"Filter bubble" algorithmic recommendations have limited effects on polarization: Naturalistic, short-exposure experiments on YouTube*

Naijia Liu[a,1], Xinlan Emily Hu[b,1], Yasemin Savas[c,1], Matthew A. Baum[d], Adam J. Berinsky[e], Allison J.B. Chaney[f], Christopher Lucas[i], Rei Mariman[b], Justin de Benedictis-Kessner[d,2], Andrew M. Guess[g,2], Dean Knox[b,2], and Brandon M. Stewart[h,2]

[a]Department of Government, Harvard University
[d]John F. Kennedy School of Government, Harvard University
[e]Department of Political Science, Massachusetts Institute of Technology
[f]Fuqua School of Business, Duke University
[g]Department of Politics and School of Public and International Affairs, Princeton University
[b]Operations, Information, and Decisions Department, the Wharton School of the University of Pennsylvania
[i]Department of Political Science, Washington University in St. Louis
[h]Department of Sociology and Office of Population Research, Princeton University
[c]Department of Sociology, Princeton University

September 4, 2024

*NL, XEH, YS are co-first authors. Direct correspondence to JdBK, AG, DK, and BMS. Author contributions at the end of the document. This project is supported by funding from the Ash Center for Democratic Governance and Innovation and the Shorenstein Center for Media, Politics, and Public Policy at the Harvard Kennedy School; a New Ideas in the Social Sciences grant from Princeton University; an unrestricted Foundational Integrity Research: Misinformation and Polarization grant from Meta; and the National Science Foundation (Political Science Program Grant SES-1528487, "Collaborative Research: A New Design for Identifying Persuasion Effects and Selection in Media Exposure Experiments via Patient Preference Trials" 2015–2021). Thanks to Jim Kim for excellent research assistance building the video platform and L. Jason Anastasopolous for collaboration in the very early stages of the project. We thank Drew Dimmery, Aleksander Madry, and Michelle Torres for their feedback, the Wharton Behavioral Lab at the University of Pennsylvania for financial support and operational assistance and the Princeton Center for Statistical and Machine Learning for computational support.

Abstract

An enormous body of work argues that opaque recommendation algorithms drive political polarization through the creation of "filter bubbles" and "rabbit holes." Drawing on four experiments with nearly 9,000 human users, we show that manipulating algorithmic recommendations to create these conditions has limited effects on opinions. Our experiments employ a custom-built video platform with a naturalistic, YouTube-like interface that presents real YouTube videos and recommendations. We experimentally manipulate YouTube's actual recommendation algorithm to simulate "filter bubbles" and "rabbit holes" through the presentation of ideologically balanced and slanted choices. Our design allows us to intervene in a feedback loop that has confounded the study of algorithmic polarization—the complex interplay between *supply* of recommendations and user *demand* for consumption—to examine effects on policy attitudes. We use over 130,000 experimentally manipulated recommendations and 31,000 platform interactions to estimate how recommendation algorithms alter users' media consumption decisions and, indirectly, their political attitudes. Our results cast doubt on widely circulating theories of algorithmic polarization by showing that even heavy-handed (although short-term) perturbations of real-world recommendations have limited causal effects on policy attitudes. Given our inability to detect consistent evidence for algorithmic effects, we argue that the burden of proof for claims about algorithm-induced polarization has shifted. Our methodology, which captures and modifies the output of real-world recommendation algorithms, offers a path forward for future investigations of black-box artificial intelligence systems. Our findings reveal practical limits to effect sizes that are feasibly detectable in academic experiments and we conclude with implications for future research.

The ubiquity of online media consumption has led to concern about partisan "information bubbles" that are thought to increasingly contribute to an under-informed and polarized public (1). Prior work has focused on cable TV or textual news, but with the rise of new forms of media, the most pressing questions concern online video platforms where content is discovered through algorithmic recommendations. Critics argue that platforms such as YouTube could be polarizing their users in unprecedented ways (2). The ramifications are immense: more than 2.1 billion users log in to YouTube monthly, and popular political extremists broadcast to tens of millions of subscribers.

Empirical research in this setting has long been stymied by enduring challenges in the causal analysis of media consumption and its effects. While observational studies allow researchers to study media in realistic settings, they often conflate the content's persuasiveness with selective consumption by those who already believe its message. Experiments mitigate the issue of self-selection by randomly assigning participants to view specific videos, but this comes at a cost: forced assignment often eliminates freedom of consumption or limits choices in ways that do not reflect real-world settings (3, 4). In turn, this makes experimental results difficult to generalize to the real-world challenges of greatest importance—whether media causes polarization *among the people who choose to consume it*. The challenges of studying this phenomenon are heightened for social-media platforms—such as YouTube, Facebook, X (Twitter), or TikTok—because their underlying recommendation algorithms are black boxes the inner workings of which academic researchers cannot directly observe. While work such as www.their.tube has powerfully demonstrated that recommendation systems can in theory

supply politically polarized recommendations, evidence on the prevalence of this polarized supply has been limited. More importantly, few existing research designs attempt to connect this algorithm-induced supply of polarized media to demand-side changes in consumer watching decisions, much less the effects of this consumption in terms of polarized attitudes and behavior. The result is a contradictory set of findings providing differing estimates of the amount of potentially-polarizing content, but few investigations of the effects of that content (5–12).

To test widely circulating theories about this phenomenon, we develop a new experimental platform and design to estimate the causal effects of black-box recommendation systems on media consumption, attitudes, and behavior. We designed and built an online video interface that resembles YouTube and allows users to navigate a realistic network of recommendations—the set of options shown after an initial "seed" video, the subsequent options that follow after the chosen second video, and so on—that are directly scraped from the existing YouTube algorithm. Starting with this naturalistic reproduction, which maximizes the ecological validity of the study, we randomly perturb the recommendations shown to users after each video. We continuously track demand-side behaviors such as choices among the recommended videos, skipping decisions, likes, dislikes, and "save to watchlist" actions during their 15–30 minute watch session.

Existing theories of polarizing recommendations come in two variations: "filter bubbles," which serve mostly *static* recommendations based on prior behavior (13), and "rabbit holes" which offer *dynamically* offer increasingly extreme content (2). We address both of these phenomena in separate experiments—focusing on "filter bubbles" which we find to be more empirically common on YouTube.

In the "filter bubble" experiments (Studies 1–3) we use a multi-wave survey to explore how experimental intervention causes individuals to change policy opinions, increase partisan animosity, or alter attitudes toward mainstream media in two issue areas. Figure 1 provides a graphical overview of the design. We then evaluate the "rabbit hole" hypothesis by constructing curated sets of video sequences that are either constant or increasing in extremity and randomly assign participants to watch them in a single-wave study. Below, we present the results of these four studies with a combined N of 8,883. Our analyses draw on over 130,000 experimentally manipulated supply-side video recommendations; more than 31,000 demand-side user decisions to watch, like, dislike, and save to watchlists; and a host of outcomes that measure recommendation-system effects on affective polarization, media trust, and policy attitudes. All experiments were preregistered with the Open Science Framework (see SI 3).

We consistently find that while changes in the recommendation algorithm do affect user demand by shifting the types of videos consumed and the amount of time spent on the platform, they ultimately did not produce the theorized effects on political attitudes in a substantial way. This is despite the fact that we do see effects from the assignment of initial seed videos to ideological moderates. We emphasize that this evidence does not rule out the possibility that YouTube is a radicalizing force in American politics, because our design does not address long-term exposure or potential effects in particularly susceptible sub-populations. Yet, in the most credible study of algorithmic polarization to date, we observe only minimal attitudinal shifts as a result of more extreme recommendations, calling into question widely circulating, unequivocal claims about the influence of algorithmic

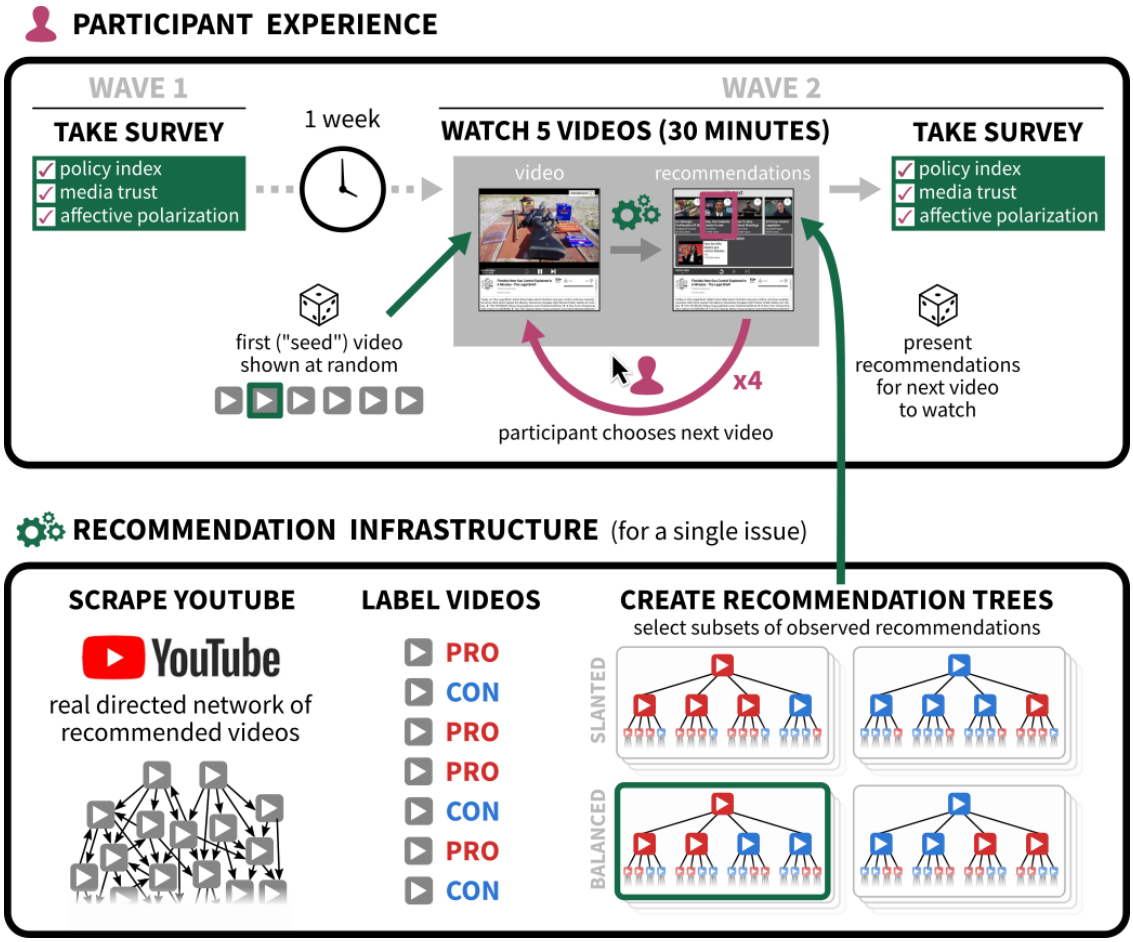

Figure 1: An overview of the design in Studies 1–3. In the first wave, participants answer a series of questions. One week later in the second wave, participants are randomized to a seed video and a recommendation system from which they choose future videos to watch. After watching five videos, they take a followup survey. Experiment 4 uses a similar design in one wave but participants are randomly assigned to a sequence of either constant or increasingly extreme content.

recommendations on political polarization. We are not claiming that polarization from recommendation systems cannot be found *anywhere*, but the consistent lack of short-term effects suggests that it is not *everywhere*.

In the next section, we briefly review the related literature and describe the testable implications of existing theories that characterize YouTube as a radicalizing system, both in terms of shifts in user demand and the effects of those shifts on political attitudes. In Section 2, we describe our survey experimental design, the video-recommendation platform that we built to conduct it, and a manipulation check we conducted to evaluate whether users perceive partisan signals in thumbnails. In Section 3, we present the results from four studies on two policy issues—gun control and minimum wage—detailing the lack of evidence for claims about algorithmic polarization. In the final section, we place these findings in a broader context—re-emphasizing the limitations of what we can know about long-term effects or small, but vulnerable, populations—and propose directions for future work.

1 The Radicalizing Potential of Algorithmic Recommendations

One of the primary theoretical perspectives on YouTube—and algorithmic recommendation systems more generally—contends that users' initial preferences trigger algorithmic personalization, which can generate polarization (see e.g. 2). Recommendation algorithms maximize certain outcomes (watch time, engagement) at the expense of others (long-term satisfaction, information quality). However, the inner workings of these systems are generally opaque apart from occasional published technical details (14–16). Prior work has noted that the circular logic of recommendation-system development, which trains recommendation algorithms on user data that is itself driven by prior algorithmic recommendations, can lead to unanticipated consequences such as homogenization of user behavior (17).

1.1 Theories of Polarization

We draw a distinction between two forms of the argument that recommendation systems contribute to polarization. One is that "filter bubbles" can form when ranking systems are optimized for predicted engagement, resulting in potentially polarizing effects of consuming information from the resulting like-minded sources that appear on the feed (1, 13). Research on the filter-bubble effect has often focused on personalized search results on specific queries (18), with recent studies finding a strong role for user preferences and heterogeneity across topics (19). Looking specifically at social media, the most recent evidence shows that content from congenial or "like-minded" sources constitute a significant share of what users see on Facebook (20), and this is driven in part (though not mostly) by algorithmic personalization (21).

The second form of the argument leverages the "rabbit hole" concept which posits a sequential element to algorithmic curation. Whereas filter bubbles are conceptualized as static—users encounter more ideologically congenial content, compared to an uncurated platform—rabbit holes imply that the curation process serves up content from one's preferred side but with increasing extremity or intensity. For example, Tufekci (2) argues that

YouTube's recommendation algorithm "promotes, recommends and disseminates videos in a manner that appears to constantly up the stakes." She suggests that this occurs via a feedback mechanism in which algorithmic curation reinforces users' preferences, which then drive even more extreme content over time. Similarly, a Wall Street Journal feature on YouTube recommendations found that "[w]hen users show a political bias in what they choose to view, YouTube typically recommends videos that echo those biases, often with more-extreme viewpoints" (22). To show the existence of "rabbit hole" phenomena, both a static (correlation between user preference and curation of congenial videos or channels) and a dynamic (algorithmically served videos become more extreme over time) element must be established. Many existing studies of YouTube recommendations focus on the static element. For example, Hosseinmardi et al. (8) find a correlation between preferences for content elsewhere on the internet and political video channels on YouTube. Other studies attempt to estimate "pathways" between categories of content on YouTube, such as channels classified as mainstream or radical (6, 7). We are aware of one study that attempts to estimate whether algorithmically driven video consumption becomes more extreme over time: Haroon et al. (23) show a significant—but slight—increase in the average extremity of videos shown to sock-puppet accounts as more up-next recommendations are followed. However, extremity in this study was determined by estimating the ideology of Twitter accounts that share links to specific videos, a method that may be sensitive to the sparsity of the data.

1.2 Approaches to Studying Opinion Change

The circular interaction between past preferences (which shape the set of recommended videos and how users choose among them) and consumption (which shapes future preferences by changing recommendations and user tastes) leads to severe challenges in the study of media persuasion and preference formation. There is a venerable social-science tradition that has used experiments to understand the persuasive effects of films and videos (24). The standard "forced-choice" design assigns one group to a video condition with another assigned to a control or placebo condition, with neither group provided alternatives or given the option to avoid the stimulus (e.g., 25). This allows analysts to cleanly estimate the effect of forcing the entire population to consume one piece of media instead of another. Yet this counterfactual quantity focuses entirely on media supply and neglects the interplay with user demand. As a result, it is of limited value in studying high-choice environments when self-selection is the primary determinant of media selection. More recently, scholars have studied the interaction of user choice and media effects in a related literature on partisan cable news (3, 4, 26). A key insight of these works is that the persuasiveness of partisan news varies across individuals with different preferences: effects are different for those who prefer entertainment, compared to those who prefer ideologically congenial news sources (27). Related insights inform the current literature on the effects of digital media and social media (28–30).

To account for the role of user demand in persuasion, Arceneaux et al. (3) develop active audience theory, which emphasizes people's goals and conscious habits in deciding what types of content to consume. On the one hand, some people may prefer to consume partisan or biased media (26, 31, 32); on the other, this media can alter future preferences. Crucially, the interaction of these phenomena could unleash a spiral of rising polarization and self-isolation

(33). Recent work has sought to estimate the causal effect of partisan media *specifically on those who choose to consume it* (3, 34, 35)—the quantity that matters most in real-world polarization, since a substantial part of the population voluntarily opts out of exposure.

The existing literature on algorithmic recommendations can similarly be broken down in terms of media recommendations (supply), media consumption (user demand), and the effects of this consumption on user preferences and attitudes. Existing work has generally focused on understanding the demand side of the problem. In an influential study, Ribeiro et. al. (7) collect video metadata, comments, and recommendations covering 349 channels, more than 330,000 videos, and nearly 6 million commenting users. By connecting commenters across videos and following networks of recommendations, the authors find that commenters in less-extreme "alt-lite" and "intellectual dark web" (IDW) channels are more likely to subsequently comment on more extreme "alt-right" channels. They also observe a substantial share of channel recommendations from alt-lite and IDW videos to alt-right channels, but they find no evidence of direct recommendations from mainstream media to alt-right channels. These findings are consistent with alternative but less extreme sources serving as a "gateway" to more extremist content—but this observational audit methodology cannot disentangle the role of the algorithm from that of user preferences, nor can it assess the effect of consumption on attitudes or behavior. Brown et. al. (9) use a different design to examine the correlation between the supply of algorithmic recommendations and policy attitudes at a particular moment in time, breaking into the supply-demand loop by eliminating the role of user choice. Participants log into their own accounts, are given a starting "seed" video and instructions to click on the first, second, etc. video recommendation video; the network of recommendations is then explored to a depth of over 20 choices. They estimate a modest correlation between self-reported ideology and the average slant of recommended videos but, counterintuitively, find a consistent center-right bias in the ideological slant of recommended videos for all users. Haroon et. al. (11) extend this approach to examine the interaction between supply and demand, using 100,000 automated "sock-puppet" accounts to simulate user behavior; they argue that YouTube's recommendation algorithm direct right-wing users to ideologically extreme content. However, in another experiment using sock-puppet accounts that initially mimic the browsing history of real users, Hosseinmardi et. al. (12) show that YouTube's recommendations quickly "forgets" a user's prior extremist history if they switch back to moderate content. Haroon et. al. (23) show through a sock-puppet study and a longitudinal experiment on 2,000+ frequent YouTube users that nudges can increase consumption of balanced news and minimize ideological imbalance, but that there is no detectable effects on attitudes.

Other work has used observational methods to study the correlation between demand and policy attitudes, rather than seeking to estimate how an intervention would change those attitudes. Hosseinmardi et. al. (8) examine the broader media ecosystem by tracking web-browsing behavior from a large representative sample; they show that video views often arise from external links on other sites, rather than the recommendation system itself, and conclude that consumption of radical content is related to both on- and off-platform content preferences. Chen et. al. (10) similarly combine a national sample and browser plugins to show that consumption of alternative and extreme content, though relatively rare, is associated with attitudes of hostile sexism; they further show that viewers tend to be subscribed to channels that deliver this content. This suggests that personal attitudes and

preferences—as reflected in the decision to subscribe to a channel—are important factors driving consumption of extremist content, though it does not rule out the possibility that algorithmic recommendation systems play a role in initially exposing viewers to this content.

Taken together, the results imply that though algorithmic recommendations may shape the experience of using video platforms, their effects may be subtler and more complex than we might expect from a simple "rabbit hole" model of radicalization. At a minimum, observational evidence suggests that users' choices to consume content can also reflect their preexisting attitudes and *non-platform* preferences. There is also limited evidence that "rabbit holes" exist in practice. While much of the work has focused on the recommendation or consumption of ideological content, *there is very little research on the causal persuasive effects of the self-selected content or the algorithms that recommend it.*

1.3 Testable Implications

We build on these existing lines of work by developing a realistic experiment to estimate how changes in recommendation-system design (a supply-side intervention) affect user interactions with the platform (demand for content) and, through changes in the content consumed, ultimately cause changes in political attitudes. In our main design, participants are presented with an initial "seed" video and, after choosing to watch or skip it, are offered four videos to select for the next round. By carefully pruning and rewiring the real-world YouTube recommendation network, we create two realistic recommendation algorithms: a "slanted" algorithm (which we call 3/1) that primarily gives options from the same ideological perspective as the most recently watched video (mirroring a "filter bubble"), and a "balanced" algorithm (which we call 2/2) that presents an equal mix of supporting and opposing perspectives. Unlike existing work on the persuasive effects of partisan media, we allow users to choose up to five videos in a single, continuous viewing session. This design mimics real-world viewing behavior and allows us to account for how demand-side choices shape the supply of videos subsequently available to view in a sequence. By experimentally manipulating actual YouTube recommendation networks, our approach combines the causal identification of recent media-persuasion experimental research with the realism of recommendation-system audit research. This produces a research design that can credibly estimate the causal persuasive effects of recommendation algorithms. It allows users to choose the content that they wish to consume, but it prevents this freedom of choice from confounding inferences about the algorithm's downstream effects. By increasing the slant of the algorithm beyond the current levels, we also side-step a challenge inherent in observational studies conducted after YouTube's 2019 algorithm updates—the fact that they are limited in what they can say about algorithm's polarizing potential *before* those changes were made (10). Platforms like YouTube are a moving target (36, 37) but our design suggests that even implementing a dramatically more slanted algorithm has limited effects on opinion formation.

In the analyses that follow, we argue that widely circulating claims about algorithmic polarization imply four testable hypotheses. First, because user behavior is heavily shaped by platform affordances and recommendation systems are designed to influence video consumption, prior observational work (7) suggests that random assignment to a balanced or slanted algorithm will powerfully affect user demand, as measured by the content that users immediately choose to consume. Second, since online video systems are part of a broader

alternative-media ecosystem (38), supply-side changes in the recommended content may affect other, second-order components of demand, such as the trust they place in various types of news sources (29, 39, 40). This builds on previous work that found one-sided media consumption drives suspicion and distrust of the news media more generally (39–41). One-sided media consumption can eventually lead to even more worrisome outcomes, such as people relying less on new information and lowering their opinion of out-party politicians (26).

Because slanted videos are believed to have a persuasive effect, a third testable hypothesis is that randomized assignment to different algorithms will indirectly cause changes in users' specific attitudes on the topic of the videos—in our studies, gun control or minimum wage. Such effects could unfold through a variety of mechanisms, including framing of the issue (42), cue-taking (43), or new policy-relevant facts (44). Finally, we examine whether manipulating the recommendation algorithm has a more general second-order impact on affective polarization, rather than just issue-specific polarization. This is because prior work has shown traditional media's role in affective polarization (45)—emotional attachments to one's partisan ingroup, as well as distaste for the outgroup—which may be heightened by the slanted and inflammatory content that recommendation systems often suggest.

While existing claims imply these four testable hypotheses, a pressing claim is whether we would expect those effects to appear in a short-exposure experiment. We describe our study as short-exposure because it is not positioned to identify effects that might come from prolonged exposure of watching videos over many months or years. Aside from innovative encouragement experiments (28) which *encourage*, but do not *force*, participants to consume media outside of real-world settings, the majority of the experimental literature is based on short exposures. We conducted an expansive review of all *PNAS* studies in the last decade that met two criteria: they (1) presented a treatment (e.g., video clips, reading materials, or images) in a human-subjects experiment; and (2) examined participants' decisions and opinions following the intervention (see SI 15 for details). The median length of exposure to persuasion stimuli was 101 seconds. Many of these studies deliver quite strong effects such as Tappin et. al. (46), which examined the persuasiveness of microtargeted videos on policy attitudes. The average duration of their video stimuli was 52 seconds and their maximum exposure was 70 seconds. Like many other studies with short exposure to media stimuli, they demonstrate that these interventions can indeed have significant effects on deeply entrenched political attitudes (they studied immigration and welfare policy, which we view as roughly comparable to our gun-rights issue and far more entrenched than our minimum-wage issue). At an average of 18 minutes, our "short" exposure is an order of magnitude longer than these prior studies, providing a credible empirical evaluation of the first-order implications of the existing narrative on algorithmic polarization.

2 Experimental Design

To address the challenges of research in this setting, we developed a new experimental design that randomly manipulates algorithmic video recommendations through a custom-built, YouTube-like platform. We provide brief details below for Studies 1–3, deferring additional details to the Materials & Methods Section.

We gathered real YouTube videos on two policy issues (more on the issues below), collect

actual YouTube recommendations for these videos, experimentally manipulated these recommendations to be slanted or balanced, and then sequentially presented the videos and their following recommendations to experimental subjects in a realistic choice environment. We continuously monitored how users chose among recommended videos, whether they skipped forward or watched videos in their entirety, and how they otherwise positively or negatively interacted with the video. To test whether recommendation algorithms had an effect on attitudes, subjects were surveyed in two waves occurring roughly one week before and immediately after using the video platform.[1]

Our platform and its recommendations were designed to closely approximate both the viewing experience and the algorithmic recommendations of YouTube. Upon entrance to the platform, respondents were shown a "seed" video on a topical policy issue: on gun control in Study 1, or on the minimum wage in Studies 2 and 3. At the conclusion of the video, respondents were presented with four recommended videos to watch next, drawn from the actual YouTube recommendation network. Respondents selected another video from the recommendations, watched that video, and then were presented with another set of recommendations. Each respondent watched up to five videos, with four opportunities to choose among different sets of recommendations. Respondents were required to watch at least 30 seconds of each video before they were allowed to skip ahead to the end of the video. Throughout their time on the video platform, respondents could interact with the platform by indicating whether they liked or disliked the video they were watching, and they could save the current or recommended videos to watch later.

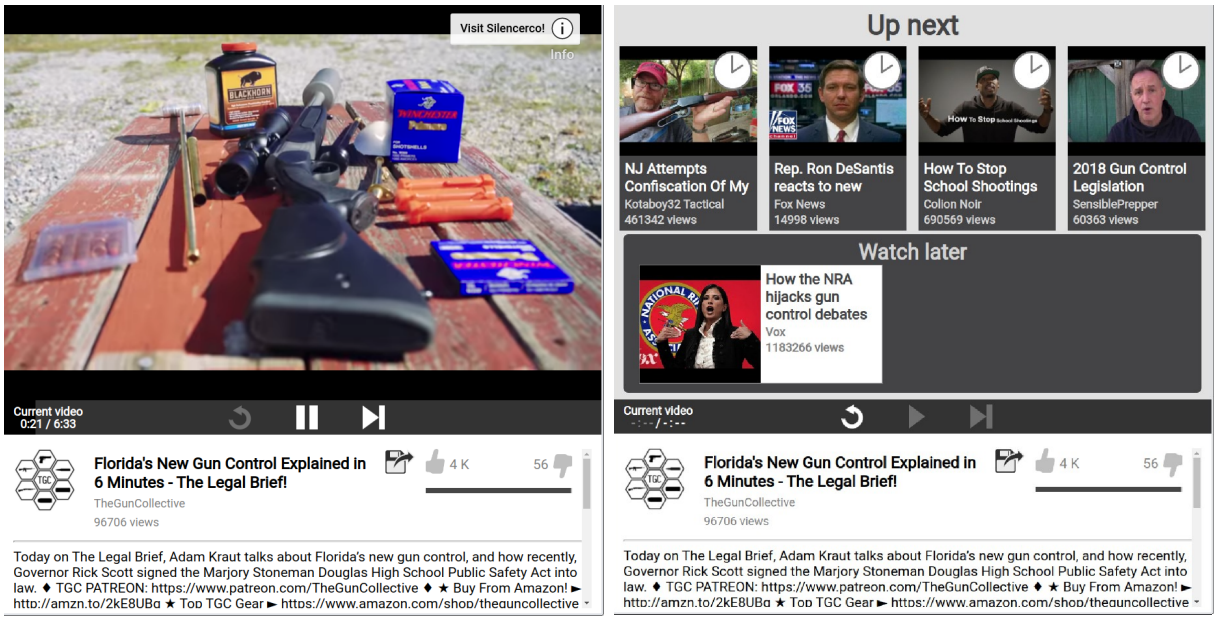

Figure 2: **Video platform interface and recommendations.** The left panel shows the video-watching interface for an example video in Study 1, and the right panel shows an example of recommendations that were presented to respondents after finishing this video.

[1]Studies 1 and 2 had a third, follow-up survey wave occurring approximately one week after the experimental video-platform session.

Videos on the selected policy topics, along with their recommendations, were identified via the YouTube API, validated, and classified for valence. Our experiments manipulated both the slant of the initial "seed" video (liberal or conservative) and the mix of recommendations presented to subjects after they watched each video (balanced or slanted in the direction of the previous video), for a total of four conditions. Based on the pre-treatment political attitudes, respondents were divided into liberal, moderate, and conservative terciles with the ideologues (liberals and conservatives) only being shown the like-minded seed. After watching or skipping each video, respondents were presented with four recommended videos that were either "balanced" (two recommendations matching the ideological direction of the previous videos and two from the opposite perspective) or "slanted" (three matching and one opposing).

Our main outcome, policy attitude, is measured with an index formed from responses to five (Study 1) or eight (Studies 2–4) survey questions on the relevant policy, which we averaged into a measure that ranged from 0 (most liberal) to 1 (most conservative). We also include measures of media-trust, behavior on platform (interactions with the video platform) and affective polarization. We analyze post-treatment attitudes using regressions that control for a set of attitudes and demographic characteristics that were measured pre-treatment per our pre-analysis plan. Our main analyses examine the effect of the slanted recommendation algorithm (vs. the balanced algorithm) on our outcomes.

We recruited large and diverse samples (within the confines of modern survey sampling) across all studies using MTurk via CloudResearch and YouGov (Studies 1–3 include approximately 2,500 participants each). Study 1 was started in June 2021, Studies 2–3 were started in April 2022, and Study 4 was started in May 2024.

2.1 Policy Issues

In order to have a well-defined measure of video valence, extremity, and policy attitude, we limit our studies to one policy issue each. Study 1 covers gun control and Studies 2–4 covers minimum wage. This naturally induces a limitation that we can only speak directly to these topic areas. The claims in the polarization literature have largely not been qualified by topic. Indeed, Tufekci (2) argues that (at least in 2018) YouTube was "radicaliz[ing] billions of people" across countless issue areas—vaccines, diet, nutrition, exercise, gun policy, white supremacy, 9/11, and more. In our choice of issues we had to trade-off between issues raised in the polarization literature but where there would be serious ethical implications (e.g. white nationalism, pro-ISIS videos, and vaccine skepticism) and more common policy topics. We chose gun control because it connects with some of the most visceral examples of rabbit holes (e.g. conspiracies in school shootings). We chose minimum wage to find a case that was high profile, but but less divided along partisan lines. In the qualitative case-selection language, the strong and weak partisan divisions on these topics of gun control and minimum wage policy, respectively, mean they could perhaps be regarded as "least likely" and "most likely" issues for persuasion effects (at least among high-profile topics). Regardless, we emphasize that our evidence is specific to the gun control and minimum wage debate; it could be that effects exist in other topic areas.

2.2 "First Impressions" Experiment

Our design changes the balance of recommendations and allows users to choose videos in an ecologically valid way—by observing the thumbnail, channel name, and view count. This does not ensure that they are able to select content based on valence if they are not able to perceive the valence from the thumbnail. In an experiment reported in SI 11, we use the video recommendation interface to collect participant evaluations of the partisan leaning of a video. Our results show that participants have a higher-than-chance ability of discerning the political leaning of a video based on the recommendation page. However, there is substantial heterogeneity across topics and video-ideology, with conservative minimum wage videos being particularly easy to guess and liberal gun control videos being particularly challenging. We also use a computational baseline (GPT-4V) to assess how much visual information is present even if participants don't discern it. We find that GPT-4V is able to achieve 84% accuracy overall (91% for minimum wage and 69% for gun control)—substantially exceeding human performance.

2.3 "Rabbit Hole" Experiment (Study 4)

Studies 1–3 take the existing YouTube algorithm as a starting point and artificially "slant" it to boost prevalence of similar ideological position (magnifying the "filter bubble" phenomenon). Our real-world recommendation data suggests that this captures real-world patterns on YouTube well. We analyzed video transcripts to measure their ideological extremity and found that recommendations did not get increasingly extreme—in fact, we found that extreme videos led to recommendations that were slightly more moderate (see SI 14). Consequently, the experiments derived from this real-world data also do not get more extreme; in other words, Studies 1–3 capture the "filter bubble" phenomenon but not the "rabbit hole." This is consistent with observational work on YouTube using sock-puppets by Haroon et al. (11) who found only "substantively small" extremity increases over video sequences.

For Study 4, we developed an experiment that would artificially intensify the extremity of videos to assess the effects that viewing such sequences might have on minimum wage political opinions. In this experiment, we again divided participants into three groups (conservatives, liberals and moderates). Conservatives and liberals were assigned to an ideologically aligned sequence that was constructed to be either constant in extremity or increasing. Moderates were assigned one of the four types of sequences. In contrast with Studies 1–3, this study was conducted entirely in one wave (asking opinions before and after the video viewing) and did not involve choosing videos to view. In this sense, it provided a substantially stronger, but less ecologically valid, treatment.

3 Results

We first present side-by-side results from Studies 1–3 to permit comparisons across issue areas and sampling frames. Our first two sets of results examine the "algorithmic effect of being assigned to an ideologically slanted recommendation system, compared to a balanced one (corresponding theoretically to a "filter bubble effect"). We begin with algorithmic effects on liberal and conservative "ideologue" respondents in Section 3.1 before proceeding

to algorithmic effects on "moderate" respondents in Section 3.2. In Section 3.3, we present a second set of results that examine the effect of assigning moderate respondents to a liberal seed video, compared to a conservative one, when users are subsequently allowed to freely navigate the recommendation system. Finally in Section 3.4 we present the results from the "rabbit hole" study.

Each section below presents estimated effects across a variety of outcome measures. We group these outcomes into four families, based on the hypotheses described in Section 1: (1) demand-side outcomes relating to media consumption and user interaction with the platform; (2) demand-side outcomes about trust in media; (3) attitudinal outcomes measuring issue-specific polarization; and (4) attitudinal outcomes relating to general affective polarization. Throughout, all hypothesis tests reflect multiple-testing corrections as described in the Materials & Methods section. Plots show 90% and 95% confidence intervals with robust standard errors; we use color to denote the results of hypothesis testing and emphasize that readers should only interpret results that remain significant after multiple-testing correction.

3.1 Algorithmic Effects Among Ideologue Respondents

We first examine these algorithm-driven effects among ideologues (i.e. those in the lowest and highest terciles of pre-treatment policy attitudes). Figure 3 shows the effects of a more extreme recommendation system for liberal respondents (on the left) and conservative respondents (on the right). Each symbol denotes one of our three studies: filled circles are estimates from our first study, on gun policy; triangles are estimates from the second study, on minimum wage policy with a Mechanical Turk sample; and diamonds are estimates from our third study on minimum wage policy with a YouGov sample.

The top panel in both sets of results shows the effects on respondents' platform interactions. For both sets of respondents, we find that a more extreme recommendation system caused respondents to choose more videos from the same ideological slant as the video they had just watched, relative to a balanced set of recommendation videos. The liberal fraction of videos chosen by liberal respondents assigned to the slanted (3/1) algorithm was 5 percentage points higher than liberal respondents assigned to the balanced (2/2) algorithm. Similarly, the liberal fraction of videos chosen by conservative respondents assigned to the slanted algorithm was 13 percentage points lower than those receiving balanced recommendations. This is consistent with the increased availability of videos: if respondents were choosing randomly, it would be about 12 percentage points higher in the ideological direction of the seed video (which, by design, was matched to the ideological orientation of liberal and conservative respondents).

The lower panels of Figure 3 respectively show the effects of the recommendation slant on policy attitudes, media trust, and affective polarization. We find few significant effects on any other outcome among ideologues. The one exception is the effect on policy attitudes in Study 3 among conservatives. In this study, respondents assigned to view more slanted recommendation videos reported post-treatment attitudes that were slightly more conservative (0.03 units on a 0–1 policy index) than respondents assigned to view balanced recommendation videos. Importantly, the estimated effects are quite small. For instance, the upper limit of this 95% confidence interval for the effect of the recommendation system on conservative respondents in Study 1 is 0.04 units on this 0–1 policy index, equivalent to

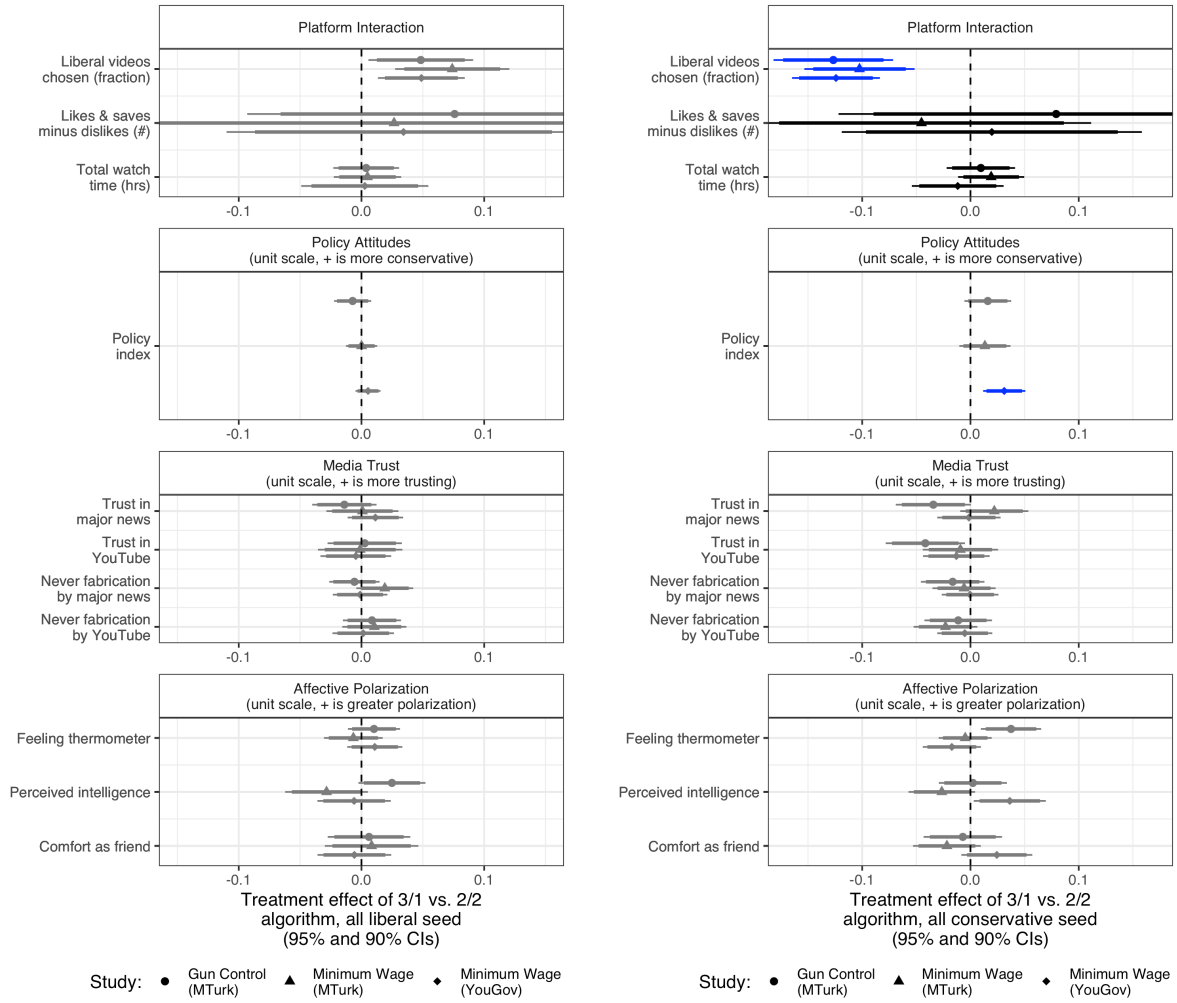

Figure 3: **Effects of recommendation algorithm among ideologues.** Both panels show the results of more algorithmic recommendation slant (vs. balance) on behaviors and attitudes among ideologues (those in the first and third tercile of pre-treatment policy attitudes). The left panel shows effects among more liberal (i.e. lowest tercile) respondents, and the right panel shows effects among more conservative (i.e. highest tercile) respondents. Grey points and error bars represent estimated effects that are not statistically significant after implementing multiple testing corrections, while blue points and error bars represent those effects that are still statistically significant after multiple testing corrections.

16% of the respondents moving one level up on each of the index's five-point components.[2]

3.2 Algorithmic Effects Among Moderate Respondents

Our results examining the effects of recommendation algorithms among moderates appear similar. Again, the more slanted (3/1) recommendations appear to influence respondents'

[2]Because we find no substantial effects on attitudes in the wave 2 data from studies 1 and 2, we did not analyze the wave 3 data.

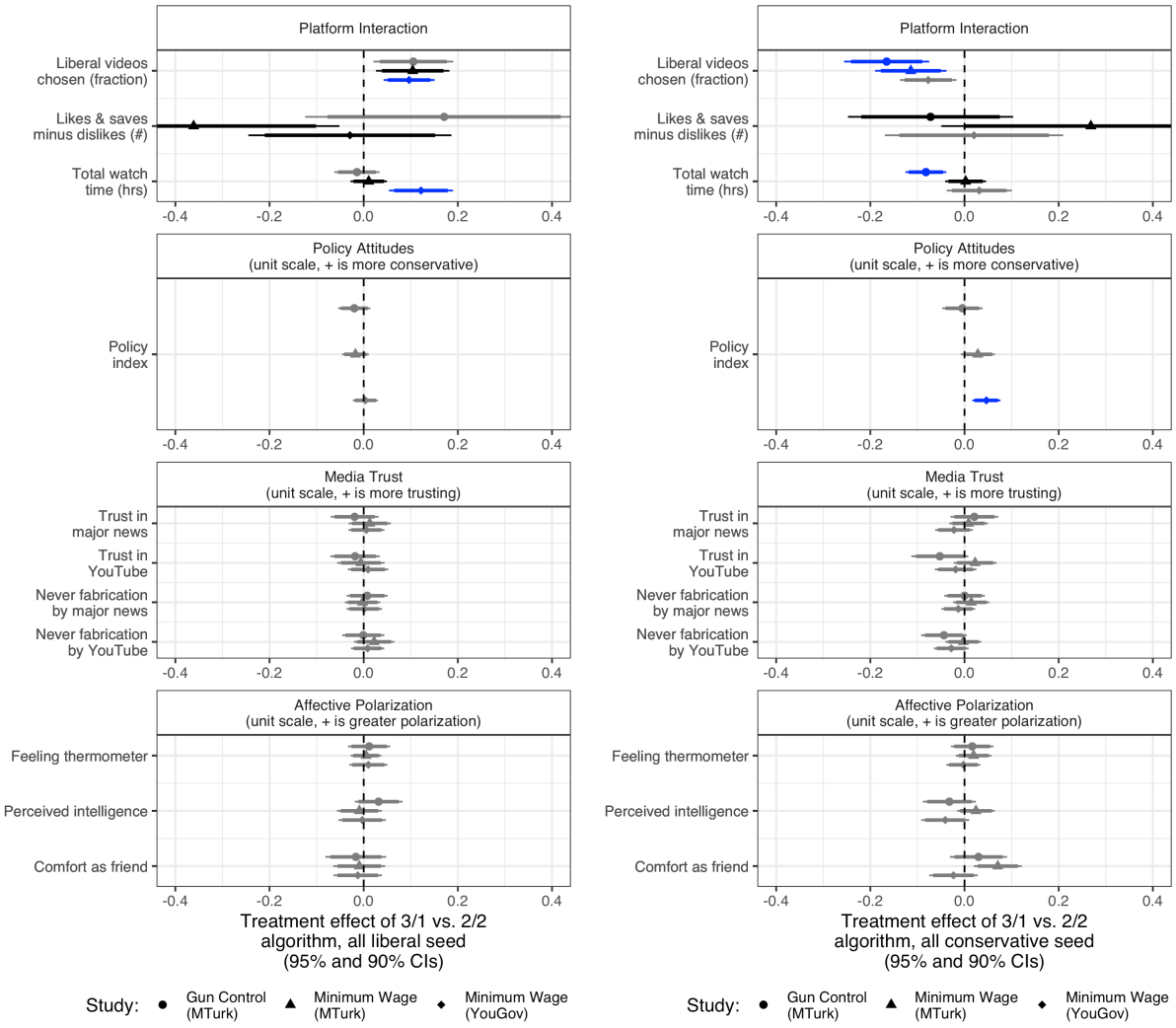

Figure 4: **Effects of recommendation algorithm among moderates.** Both panels shows the results of more algorithmic recommendation extremity (vs. balance) on behaviors and attitudes among moderates (those in the middle tercile of pre-treatment policy attitudes). The left panel shows effects among those respondents assigned to a liberal (i.e. pro-gun control or pro-minimum wage) seed video, and the right panel shows effects among respondents assigned to a conservative (i.e. anti-gun control or anti-minimum wage) seed video. Grey points and error bars represent estimated effects that are not statistically significant after implementing multiple testing corrections, while blue points and error bars represent those effects that are still statistically significant after multiple testing corrections.

choices of videos, compared to the balanced (2/2) ones, and in two instances significantly affected the amount of time respondents spent on the platform. Figure 4 shows the effect of the more slanted recommendation system for respondents assigned to the liberal seed videos (on the left) and the conservative seed videos (on the right). As in the previous section, respondents assigned to the slanted algorithm chose to watch a higher proportion of videos that resembled the seed video. In other words, respondents assigned to a liberal seed and

slanted recommendations were more likely to choose liberal videos, compared to other liberal-seed respondents who received balanced recommendations. Similarly, respondents assigned to a conservative seed and slanted recommendations chose liberal videos at a lower rate, compared to other conservative-seed respondents with balanced recommendations. Among moderates assigned a liberal seed in Study 3, being assigned the slanted recommendations appears to have increased the total time respondents spent on the platform by 7.3 minutes on average, while moderates assigned a conservative seed video in Study 1 with slanted recommendations appear to have spent 4.9 minutes *less* time watching videos on average than those assigned a balanced set of recommendations. These effects are quite large given the average watch time of 18 minutes. This may be because the sample skews liberal overall, meaning that the "moderate" tercile is still somewhat liberal. In this case, being forced to watch a conservative video and then being presented with three more conservative videos in the first set of recommendations could plausibly decrease satisfaction and time spent on the platform, despite subsequent freedom of choice.

Despite these large effects on media consumption, the slant in recommendations appears to affect political attitudes only minimally among moderates. Nearly all the effects of the recommendation algorithm on policy attitudes, media trust, and affective polarization appear statistically indistinguishable from zero. The one exception is again in Study 3, where it appears that moderate respondents assigned the conservative seed video and slanted recommendations reported opinions that were slightly more conservative (0.05 units on a 0–1 scale) than respondents assigned to balanced recommendations. The small size of these estimates and their relatively narrow confidence intervals suggest that the general lack of statistical significance is not simply due to small-sample noise, but rather a genuinely small or nonexistent short-term attitude change. That is, we can rule out anything greater than these quite-modest immediate effects on policy attitudes caused by more extreme recommendation algorithms.

3.3 Forced-Exposure Effects Among Moderate Respondents

We assess the effects of the randomized *seed video* among moderates. These effects most closely mirror the effects of a traditional randomized forced-exposure study, as they measure the effects of being assigned a conservative rather than a liberal initial video—often referred to as attitudinal persuasion. However, our results differ in that after this forced exposure, we allow users to freely interact with the platform and choose which videos to consume. The results of these analyses are presented in Figure 5, showing the difference in outcomes between those respondents assigned to a conservative seed video compared to those assigned to a liberal seed video, among those respondents who received recommendations in a more slanted mix (3/1, on the left) or a more balanced mix (2/2, on the right). In the slanted recommendation system, being assigned to a conservative video led moderate respondents to choose a much lower fraction of subsequent liberal videos than those assigned to a liberal video, as the top left panel shows. This effect disappears when moderate respondents are assigned to the balanced recommendations: watching a conservative seed video made respondents no more or less likely to choose liberal videos from the recommendations presented to them, as shown in the top right panel.

The effects of the assigned seed video on moderates' attitudes, presented in the lower

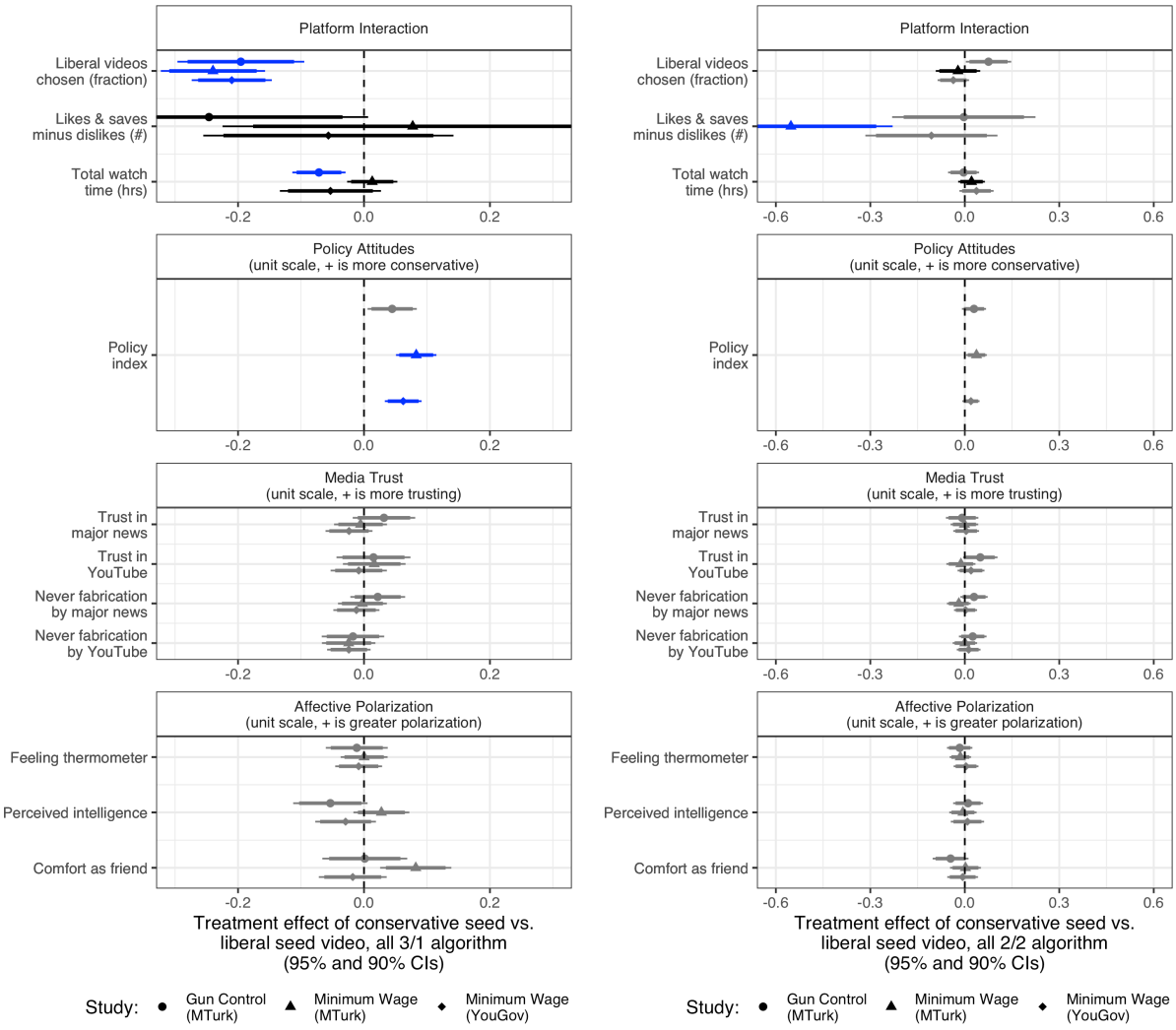

Figure 5: **Effects of seed video slant among moderates.** Both panels shows the results of a more conservative seed video on behaviors and attitudes among moderates (those in the middle tercile of pre-treatment attitudes). The left panel shows effects among those respondents assigned to a 3/1 recommendation algorithm, and the right panel shows effects among respondents assigned to a 2/2 recommendation algorithm. Grey points and error bars represent estimated effects that are not statistically significant after implementing multiple testing corrections, while blue points and error bars represent those effects that are still statistically significant after multiple testing corrections.

panels of Figure 5, suggest slight persuasion effects. Respondents assigned to the slanted recommendations who were assigned a conservative seed video reported slightly more conservative policy attitudes than those who were assigned a liberal seed video, as shown in the second panel on the left side of Figure 5. These effects, again, are muted among those respondents who were assigned to the balanced recommendations. These respondents reported policy attitudes that were not discernibly different when assigned to either the conservative or liberal seed video. We observed no other effects on attitudes that were statistically dis-

tinguishable from the null hypothesis. That there are some detectable effects on policy from the forced choice assignment gives us confidence that the algorithmic assignment would be able to detect an effect if one existed.

3.4 "Rabbit Hole" Effects (Study 4)

Finally, we present the results of Study 4 which constructed artificial sequences on the minimum wage which were increasing in extremity or held constant in order to test the "rabbit hole" hypothesis. This design is distinct from Studies 1–3 in that it is a single-wave study and respondents are assigned to a fixed sequence of videos (they do not choose recommendations, comparable to the YouTube "Autoplay" experience or the "YouTube Shorts" interface). The results of these analyses are presented in Figure 6, showing the effects on policy attitudes for different causal contrasts. Assignment to the increasing (vs. constant) sequences appears to have no effect on ideologues or moderates. The only discernible effect is a modest effect of the seed assignment for moderates, consistent with the results in the previous section. This suggests that any algorithmic effect for "rabbit holes" that exists is likely far smaller than simply being assigned to watch conservative or liberal video sequences.

Despite these null short-term results on our overall attitudinal index, it remains possible that recommendation algorithms expose viewers to new ways of understanding or interpreting a policy issue that might eventually lead to long-term persuasion. An exploratory reanalysis of Studies 1–3 proposed by a reviewer suggested that this might be the case: when extracting more-conceptual survey questions from the overall index to analyze individually, we found patterns that were consistent with algorithmic effects on conservative participants' understanding of minimum-wage issues. In Study 4, we therefore sought to assess whether algorithmic interventions exposed viewers to unfamiliar information by asking participants about whether they learned anything from watching the videos. Almost 90% of participants reported learning something new. An analysis of the open-ended responses suggests that this learning was diverse including general knowledge, impact on businesses/the economy/poverty, automation, wage stagnation and political dynamics. We did not find evidence that algorithmic interventions affected the amount of self-reported learning. For details on the exploratory reanalysis of Studies 1–3, see SI 9; for learning in Study 4, see SI 13.

4 Discussion and Conclusion

In her 2018 *New York Times* opinion piece, Zeynep Tufekci provides one of the clearest articulations of YouTube's role as a radicalizing force in American politics. She paints a picture of YouTube's ability to recommend users ever more extreme views of what they are already watching—Donald Trump rallies lead to white supremacist rants, Hillary Clinton videos lead to leftist conspiracies, and even jogging leads to ultramarathons. She writes,

> It seems as if you are never "hard core" enough for YouTube's recommendation algorithm. It promotes, recommends and disseminates videos in a manner that appears to constantly up the stakes. Given its billion or so users, YouTube may be one of the most powerful radicalizing instruments of the 21st century. (2)

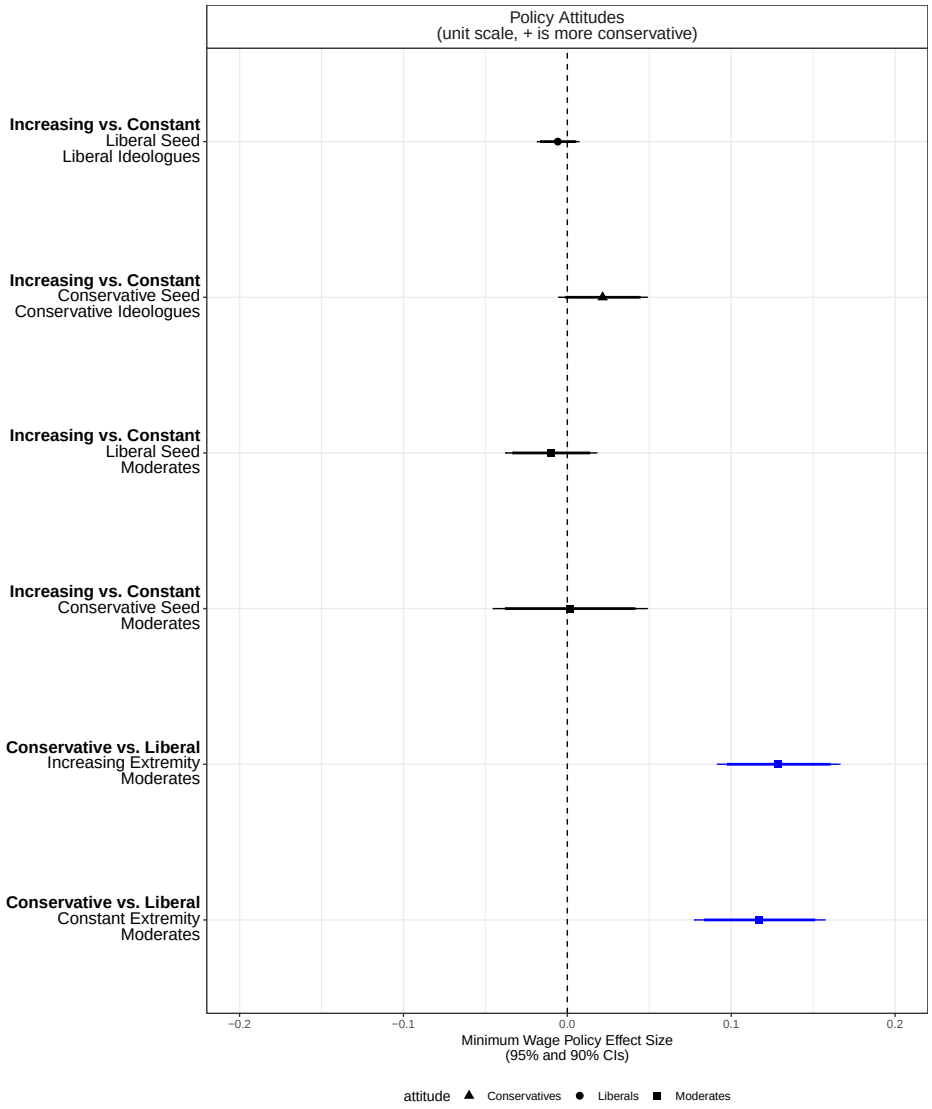

Figure 6: **Effects of "Rabbit Hole" treatment on policy attitudes.** This panel shows the causal contrasts in Study 4. The top two rows show the effects of being assigned to an increasing extremity sequence on ideologues (those in the lowest and highest tercile of pre-treatment attitudes). The next two rows show the effects of being assigned to increasing sequences amongst moderates within each assigned seed. The last two rows show the only significant effects which are the effects of the seed conservative seed assignment among moderates given that the sequences are increasing or constant. Grey points and error bars represent estimated effects that are not statistically significant after implementing multiple testing corrections, while blue points and error bars represent those effects that are still statistically significant after multiple testing corrections.

The implication of this argument—and the assumption of many scientific studies that followed—is not only that YouTube's recommendation algorithm presents more extreme content to consumers, but that the presentation of this extreme content also *changes their opinions*

and behaviors. This is a worrying claim that applies not only to YouTube, but to any of the increasingly numerous online systems that rely on similar recommendation algorithms and, it is claimed, all pose similar potential risks to a democratic society (47). The weaker claim is that recommendation algorithms induce "filter bubbles" which could produce similar types of opinion changes. Yet if these claims were true, one would imagine that users in our study who were recommended gun-rights videos would have shifted their attitudes substantially toward support of gun rights, and those who were recommended gun-control videos would have moved substantially toward support for gun control.

Of course, in many ways the situation we can test with our experimental design is not the entirety of the story that Tufekci (2) and others describe. It remains possible that months- or years-long exposure to personalized recommendation systems could lead to the conjectured radicalization. Work by Centola (48) has shown that repeated exposure is important to behavior contagions. It also remains possible that there are heterogeneous effects—though we failed to detect heterogeneity in exploratory analyses examining the moderating role of age, political interest, YouTube consumption, and college education.[3] Finally, we cannot rule out the existence of a small—but highly susceptible—-population that cannot be detected with our sample sizes.

Nevertheless, by providing real subjects with naturalistic choices over the media they consume, based on actual recommendations from YouTube in nearly 9000-person randomized controlled trials, our study arguably represents the most credible test of the phenomenon to date. Widespread discussion of YouTube's radicalizing effects are difficult to reconcile with the fact that we fail to detect consistent evidence of algorithmic polarization in this experiment of either the "filter bubble" or "rabbit hole" form. Notably, the narrow confidence intervals on attitudinal effects show that even the maximum effect sizes consistent with our algorithm system results are small, relative to recent experiments on media persuasion with approximately comparable stimuli and our own seed effect estimates. Experiments that allow for respondent choice in videos may tend to have smaller persuasive effects than in traditional forced-choice settings, in part for the simple reason that allowing realistic choice in media consumption leads to fewer users consuming the opposing viewpoints that could persuade them. Our results also align with recent work showing the limits of selective exposure in online media consumption (49, 50), which implies that only a limited set of people will consume highly imbalanced media when given the opportunity. Our results with forced exposure in Study 4 provide larger seed effects, but still no system effects.

Although our study does not provide convincing evidence that the recommendation-system manipulation affected attitudes, we do observe changes in behavior: the balance of recommended videos appears to influence subsequent video selection among moderates and (depending on the seed) total watch time on the platform. Potential *decreases* in platform watch time as a result of unwanted or unexpected content exemplify the kind of problem that recommendation algorithms are likely intended to solve. This kind of divergence between attitudinal and behavioral effects on social platforms is a potential area for future research. One shortcoming that our study shares with nearly all research on YouTube is that, by taking existing platform recommendations as a starting point, we hold the set of potential videos that *could* be shown—the supply—as largely fixed, apart from the experimental per-

[3]The only moderating factor identified in these exploratory analyses was gender.

turbations in exposure that we induce. Yet like users' behavior, the production of content is dynamic and subject to incentives. As Munger elaborates (51, 52), the interplay of supply and demand may be an underappreciated factor shaping the choices available to users as they experience the platform, regardless of the specifics of any recommendation system. A full understanding of the impact of streaming video platforms such as YouTube requires simultaneous consideration of interacting and self-reinforcing processes in the supply, demand, and effects of media consumption.

Finally, while our experiments cannot rule out the possibility of some level of radicalization on some subset of the population on YouTube, it provides some guidance on the complexity and scale of an experiment that *would* be necessary to detect such an effect. Our multiple large-scale survey samples appear to approach the limit of the number of experimental subjects that can currently be recruited for studies as time-intensive as the ones presented here—suggesting that if algorithmic polarization has smaller effects than we were powered to detect, it may be difficult to ever identify them under controlled conditions. Sobering though this conclusion may be, our goal throughout the design and execution of this study has been to maximize our chances of observing a true effect despite hard budgetary constraints. If radicalization were possible, our choice of policy areas—both of which were selected for their low to moderate levels of preexisting polarization—should have enabled us to observe attitudinal change. Similarly, our selection of real-world video recommendations from YouTube represents the most realistic attempt that we know of to replicate the slanted recommendation algorithms of social media platforms. The results from our four studies thus collectively suggest that extreme content served by algorithmic recommendation systems has a limited radicalizing influence on political attitudes and behavior, if this influence even exists.

5 Materials and Methods

This study has been approved by Princeton University IRB (#12989) and the other institutions via Smart IRB (ID: 3931). All participants consented to the experiment at the before the initial survey, with consent materials provided in SI 1. All replication data and code will be made available in Dataverse on publication.

5.1 Collecting the Videos

We base our experiment recommendations on real recommendations from the YouTube API. To obtain a videos recommendations we start with the related videos that the YouTube API provides for each video. From these we selected the subset of recommendations that were on the same policy topic and took either a liberal or conservative stance on the policy, as determined by a combination of hand coding and supervised machine learning. For both topics, we first conducted a round of coarse screening for topicality. For gun control, we used crowd workers on MTurk to create a hand-labeled training set for a cross-validated support vector machine, which was then used to select videos for inclusion. For minimum wage, we used crowdsourcing to classify all videos. Inter-rater agreement ranged from 80% to 85% across multiple rounds of classification. The authors then conducted a final round of

manual validation. We arrived at our 3/1 and 2/2 experimental proportions after analyzing YouTube recommendations on the gun-control topic and finding that, among videos with a discernible ideological direction, roughly 60% of recommendations of the same ideology. The 3/1 and 2/2 experimental conditions thus bracket the average real-world proportions, increasing realism. SI 2 contains more details about construction of the recommendation trees, hand-coding, classification, and validation.

Two aspects of this process deserve additional discussion. First, to our knowledge, there is no formal documentation explaining the relationship between the recommendations obtained from the YouTube API and those that are shown to actual users in the web or app interface. To investigate this, we conducted a validation exercise comparing API recommendations to those presented on the YouTube web interface in actual browser sessions to an anonymous user, both starting from the same video. Aside from some instances in which the web interface deviated to off-topic recommendations that would have been eliminated by our trimming procedure, the two sets of recommendations are largely the same. See the supplement for more details. A second point is that, like most prominent audits of the YouTube recommendation algorithm (e.g., 6, 7), we do not observe personalization based on a user's watch histories or past engagement. This is an important scope condition, as Haroon et. al. (11) find (modestly) increasingly ideological recommendations for automated sock-puppet accounts. With that said, our design allows us to experimentally manipulate a recommendation algorithm that is actually deployed in the real world—the generic YouTube algorithm that makes suggestions *based on the currently selected video*—allowing us to target a well-defined estimand that remains informative for policy questions about algorithmic recommendations, even if we cannot study the personalization process directly (53). Specifically, the videos recommended by our design remain relevant as long as personalization does not fundamentally change the type of recommendations made, but rather only shifts their relative rankings.

5.2 Additional Study Details

Studies 1 and 2 respectively recruited 2,583 and 2,442 respondents on MTurk via CloudResearch, Study 3 drew 2,826 respondents from YouGov, and Study 4 recruited 1,032 respondents on MTurk via CloudResearch. In recruiting our experimental subjects, we used approval requirement qualifications and attempted to recruit a balanced set of political opinions on Mechanical Turk. We had difficulty recruiting respondents that fit these criteria, suggesting that we might be reaching the upper limit of how many people can be recruited on Mechanical Turk for such time-intensive studies. Our sample from a larger and more expensive subject pool, YouGov, ran into similar issues, suggesting that there are limits to the subject pool available for interrogating these questions more broadly.

Our main policy attitude outcome is an index formed from responses to five (Study 1) or eight (Studies 2 and 3) survey questions on the relevant policy, which we averaged into a measure that ranged from 0 (most liberal) to 1 (most conservative). These scales were quite reliable: for Study 1, $\alpha = 0.87$; study 2, $\alpha = 0.94$; and study 3, $\alpha = 0.94$. We also pre-registered an exploratory factor analysis with varimax rotation for these questions. The proportion of variance explained by a single dimension is 0.68, 0.72, and 0.73, respectively. Refer to SI 8 for question wordings. Our media-trust questions were taken from standard

batteries used in research on political communication (e.g. 3), while our measures of affective polarization were similarly taken from validated measures of out-party animosity (e.g. 54). Following our pre-registered plan, we assessed the effects of the video recommendation algorithm by comparing the post-treatment attitudes of respondents in different experimental conditions, based on the same liberal-ideologue, moderate, and conservative-ideologue subgroups used in treatment assignment. We analyze post-treatment attitudes using regressions that control for a set of attitudes and demographic characteristics that were measured pre-treatment per our pre-analysis plan. Our main analyses examine the effect of the slanted recommendation algorithm (vs. the balanced algorithm) on respondents' video choices; their platform interactions; and their survey-reported policy attitudes, media trust, and affective polarization.

Specifically, in the policy-attitude, media-trust, and affective-polarization analyses, we control for pre-treatment versions of all outcomes in the hypothesis family, defined below. In the platform-interaction analyses, we control for age, gender, political interest, YouTube usage frequency, number of favorite YouTube channels, whether popular YouTube channels are followed, text/video media consumption preference, a self-reported gun enthusiasm index, and perceived importance of the gun policy issue. We pre-registered the use of the Lin (55) estimator (using demeaned controls, all interacted with treatment) but found this to produce an infeasible number of parameters. As a result, we instead use controls in an additive (non-interacted) regression with robust standard errors. These results are substantively similar to the unadjusted results.

Study 1 and the MTurk sample for Study 2 contained an additional "pure control" condition that involved watching no videos. Per our pre-registration, we committed to only using this control condition if there was a newsworthy event related to the policy issue under study, which did not occur during either study. We conducted stratified randomization to these experimental conditions based on respondents' pre-treatment political attitudes on the policy subject. Respondents in the most liberal tercile ("liberal ideologues") were only shown a liberal seed video, meaning that the only randomization for these subjects was between the balanced and slanted recommendation algorithm. This avoided forcibly exposing liberal participants to conservative viewpoints that they did not voluntarily consume, improving the realism of the study. Similarly, "conservative ideologues" initially in the most conservative tercile were only exposed to conservative seed videos. "Moderate" respondents, defined as the middle tercile of pre-treatment attitudes, were randomly presented with either liberal or conservative seed videos.

5.3 Multiple Testing Correction

To account for the four families of outcomes, we conduct multiple-testing corrections following our pre-analysis plan and the recommendations of the literature (56, 57) to control the false discovery rate while properly accounting for the nested nature of the tests. We examine three layers of hypotheses: (1) whether the experiment had any effect on a family of outcomes, broadly construed; (2) which subgroup and treatment contrast generates the effect; and (3) the specific outcome on which the effect manifests. The correction proceeds as follows. Within hypothesis families that survive the first-stage assessment of overall significance, we proceed to disaggregated examination of individual hypotheses. The initial

"layer-1" family-level filtering is conducted using Simes' method (58) to combine layer-2 p-values (defined below) across the six treatment contrasts. This tests the intersection null that no version of the treatment had any effect on any outcome in the family. Because four hypothesis families are tested, an additional Benjamini-Hochberg (BH) correction (59) is applied to the family's Simes p-value before interpreting the layer-1 results. We say that a family "survives" if its BH-corrected Simes p-value is less than 0.05. Within each hypothesis family and treatment contrast, layer-2 p-values are obtained by an F-test from a multiple-outcome regression, testing the null that the contrasted treatment groups are identical on all outcomes in the family. (If an F-test for joint significance cannot be computed for the multiple-outcome regression due to numerical issues in the variance-covariance matrix, we will fall back on an alternative, more conservative procedure in which we conduct separate regressions for each outcome and combine them with the Simes method.) We only seek to interpret a family's layer-2 p-values (which correspond to specific treatment contrasts) if the family survives layer-1 filtering (indicating that some effect exists for some treatment contrast). To interpret layer-2 p-values, we first apply a BH correction to the F-test results, then multiply by an additional inflation factor (one over the proportion of surviving families) to account for selection at layer 1. Finally, for treatment contrasts that survive layer-2 filtering, we examine which specific outcomes in the family are affected. These layer-3 p-values are obtained by disaggregating the previous analysis into single-outcome regressions. As before, a BH correction is applied to account for the fact that multiple outcomes are evaluated; in addition, inflation factors for layer-1 and layer-2 selection are also applied.

5.4 Measuring extremity

We tested numerous approaches for measuring the extremity of the content and ultimately determined that GPT annotations of political extremity—based on full transcript, channel name, and thumbnail image—appear to perform best when compared with human annotations. Specifically, we utilized OpenAI's GPT-4V, which can incorporate visual information from a video's preview thumbnail, which can often be informative. We evaluated a number of other approaches from recent work, including Lai et. al.'s (60) pretrained model using title/description metadata and Hosseinmardi et al. (12) expert classification of the channel/creator extremity. However, we found that these approaches performed poorly in recovering our own human labels, and we ultimately concluded that it was essential to incorporate the actual transcript of arguments made in the video. See additional details in SI 12.B.1.

Author Contributions: NL, XEH, YS, JdBK, AG, DK, and BMS developed the research designs. NL, AJBC, and CL collected web data. NL, XEH, and YS gathered, classified, and experimentally manipulated recommendation networks, with support from RM and DK. JdBK, AG, BMS, and RM designed and implemented the main survey, with support from MAB, AJB, XEH and YS. NL and DK designed the video platform, and NL collected platform browsing data. NL, XEH, YS, JdBK, AG, DK, and BMS designed and implemented analyses. JdBK, AG, and BMS drafted the original manuscript, and everyone edited it. Corresponding authors are listed alphabetically.

[1] CR Sunstein, *#Republic: Divided Democracy in the Age of Social Media*. (Princeton University Press), (2017).
[2] Z Tufekci, Youtube, the great radicalizer. *The New York Times* **10**, 2018 (2018).
[3] K Arceneaux, M Johnson, *Changing Minds or Changing Channels?: Partisan News in an Age of Choice*, Chicago Studies in American Politics. (University of Chicago Press), (2013).
[4] J de Benedictis-Kessner, MA Baum, AJ Berinsky, T Yamamoto, Persuading the enemy: Estimating the persuasive effects of partisan media with the preference-incorporating choice and assignment design. *Am. Polit. Sci. Rev.* **113**, 902–916 (2019).
[5] K Papadamou, et al., Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children. *Proc. Int. AAAI Conf. on Web Soc. Media* **14**, 522–533 (2020).
[6] M Ledwich, A Zaitsev, Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization. *arXiv:1912.11211 [cs]* (2019) arXiv: 1912.11211.
[7] MH Ribeiro, R Ottoni, R West, VA Almeida, W Meira Jr, Auditing radicalization pathways on youtube in *Proceedings of the 2020 conference on fairness, accountability, and transparency*. pp. 131–141 (2020).
[8] H Hosseinmardi, et al., Examining the consumption of radical content on youtube. *Proc. Natl. Acad. Sci.* **118** (2021).
[9] MA Brown, et al., Echo chambers, rabbit holes, and algorithmic bias: How youtube recommends content to real users. *Available at SSRN 4114905* (2022).
[10] AY Chen, B Nyhan, J Reifler, RE Robertson, C Wilson, Subscriptions and external links help drive resentful users to alternative and extremist youtube channels. *Sci. Adv.* **9**, eadd8080 (2023).
[11] M Haroon, et al., Auditing youtube's recommendation system for ideologically congenial, extreme, and problematic recommendations. *Proc. Natl. Acad. Sci.* **120**, e2213020120 (2023).
[12] H Hosseinmardi, et al., Causally estimating the effect of youtube's recommender system using counterfactual bots. *Proc. Natl. Acad. Sci.* **121**, e2313377121 (2024).
[13] E Pariser, *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. (Penguin), (2011) Google-Books-ID: wcaIrOI1YbQC.
[14] J Davidson, et al., The youtube video recommendation system in *Proceedings of the fourth ACM conference on Recommender systems*. pp. 293–296 (2010).
[15] P Covington, J Adams, E Sargin, Deep Neural Networks for YouTube Recommendations in *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16. (Association for Computing Machinery, New York, NY, USA), pp. 191–198 (2016).
[16] Z Zhao, et al., Recommending what video to watch next: a multitask ranking system in *Proceedings of the 13th ACM Conference on Recommender Systems*. pp. 43–51 (2019).
[17] AJB Chaney, BM Stewart, BE Engelhardt, How algorithmic confounding in recommendation systems increases homogeneity and decreases utility in *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18. (Association for Computing Machinery, New York, NY, USA), pp. 224–232 (2018).
[18] A Hannak, et al., Measuring personalization of web search in *Proceedings of the 22nd international conference on World Wide Web*. pp. 527–538 (2013).
[19] RE Robertson, et al., Users choose to engage with more partisan news than they are exposed to on google search. *Nature* **618**, 342–348 (2023).
[20] B Nyhan, et al., Like-minded sources on facebook are prevalent but not polarizing. *Nature* **620**, 137–144 (2023).
[21] AM Guess, et al., How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science* **381**, 398–404 (2023).
[22] J Nicas, How youtube drives people to the internet's darkest corners. *Wall Str. J.* **7** (2018).
[23] M Haroon, X Yu, E Menchen-Trevino, M Wojcieszak, Nudging the recommendation algorithm increases news consumption and diversity on youtube. (2023).
[24] CI Hovland, IL Janis, HH Kelley, *Communication and persuasion*. (Yale University Press), (1953).
[25] S Iyengar, DR Kinder, *News that matters: Television and American opinion*. (University of Chicago Press), (2010).
[26] M Levendusky, *How Partisan Media Polarize America*. (University of Chicago Press), (2013).
[27] M Prior, *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*. (Cambridge University Press), (2007).
[28] CA Bail, et al., Exposure to opposing views on social media can increase political polarization. *Proc. Natl. Acad. Sci.* **115**, 9216–9221 (2018).
[29] AM Guess, P Barberá, S Munzert, J Yang, The consequences of online partisan media. *Proc. Natl. Acad. Sci.* **118**, e2013464118 (2021).
[30] R Levy, Social media, news consumption, and polarization: Evidence from a field experiment. *Am. economic review* **111**, 831–870 (2021).
[31] S Iyengar, KS Hahn, Red media, blue media: Evidence of ideological selectivity in media use. *J. communication* **59**, 19–39 (2009).
[32] NJ Stroud, Media use and political predispositions: Revisiting the concept of selective exposure. *Polit. Behav.* **30**, 341–366 (2008).
[33] KH Jamieson, JN Cappella, *Echo chamber: Rush Limbaugh and the conservative media establishment*. (Oxford University Press), (2008).
[34] BJ Gaines, JH Kuklinski, PJ Quirk, B Peyton, J Verkuilen, Same Facts, Different Interpretations: Partisan Motivation and Opinion on Iraq. *J. Polit.* **69**, 957–974 (2007).
[35] D Knox, T Yamamoto, MA Baum, AJ Berinsky, Design, Identification, and Sensitivity Analysis for Patient Preference Trials. *J. Am. Stat. Assoc.* **114**, 1532–1546 (2019) Publisher: Taylor & Francis.
[36] K Munger, The limited value of non-replicable field experiments in contexts with low temporal validity. *Soc. Media+ Soc.* **5**, 2056305119859294 (2019).
[37] A Shaw, Social media, extremism, and radicalization. *Sci. Adv.* **9**, eadk2031 (2023).
[38] R Lewis, Alternative influence: Broadcasting the reactionary right on youtube. *Data & Soc.* **18** (2018).
[39] K Arceneaux, M Johnson, C Murphy, Polarized political communication, oppositional media hostility, and selective exposure. *The J. Polit.* **74**, 174–186 (2012).
[40] JM Ladd, *Why Americans Hate the News Media and How It Matters*. (Princeton University Press), (2012).
[41] K Arceneaux, M Johnson, How does media choice affect hostile media perceptions? evidence from participant preference experiments. *J. Exp. Polit. Sci.* **2**, 12–25 (2015).
[42] D Chong, JN Druckman, Framing theory. *Annu. Rev. Polit. Sci.* **10**, 103–126 (2007).
[43] JN Druckman, E Peterson, R Slothuus, How elite partisan polarization affects public opinion formation. *Am. Polit. Sci. Rev.* **107**, 57–79 (2013).
[44] JL Kalla, DE Broockman, "outside lobbying" over the airwaves: A randomized field experiment on televised issue ads. *Am. Polit. Sci. Rev.* **116**, 1126–1132 (2022).
[45] JN Druckman, S Gubitz, AM Lloyd, MS Levendusky, How incivility on partisan media (de) polarizes the electorate. *The J. Polit.* **81**, 291–295 (2019).
[46] BM Tappin, C Wittenberg, LB Hewitt, AJ Berinsky, DG Rand, Quantifying the potential persuasive returns to political microtargeting. *Proc. Natl. Acad. Sci.* **120**, e2216261120 (2023).
[47] C O'Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. (Crown), (2017).
[48] D Centola, *How behavior spreads: The science of complex contagions*. (Princeton University Press Princeton, NJ) Vol. 3, (2018).
[49] AM Guess, (almost) everything in moderation: New evidence on americans' online media diets. *Am. J. Polit. Sci.* (2021).
[50] C Wittenberg, MA Baum, AJ Berinsky, J de Benedictis-Kessner, T Yamamoto, Media measurement matters: Estimating the persuasive effects of partisan media with survey and behavioral data. *The J. Polit.* **85**, 1275–1290 (2023).
[51] K Munger, J Phillips, Right-wing youtube: A supply and demand perspective. *The Int. J. Press.* **27**, 186–219 (2022).
[52] K Munger, *The YouTube Apparatus*. (Cambridge University Press), (2024).
[53] I Lundberg, R Johnson, BM Stewart, What is your estimand? defining the target quantity connects statistical evidence to theory. *Am. Sociol. Rev.* **86**, 532–565 (2021).
[54] JN Druckman, MS Levendusky, What do we measure when we measure affective polarization? *Public Opin. Q.* **83**, 114–122 (2019).
[55] W Lin, Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *Annals Appl. Stat.* **7**, 295–318 (2013).
[56] CB Peterson, M Bogomolov, Y Benjamini, C Sabatti, Many phenotypes without many false discoveries: error controlling strategies for multitrait association studies. *Genet. epidemiology* **40**, 45–56 (2016).
[57] M Bogomolov, CB Peterson, Y Benjamini, C Sabatti, Hypotheses on a tree: new error rates and testing strategies. *Biometrika* **108**, 575–590 (2021).
[58] RJ Simes, An improved bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–754 (1986).
[59] Y Benjamini, Y Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).
[60] A Lai, et al., Estimating the ideology of political youtube videos. *Polit. Analysis* pp. 1–16 (2024).