) that cluster around a moderate ideology. In our subsequent analyses, we compare the results from the preference condition to those in the audit condition to shed light on the experiences of real YouTube users, and the extent to which the recommendation algorithm influences those experiences. All analyses focus on videos categorized as either "News & Politics" or "People & Blogs", although our substantive conclusions persist when dropping this restriction.

4 Results

We assess the prevalence of echo chambers, rabbit holes, and radicalization pathways in our data. Our analyses combine descriptive visualizations of the raw data with linear regression models described above.

4.1 H1: Echo Chambers

We start by plotting the average ideology of recommendations shown to users, broken out by self-reported partisanship and treatment group, in Figure 4. Based on this visualization, there is no evidence that the recommendation algorithm exerts an independent influence on whether users exist in ideological echo chambers. There is slightly more evidence of a separation by partisanship in the preference condition, suggesting that – to the extent they do exist – ideological echo chambers are demand, not supply, driven. But even here the difference is small (roughly 0.17 units on a scale ranging from -1.5 to +1.5), and there is substantial overlap.

To more formally test this descriptive pattern, we run the regression specified in Equation 1 – focusing only on the user-level characteristics of interest – and summarize the findings in Table 1. Even columns present the results from the respondents in the preference condi-

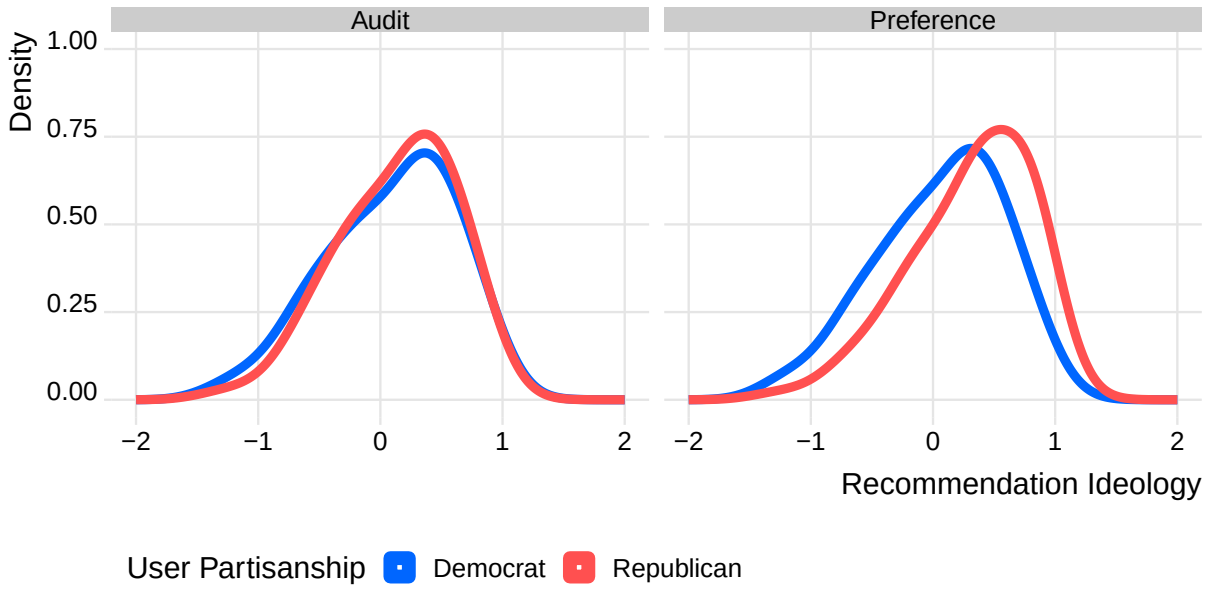

Figure 4: Distribution of ideology of all recommendations (x-axis) shown to users by self-reported partisanship. Left panel displays results from audit condition in which participants were randomly assigned to follow a traversal rule. Right panel displays results from the preference condition in which participants could click on which every video they found most interesting.

tion while odd columns are those in the audit condition. The first two columns look only at user characteristics, comparing self-reported Republicans to non-Republicans (Democrats and Independents). Note that positive coefficients mean the content recommended is more conservative. Consistent with the descriptive visualization presented in Figure 4, Republicans are recommended significantly more conservative content than non-Republicans, but only when they are allowed to express their preferences (column 1). This difference disappears in the audit condition (column 2) suggesting that the recommendation algorithm exerts no independent influence on the difference in ideology of recommendations. A similar story obtains when we replace self-reported partisanship with the average ideology of the users' watch histories among those who opted to provide them, although the positive coefficients are no longer statistically significant (columns 3 and 4). Nevertheless, the coefficients on watch history in the preference condition are between 2 and 3 times as large as those in the audit condition. The summary conclusion persists when we combine both user characteristics and

information from their watch histories (columns 5 and 6): there is evidence of Republicans being recommended more conservative videos than non-Republicans, but only when they are allowed to choose which video to watch next during the traversal task. Thus we conclude that YouTube's recommendation algorithm – on its own – doesn't systematically push users into ideological echo chambers on the basis of self-reported partisanship.[17]

Table 1: Recommendation Ideology by User Characteristics

	User Info		Watch History		Combined	
	Pref	Audit	Pref	Audit	Pref	Audit
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
GOP	0.162***	0.012			0.219***	-0.018
	(0.035)	(0.033)			(0.060)	(0.054)
History Ideo			0.143	0.047	0.082	0.046
			(0.107)	(0.099)	(0.098)	(0.099)
n Users	476	467	188	180	188	180
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Fixed-effects						
Seed Video	Yes	Yes	Yes	Yes	Yes	Yes
Traversal Step	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	43,914	42,442	17,739	15,743	17,739	15,743
R^2	0.105	0.097	0.093	0.094	0.111	0.094
Within R^2	0.025	0.003	0.028	0.018	0.048	0.018

Clustered (User) standard-errors in parentheses
*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05*

[17]Similar conclusions hold when replacing self-reported partisanship with self-reported ideology, although the results are more noisily estimated, likely due to the greater measurement error associated with this approach to characterizing user ideology.

4.2 H2: Rabbit Holes

If the recommendation algorithm, left to its own devices, doesn't push users toward ideological echo chambers, then how does it operate? Here we examine the evidence of rabbit holes: dynamic processes in which the recommendations suggested to a user are predominantly influenced by the *context* in which the user is engaged with the platform. We first provide descriptive visualizations of the relationship between the average ideology of all recommendations shown to a user and the ideology of the seed video to which they were randomly assigned (left panel of Figure 5), and the ideology of the current video they are watching (right panel of Figure 5). As illustrated, there is a modest positive association between the randomly assigned seed video and the recommendations a user was suggested over the course of their time in our study (left panel). But the strength of this association increases dramatically when looking at the relationship between the average ideology of the recommendations suggested and the ideology of the current video being played, although only when we focus our analysis on videos categorized as News & Politics.

To more rigorously test the evidence of rabbit holes, we turn to linear regression analysis. As described in Equation 1, we are interested in the ρ coefficients which capture the association between recommendation ideology and the ideology of the video currently being watched (ρ_1) and the ideology of the seed video to which the user was randomly assigned (ρ_2). We present the results in Table 2, dividing our focus between each context predictor (the seed video and current video ideologies) in isolation, before combining them. Here we find striking evidence of the role played by both the seed video ideology and the current video ideology in predicting the ideology of the videos recommended, regardless of whether the user is in the preference or audit conditions.

Note that the substantive and statistical significance of the seed video ideology disappears in columns 5 and 6 when we estimate the specification with both measures of context included. This suggests that the recommendation system on YouTube – at least over the

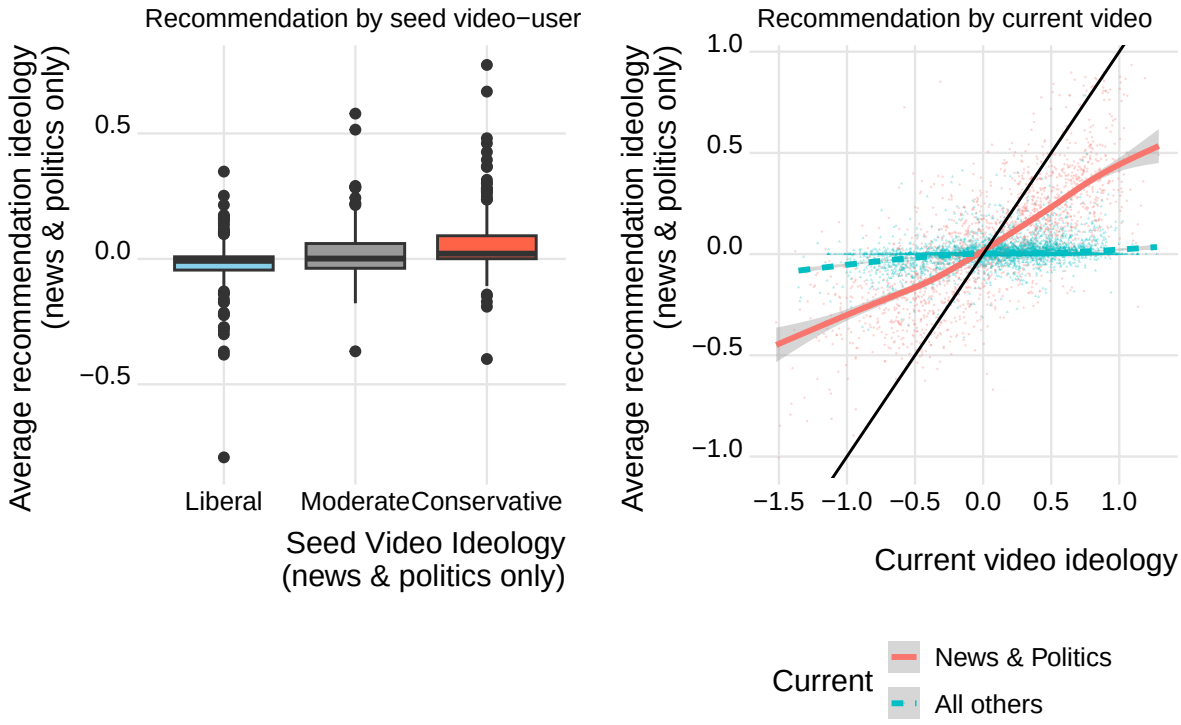

Figure 5: Association between seed video ideology (x-axis left panel) or current video ideology (x-axis right panel) and average ideology of recommendations (y-axes), broken out by video category.

course of several traversals – operates according to a Markov process in which all that matters is the current video. To test this expectation, we subset the data to only the final traversal step and predict the average ideology of the recommendations in the final traversal step as a function of all prior videos watched, including the seed video. We plot the coefficient estimates of the current video at each traversal step on the recommendation ideology at the final traversal step in Figure 6, highlighting that it is only the two most recently watched videos that significantly predict the ideology of the final recommendations.[18] These results demonstrate evidence of "rabbit holes", operationalized as serial correlation between what a user is currently watching and what they are recommended across both the preference and

[18]We apply this specification to every traversal step separately and find substantively similar results, regardless of how deep into the traversal experience a user has traveled. In addition, there is no evidence that assignment to the preference or audit condition matters greatly to these conclusions.

Table 2: Recommendation Ideology by Context

	Current Video		Seed Video		Combined	
	Pref	Audit	Pref	Audit	Pref	Audit
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Current Ideo	0.275***	0.247***			0.409***	0.399***
	(0.016)	(0.014)			(0.019)	(0.018)
Seed Ideo			0.098***	0.068**	0.025	-0.0007
			(0.025)	(0.022)	(0.016)	(0.014)
n Users	534	519	472	465	472	465
Controls	No	No	Yes	Yes	Yes	Yes
Fixed-effects						
Seed Video	Yes	Yes				
Traversal Step	Yes	Yes	Yes	Yes	Yes	Yes
User	Yes	Yes				
Fit statistics						
Observations	47,038	44,910	43,585	42,240	43,160	41,840
R^2	0.307	0.320	0.066	0.034	0.220	0.183
Within R^2	0.065	0.053	0.056	0.016	0.213	0.168

Clustered (User) standard-errors in parentheses
*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05*

audit conditions.

4.3 RQ3: Radicalization

On net, we find little evidence of echo chambers, especially when focusing on users randomly assigned to our audit condition which – we argue – provides an ecologically valid snapshot of how YouTube recommended content to users in the fall of 2020. We do, however, find evidence of "rabbit holes", operationalized as serial correlation between what a user is currently watching and what they are recommended. The obvious concern is whether these rabbit holes are random walks or instead push users toward more extreme ideological content.

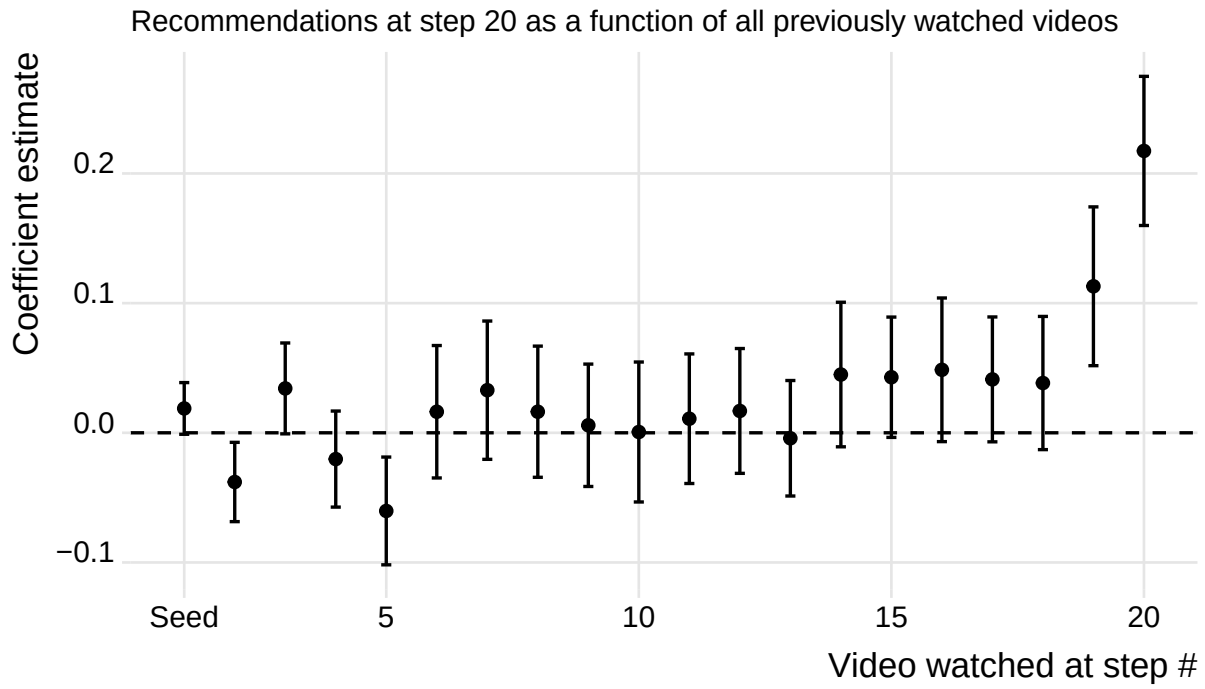

Figure 6: Coefficient estimates (y-axis) connecting the video watched at a given traversal step (x-axis) to the recommendations suggested at traversal step 20.

To evaluate this research question, we again combine descriptive visualizations of the data with regression analysis for inference. As illustrated in the descriptive plot in Figure 7, two patterns jump out. First, recommendations suggested to users who started on a conservative seed video remain roughly constant throughout their time completing the traversal task, whereas those suggested to users who started on a liberal seed video grow increasingly conservative over the course of the task. Second, the diversity of recommendations declines over traversal steps, as illustrated by the narrowing of the box capturing 80% of the data.

Substantively, these patterns suggest a mild radicalization pathway in which recommendations push users toward a right-of-center average, which grows more homogeneous as users spend more time on the platform. But this descriptive visualization obscures important variation by respondent partisanship and treatment condition. To investigate, we run an triple-interacted specification in which we predict the ideology of recommendations as a function of the traversal step, interacted with the user's self-reported partisanship and the

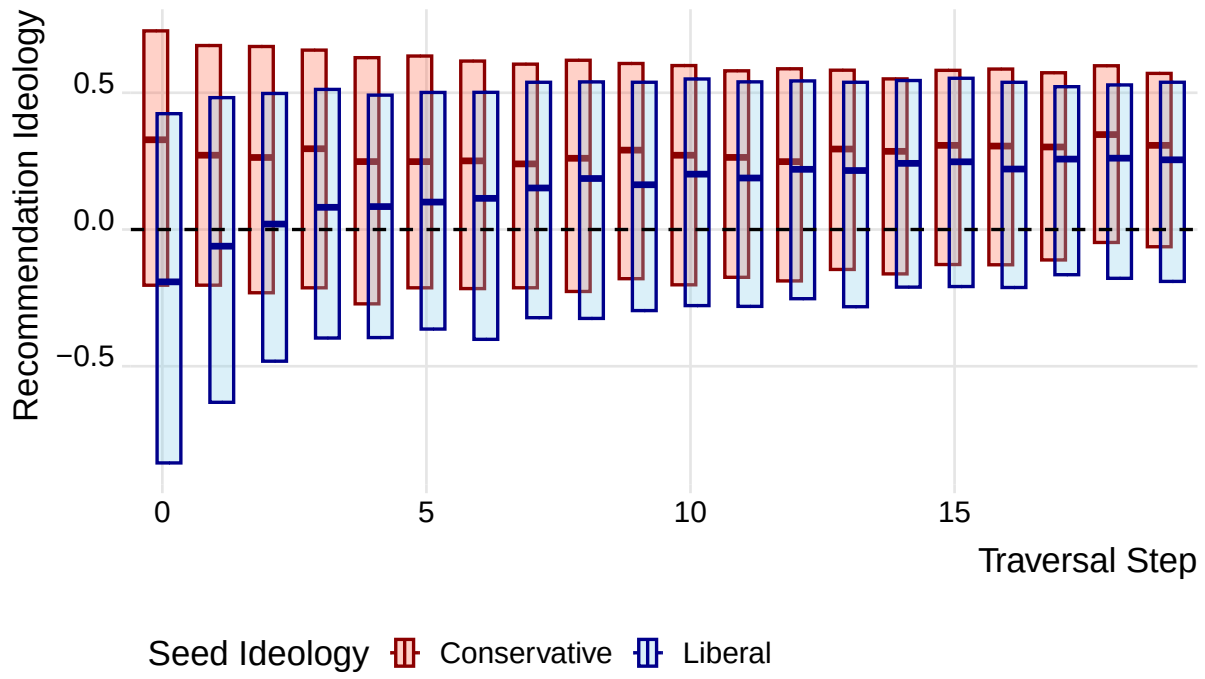

Figure 7: Distribution of ideology (y-axis) of the recommendations shown to users who started on liberal (blue) and conservative (red) seed videos over the course of 20 traversal steps (x-axis). Boxes capture 80% of the data while horizontal lines reflect the median recommendation ideology.

ideology of the randomly assigned seed. We also predict the variance of the recommendations at each step using the same specification, and plot the predicted values of both models in Figure 8. Formally:

$$
\begin{aligned}
y_{i,t,u,j,k} = \alpha_j & \\
& + \gamma_1 \text{step}_k + \rho_1 \text{seed}_{u,j} + \beta_1 \text{party}_u \\
& + \gamma_2 \text{step} * \text{seed} + \gamma_3 \text{step} * \text{party} + \gamma_4 \text{seed} * \text{party} \\
& + \gamma_5 \text{step} * \text{seed} * \text{party} \\
& + \beta_2 \text{history}_u + \rho_2 \text{current}_{t,u,j,k} \\
& + \lambda \mathbf{X}_u + \varepsilon_{i,t,u,j,k}
\end{aligned}
\tag{2}
$$

where y is either the recommendation ideology (indexed by each recommended video i) or

the variance of the recommendations at a given traversal step (dropping the index i).

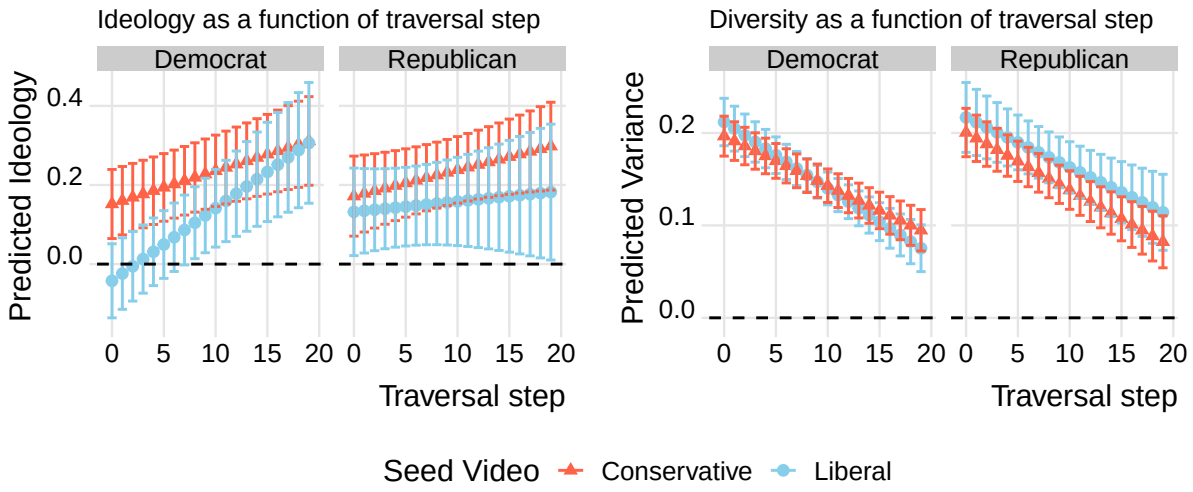

Figure 8: Predicted ideology of recommendations (left panel) and predicted variance of recommendations (right panel) as a function of traversal step (x-axes), ideology of seed video (colors, and self-reported partisanship (facet tiles). Data subset to those in the audit condition.

There are three patterns to highlight from these results. First, there is no evidence that the algorithm separates Democrats from Republicans via radicalization pathways as they spend more time on the platform. Instead it appears that both groups of users are pushed toward the same ideological content as they follow their traversal rule. Second, the diversity of the recommendations declines over the course of a user's time on YouTube. Regardless of partisanship and starting seed video, all users arrive at a more ideologically homogeneous set of recommendations compared to the videos recommended at the beginning of their time on the platform. Third, there is clear evidence of a platform-wide conservative bias. Regardless of both the user's partisanship and the starting seed video ideology, YouTube's algorithm pushes users toward more conservative content writ large.

5 Discussion

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 Republicans see content that is more conservative than Democrats, the magnitude of this difference is small, and driven by user behavior of the participants randomly assigned to the "preference" treatment condition. YouTube's recommendation algorithm exhibits no independent influence pushing Democrats and Republicans into different information environments. Conversely, we find strong evidence of what we define as "rabbit holes": dynamic processes in which each subsequent set of recommendations is heavily influenced by the video the user is currently watching. We show that these patterns exhibit properties of a Markov process in which the influence of previously played videos decays rapidly, disappearing altogether after two lags. Furthermore, we find evidence that the ideological distribution of videos recommended narrows over time, but this does not differ systematically by user partisanship.

Despite the evidence of rabbit holes, we find little support for radicalization pathways that push Democrats toward more extreme liberal content and Republicans toward more extreme conservative content. Instead, we find that, despite the mild differences between the experiences of Democrats and Republicans on the platform, all users regardless of partisanship receive more conservative and less ideologically diverse recommendations over time. We are agnostic about why YouTube's recommendation algorithm exhibited this conservative bias in the fall of 2020. One intuitive explanation would be that there is simply more conservative content on the platform. Another might be that conservative content is more attractive along other dimensions to which the algorithm is responding such as likes or views. We leave a more thorough investigation of these patterns to future work, but present some preliminary descriptive evidence in the Supporting Information Section 10, suggesting that both explanations are at play: a random sample of 1.7 million political YouTube videos skews conservative on average, and more conservative content is more popular on the platform.

However, these results should be interpreted with caution. 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. In addition, we only look at twenty traversal steps within a single session on YouTube. While our results do show a statistically significant ideological shift towards more conservative content, we urge caution in interpreting these findings as an infinitely increasing ideological shift. When we rerun our analysis with a curvelinear specification (provided in section 5 of the Supporting Information), we find that there is more movement towards conservative content in the initial traversal steps, which then tapers off the longer the users follow recommendations. Thus, we cautiously infer that users following the recommendation algorithm out one hundred or one thousand traversal steps would not be recommended infinitely increasing conservative content.

We also note that these findings are specific to the context in which we collected the data; that is, they reflect what YouTube was recommending users in the fall of 2020 when we conducted our study. Platform recommendation systems are regularly modified by the companies that generate them, which cannot be accounted for in our study. However, our study provides an analysis of what YouTube was recommending to real users, which has not previously been analyzed at scale using the audit framework we apply. Moreover, we provide a methodological framework for auditing platform algorithms that allows researchers to isolate the effects of a platform algorithm from confounders like user choice; this framework can be applied to studies in the future to assess the temporal validity of our findings, as well as to test additional hypotheses about the impact of platform algorithms.

Perhaps most importantly, our focus on isolating the independent influence of the algorithm on what is recommended to real users challenges both the ecological and external validity of our results. With respect to external validity, the nature (and resource constraints) of our study meant that users only spent a few seconds on each video during the

process of collecting data. If the algorithm adapts in near-real time to user behavior, it is possible that the lack of evidence of echo chambers found in the audit condition simply reflect the appearance of users expressing disinterest in a variety of videos. Despite this concern, our summary conclusions persist when we restrict attention to only the recommendations shown on the seed video – i.e., before the perceived lack of user interest could affect the recommendations (see SI section 6). Furthermore, if the relative influence of the context (i.e., current video characteristics) indeed dominates that of the user characteristics (i.e., partisanship), our audit setting should be a relatively hard test, since the short durations spent on each video by our participants would signal a lack of interest in a given video. That we nevertheless conclude that context dominates is perhaps a lower bound on this pattern.

Finally, our study sacrifices some important aspects of ecological validity. Separating user behavior from the algorithm is an extreme, stylized, and unrealistic reflection of how these phenomena interact in the real world. In truth, user behavior and the recommendation algorithm must be mutually constitutive in a way that precludes any tidy separation of cause and effect. The very notion of an "acyclic" relationship presumed by our interest in isolating the effect of the algorithm on recommendations does not reflect the deeply cyclic relationship between the algorithm and user behavior. Although we show that there is no difference in the ideology of recommendations shown to Democrats and Republicans in our audit condition, while demonstrating a statistically significant gap between Democrats and Republicans when they are allowed to express their preferences, this doesn't mean that online echo chambers are the 'fault' of users, nor that the black-box recommendation algorithms (and the profit-seeking social media companies who develop them) are exonerated in their role in contributing to normatively troubling information environments. By our own empirical conclusions, we show that user behaviors reveal a desire for ideological echo chambers, and that recommendation algorithms give users what they (appear to) want. Although we conclude that the recommendation algorithm is not, on its own, producing these types of outcomes, it is also clear that it is not suppressing them either.

Nevertheless, while user preferences and the recommendation algorithm are mutually constitutive phenomena, it remains important to examine them separately to the extent that it is possible in audit studies like the one we demonstrate here. Identifying to what extent user behavior versus platform algorithms generate radicalization pathways (or drive consumption of other types of harmful content such as hateful content, disinformation, or conspiracy theories) is vital to informing policy and designing platforms that promote pro-social information environments. We do not find that YouTube's recommendation algorithm is driving users into radicalization pathways, suggesting that focusing on solely YouTube's recommendation algorithm would not lead to effective interventions for reducing radicalization facilitated by content hosted on YouTube. While beyond the scope of our empirical evidence, it is worth considering alternative solutions to curating socially healthy online information environments that go beyond an exclusive focus on the recommendation algorithm. It is possible that being overly focused on the potential harms from the recommendation algorithm may have obscured the more simple potential for harms related to YouTube's function as a repository for content that can be directly accessed via links to videos.

6 Conclusion

Our study considers how interactions between utility-maximizing individuals and profit-maximizing institutions affect the consumption of political content in online information environments. We define echo chambers, rabbit holes, and radicalization pathways and theorize how online recommendation systems, when combined with the well-documented human behaviors of selective exposure and information foraging, can lead to these substantively important phenomena. We develop a novel research design to examine the prevalence of echo chambers, content rabbit holes, and radicalization pathways through an audit with real users on YouTube. We find only minimal evidence of echo chambers separating Republicans from Democrats, and further show that most of the difference between ideological content rec-

ommended to Republicans and Democrats is driven by the preferences of users themselves. However, we do find substantial evidence of what we define as content rabbit holes, dynamic processes in which each subsequent set of recommendations is heavily influenced by the video the current user is watching. In our audit, the ideology of the current video was strongly predictive of the ideology of the recommendations. Yet, these content rabbit holes we find do not interact with echo chambers to produce radicalization pathways on average. Instead, we find that YouTube recommends moderately conservative content to all users, regardless of their partisanship. Our results can inform the ongoing scholarly and public debate on the role of recommendation algorithms in our information ecosystem.

Competing Interests

No potential competing interest was reported by the author(s).

Funding

We gratefully acknowledge that the Center for Social Media and Politics at New York University is supported by funding from the John S. and James L. Knight Foundation the William and Flora Hewlett Foundation, and the Siegel Family Endowment. Funders supporting this work included the above and the Charles Koch Foundation. In addition, this work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request. The data are not publicly available due to privacy protocols that prevent the linking of raw ideology scores to video IDs, which have been anonymized in the replication materials to protect participant privacy.

References

Abi-Hassan, Sahar, Janet M Box-Steffensmeier, Dino P Christenson, Aaron R Kaufman and Brian Libgober. 2023. "The Ideologies of Organized Interests and Amicus Curiae Briefs: Large-Scale, Social Network Imputation of Ideal Points." *Political Analysis* 31(3):396–413.

Auxier, Brooke and Monica Anderson. 2021. *Social Media Use in 2021*. Pew.
URL: *https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/*

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.

Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*. Vol. 14 pp. 830–839.

Bond, Robert and Solomon Messing. 2015. "Quantifying social media's political space: Estimating ideology from publicly revealed preferences on Facebook." *American Political Science Review* 109(1):62–78.

Bonica, Adam. 2014. "Mapping the ideological marketplace." *American Journal of Political Science* 58(2):367–386.

Cain, Glen G. 1996. *Journal of Economic Literature* 34(1):165–167.
URL: *http://www.jstor.org/stable/2729440*

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.

URL: *https://www.adl.org/resources/reports/exposure-to-alternative-extremist-content-on-youtube*

Covington, Paul, Jay Adams and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*. pp. 191–198.

Datta, Amit, Michael Carl Tschantz and Anupam Datta. 2014. "Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination." *CoRR* abs/1408.6491.

URL: *http://arxiv.org/abs/1408.6491*

Davidson, James, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston et al. 2010. The YouTube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*. pp. 293–296.

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.

Goldenberg, Amit, Joseph M Abruzzo, Zi Huang, Jonas Schöne, David Bailey, Robb Willer, Eran Halperin and James J Gross. 2023. "Homophily and acrophily as drivers of political segregation." *Nature Human Behaviour* 7(2):219–230.

Guess, Andrew M. 2021. "(Almost) Everything in Moderation: New Evidence on Americans' Online Media Diets." *American Journal of Political Science* 65(4):1007–1022.

URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12589*

Guess, Andrew M, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow et al. 2023. "How do social media feed algorithms affect attitudes and behavior in an election campaign?" *Science* 381(6656):398–404.

Hannak, Aniko, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove and Christo Wilson. 2013. Measuring Personalization of Web Search. In *Proceedings of the 22nd International Conference on World Wide Web*. WWW '13 New York, NY, USA: Association for Computing Machinery p. 527–538.
URL: *https://doi.org/10.1145/2488388.2488435*

Haroon, Muhammad, Magdalena Wojcieszak, Anshuman Chhabra, Xin Liu, Prasant Mohapatra and Zubair Shafiq. 2023. "Auditing YouTube's recommendation system for ideologically congenial, extreme, and problematic recommendations." *Proceedings of the National Academy of Sciences* 120(50):e2213020120.

Hosseinmardi, Homa, Amir Ghasemian, Aaron Clauset, Markus Mobius, David M Rothschild and Duncan J Watts. 2021. "Examining the consumption of radical content on YouTube." *Proceedings of the National Academy of Sciences* 118(32):e2101967118.

Huszár, Ferenc, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer and Moritz Hardt. 2022. "Algorithmic amplification of politics on Twitter." *Proceedings of the National Academy of Sciences* 119(1):e2025334119.

Kenton, Jacob Devlin Ming-Wei Chang and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*. Vol. 1 Minneapolis, Minnesota.

Kliman-Silver, Chloe, Aniko Hannak, David Lazer, Christo Wilson and Alan Mislove. 2015. Location, Location, Location: The Impact of Geolocation on Web Search Personalization. In *Proceedings of the 2015 Internet Measurement Conference*. IMC '15 New York, NY,

USA: Association for Computing Machinery p. 121–127.
URL: *https://doi.org/10.1145/2815675.2815714*

Kosinski, Michal, David Stillwell and Thore Graepel. 2013. "Private traits and attributes are predictable from digital records of human behavior." *Proceedings of the national academy of sciences* 110(15):5802–5805.

Kulshrestha, Juhi, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi and Karrie Karahalios. 2017. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '17 New York, NY, USA: Association for Computing Machinery p. 417–432.
URL: *https://doi.org/10.1145/2998181.2998321*

Lai, Angela, Megan A Brown, James Bisbee, Joshua A Tucker, Jonathan Nagler and Richard Bonneau. 2024. "Estimating the ideology of political YouTube videos." *Political Analysis* 32(3):345–360.

Lazarsfeld, Paul F, Bernard Berelson and Hazel Gaudet. 1968. *The people's choice: How the voter makes up his mind in a presidential campaign*. Columbia University Press.

Ledwich, Mark, Zaitsev-Anna. 2020. "Algorithmic extremism: Examining YouTube's rabbit hole of radicalization." *First Monday* 25(3).
URL: *https://arxiv.org/pdf/1912.11211.pdf*

Metaxa, Danaë, Joon Sung Park, Ronald E Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, Christian Sandvig et al. 2021. "Auditing algorithms: Understanding algorithmic systems from the outside in." *Foundations and Trends® in Human–Computer Interaction* 14(4):272–344.

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*

Pirolli, Peter and Stuart Card. 1999. "Information foraging." *Psychological review* 106(4):643.

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.

Robertson, Ronald E., David Lazer and Christo Wilson. 2018. Auditing the Personalization and Composition of Politically-Related Search Engine Results Pages. In *Proceedings of the 2018 World Wide Web Conference*. WWW '18 Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee p. 955–965.
URL: *https://doi.org/10.1145/3178876.3186143*

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

Ross Arguedas, Amy, Craig Robertson, Richard Fletcher and Rasmus Nielsen. 2022. "Echo chambers, filter bubbles, and polarisation: A literature review.".

Sandvig, Christian, Kevin Hamilton, Karrie Karahalios and Cedric Langbort. 2014. "Auditing algorithms: Research methods for detecting discrimination on internet platforms." *Data and discrimination: converting critical concerns into productive inquiry* 22(2014):4349–4357.

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

Sears, David O and Jonathan L Freedman. 1967. "Selective exposure to information: A critical review." *Public opinion quarterly* 31(2):194–213.

Sutton, Robbie M and Karen M Douglas. 2022. "Rabbit Hole Syndrome: Inadvertent, accelerating, and entrenched commitment to conspiracy beliefs." *Current Opinion in Psychology* 48:101462.

Sweeney, Latanya. 2013. "Discrimination in Online Ad Delivery: Google Ads, Black Names and White Names, Racial Discrimination, and Click Advertising." *Queue* 11(3):10–29. **URL:** *https://doi.org/10.1145/2460276.2460278*

Tufekci, Zeynep. 2018. "YouTube, the great radicalizer." *The New York Times* 12:15. **URL:** *https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html*

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*