Algorithmic recommendations have limited effects on polarization: A naturalistic experiment on YouTube*†

Naijia Liu[1], Matthew A. Baum[2], Adam J. Berinsky[3], Allison J.B. Chaney[4], Justin de Benedictis-Kessner[2], Andy Guess[5], Dean Knox[6], Christopher Lucas[7], Rachel Mariman[8], and Brandon M. Stewart[9]

[1]Department of Government, Harvard University
[2]John F. Kennedy School of Government, Harvard University
[3]Department of Political Science, Massachusetts Institute of Technology
[4]Fuqua School of Business, Duke University
[5]Department of Politics and School of Public and International Affairs, Princeton University
[6]Operations, Information, and Decisions Department, the Wharton School of the University of Pennsylvania
[7]Department of Political Science, Washington University in St. Louis
[8]Analytics at Wharton, the Wharton School of the University of Pennsylvania
[9]Department of Sociology and Office of Population Research, Princeton University

September 18, 2023

*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, Drew Dimmery, Aleksander Madry, and Michelle Torres for their feedback, and the Wharton Behavioral Lab at the University of Pennsylvania for financial support and operational assistance. This study has been approved by Princeton University IRB (#12989) and the other institutions via Smart IRB (ID: 3931).

†Author contributions: NL, JdBK, AG, DK, and BMS developed the research design. NL, AJBC, and CL collected web data. NL gathered, classified, and experimentally manipulated recommendation networks, with support from RM and DK. JdBK, AG, BMS, and RM designed and implemented the survey, with support from MAB and AJB. NL and DK designed the video platform, and NL collected platform browsing data. NL, JdBK, AG, DK, and BMS designed and implemented analyses. JdBK, AG, and BMS drafted the manuscript, and NL, MAB, AJB, AJBC, JdBK, AG, DK, CL, RM, and BMS revised it. Authors after NL listed alphabetically.

Abstract

An enormous body of academic and journalistic work argues that opaque recommendation algorithms contribute to political polarization by promoting increasingly extreme content. We present evidence that challenges this dominant view, drawing on three large-scale, multi-wave experiments with a combined N of 7,851 human users, consistently showing that extremizing algorithmic recommendations has limited effects on opinions. Our experiments employ a custom-built video platform with a naturalistic, YouTube-like interface that presents real videos and recommendations drawn directly from YouTube. We experimentally manipulate YouTube's actual recommendation algorithm to create ideologically balanced and slanted variations. Our design allows us to directly intervene in a cyclical feedback loop that has long confounded the study of algorithmic polarization—the complex interplay between algorithmic *supply* of content recommendations and user *demand* for its consumption—to examine the downstream effects of recommendation-consumption cycles on policy attitudes. We use data on over 125,000 experimentally manipulated recommendations and 26,000 platform interactions to estimate how recommendation algorithms alter users' media consumption decisions and, indirectly, their political attitudes. Our work builds on recent observational studies showing that algorithm-driven "rabbit holes" of recommendations may be less prevalent than previously thought. We provide new experimental evidence casting further doubt on widely circulating theories of algorithmic polarization, showing that even large perturbations of real-world recommendation systems that substantially modify consumption patterns have limited causal effects on policy attitudes. 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. However, our findings also reveal practical limits to effect sizes that are feasibly detectable in academic experiments.

1 Introduction

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 (Sunstein 2017). 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 (Tufekci 2018). 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 (Arceneaux and Johnson 2013; de Benedictis-Kessner et al. 2019). In turn, this makes experimental results difficult to generalize to the real-world concerns 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, Twitter, or TikTok—because their underlying recommendation algorithms are black boxes 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 (1) this algorithm-induced supply of polarized media to (2) demand-side changes in consumer watching decisions, much less (3) 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 extreme content, but few investigations of the polarizing effects of that content (Papadamou et al. 2020; Ledwich and Zaitsev 2019; Ribeiro et al. 2019; Hosseinmardi et al. 2021; Brown et al. 2022; Chen et al. 2022; Haroon et al. 2022; Hosseinmardi et al. 2023).

To test widely circulating theories about this phenomenon, we develop a new experimental 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

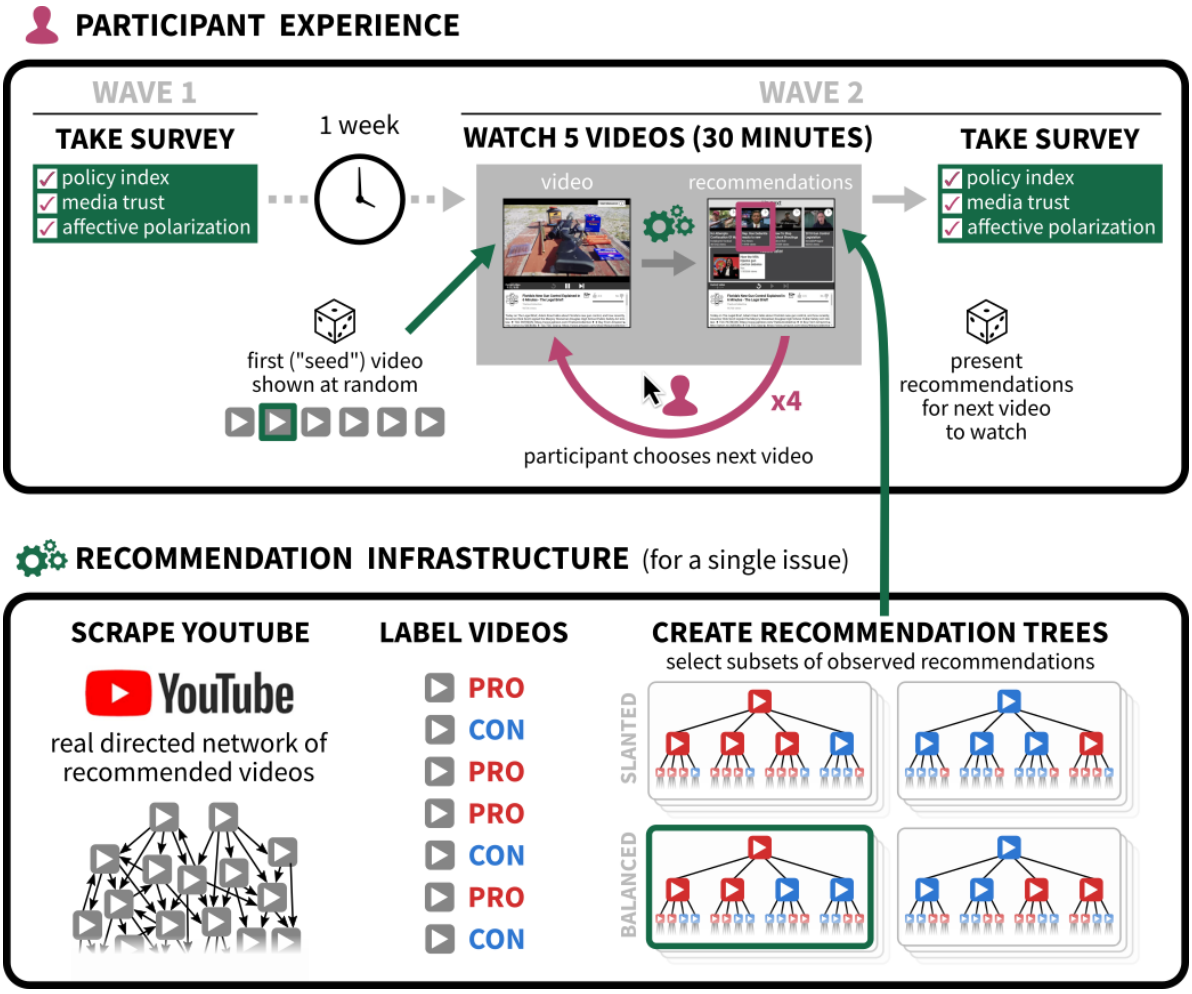


Figure 1: An overview of our new experimental design. In the first wave, participants answer a series of questions. One week later in the second wave, participants are randomized to a seed video and a recommendation system from which they choose future videos to watch. After watching five videos, they take a followup survey.

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 ideological balance of 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. Finally, using a multi-wave survey, we explore how experimental intervention causes individuals to change policy opinions, increase partisan animosity, or alter attitudes toward mainstream media. Figure 1 provides a graphical overview of the design, which was preregistered with the Open Science Framework.[1] Below, we present the results of three studies with a combined N of 7,851. Our analyses draw on over 125,000 experimentally manipulated supply-side video recommendations; more than 26,000 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 in terms of changes over a one-week period.

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. 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, lending pause to widely circulating, unequivocal claims about the influence of algorithmic recommendations on political polarization.

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 3, we describe our multi-wave survey experimental design and the video-recommendation platform that we built to conduct it. In Section 4, we present the results from three studies on the policy issues of 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 and propose directions for future work.

[1]Details of the preregistration can be seen at the following links: study 1 and studies 2–3.

2 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 to serve up increasingly extreme content over time (Tufekci 2018, see e.g.). The argument bears a close similarity to earlier warnings about "filter bubbles" that can form when ranking systems are optimized for predicted engagement, and the potentially polarizing effects of consuming information from the resulting like-minded sources that appear on the feed (Pariser 2011; Sunstein 2017). Algorithmic systems of this sort are known to 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 (Davidson et al. 2010; Covington, Adams and Sargin 2016; Zhao et al. 2019). 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 (Chaney, Stewart and Engelhardt 2018).

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. Since the pioneering work of Hovland, Janis and Kelley (1953), a venerable social-science tradition has used experiments to understand the persuasive effects of films and videos. 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., Iyengar and Kinder 2010). 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 (Arceneaux and Johnson 2013; de Benedictis-Kessner et al. 2019; Levendusky 2013). 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 (Prior 2007). Related insights inform the

current literature on the effects of digital media and social media (Bail et al. 2018; Guess et al. 2021; Levy 2021).

To account for the role of user demand in persuasion, Arceneaux and Johnson (2013) 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 (Iyengar and Hahn 2009; Levendusky 2013; Stroud 2008); 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 (Jamieson and Cappella 2008). Recent work has sought to estimate the causal effect of partisan media *specifically on those who choose to consume it* (Arceneaux and Johnson 2013; Gaines et al. 2007; Knox et al. 2019)—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. (2019) 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. (2022) 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. (2022) 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. (2023) show that YouTube's recommendations quickly "forgets" a user's prior extremist history if they switch back to moderate content.

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. (2021) 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. (2022) 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. While much of the work has focused on the recommendation or consumption of ideological content, *there is essentially no research on the causal persuasive effects of the chosen content or the algorithms that recommend it.*

We build on this line 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 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 that primarily gives options from the same ideological perspective as the most recently watched video, and a "balanced" algorithm 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 (Chen et al. 2022). Platforms like Youtube are a moving target (Munger 2019; Shaw 2023) but our design suggests that even implementing a dramatically more slanted algorithm has limited effects on opinion formation.

In the analyses that follow, we show 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 work such as Ribeiro et al. (2019) 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 (Lewis 2018), supply-side changes in the recommended content may affect other, longer-term components of demand, including the trust they place in various types of news sources (Arceneaux, Johnson and Murphy 2012; Guess et al. 2021). In turn, theories of algorithmic polarization suggest that this change in consumption will indirectly lead to a number of changes in user attitudes. 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 (Chong and Druckman 2007), cue-taking (Druckman, Peterson and Slothuus 2013), or new policy-relevant facts (Kalla and Broockman 2022). Finally, we examine whether manipulating the recommendation algorithm has a more general impact on affective polarization, rather than issue-specific polarization. This is because prior work has shown traditional media's role in affective polarization (Druckman et al. 2019)—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.

3 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. This design is used in three related studies. We gathered real YouTube videos on two policy issues, 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 chose videos on two separate policy issues, gun control and minimum wage, in order to test our hypotheses in both highly salient and less salient issue areas. 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.[2]

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 recommendedations.[3] 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.

Videos on the selected policy topics, along with their recommendations, were identified via the YouTube API. Starting with the list of recommendations that the YouTube API provided for each video, 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.[4] Two aspects of this process

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

[3]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.

[4]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,

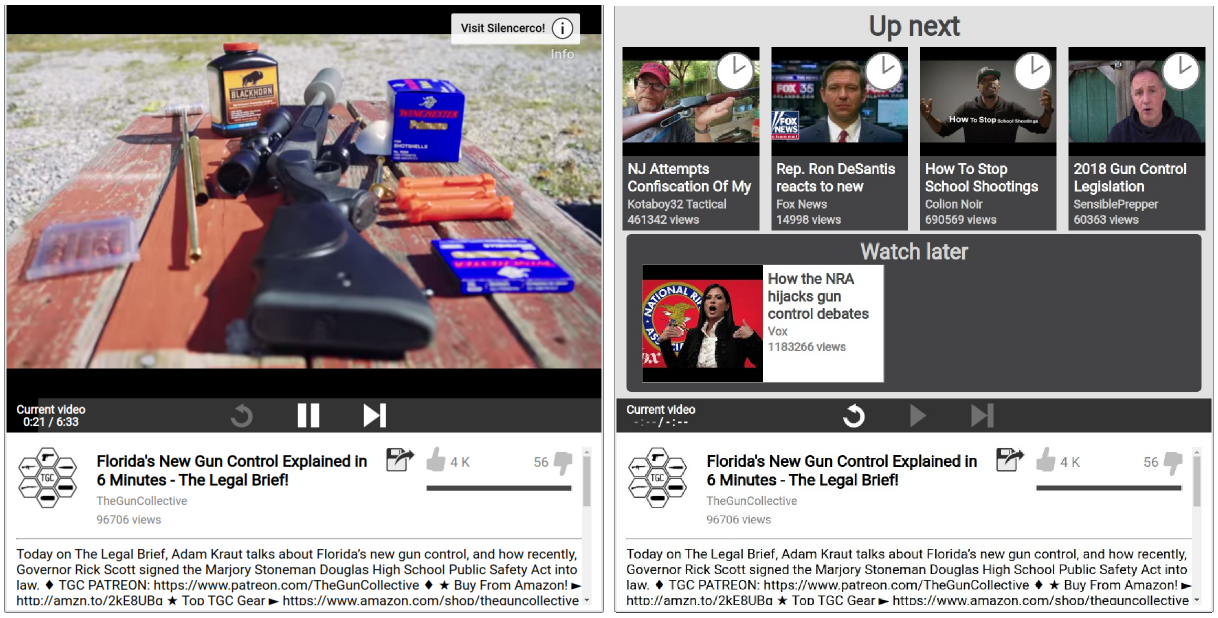


Figure 2: **Video platform interface and recommendations.** The left panel shows the video-watching interface for an example video in study 1, and the right panel shows an example of recommendations that were presented to respondents after finishing this video.

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.[5] A second point is that, like most prominent audits of the YouTube recommendation algorithm (e.g., Ribeiro et al. 2019; Ledwich and Zaitsev 2019), 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. (2022) find 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 an estimand that remains highly informa-

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. Appendix A contains more details about construction of the recommendation trees, hand-coding, classification, and validation.

[5]See Appendix C for more details.

tive for policy questions about algorithmic recommendations, even if we cannot study the personalization process directly.[6]

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.[7] 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. 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).

We recruited large and diverse samples across our three studies. Studies 1 and 2 respectively recruited $N = 2,583$ and $N = 2,442$ respondents on MTurk via CloudResearch, and study 3 drew $N = 2,826$ respondents from YouGov. Our outcomes involve both behaviors (interactions with the video platform) and attitudes (responses on a post-treatment surveys). 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).[8] Our media-trust questions were taken from standard batteries used in research on political communication (e.g. Arceneaux and Johnson 2013), while our measures of affective polarization were similarly taken from validated measures of out-party animosity (e.g. Druckman and Levendusky 2019). Following our pre-registered plan, we assessed the effects of the video recommendation algorithm by

[6]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.

[7]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.

[8]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 for question wordings.

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.[9]

4 Results

Below, we present side-by-side results from all three studies 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. We begin with algorithmic effects on liberal and conservative "ideologue" respondents in Section 4.1 before proceeding to algorithmic effects on "moderate" respondents in Section 4.2. Finally, in Section 4.3, 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.

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 2: (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.[10] Plots show 90% and

[9]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 (2013) 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.

[10]To account for the four families of outcomes, we conduct multiple-testing corrections following our pre-analysis plan and the recommendations of Peterson et al. (2016) and Bogomolov et al. (2021) 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 (Simes 1986) to combine layer-2 p-

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.

4.1 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

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 (Benjamini and Hochberg 1995) 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.

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 16% of the respondents moving one level up on each of the index's five-point components.[11]

4.2 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' 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

[11] 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.

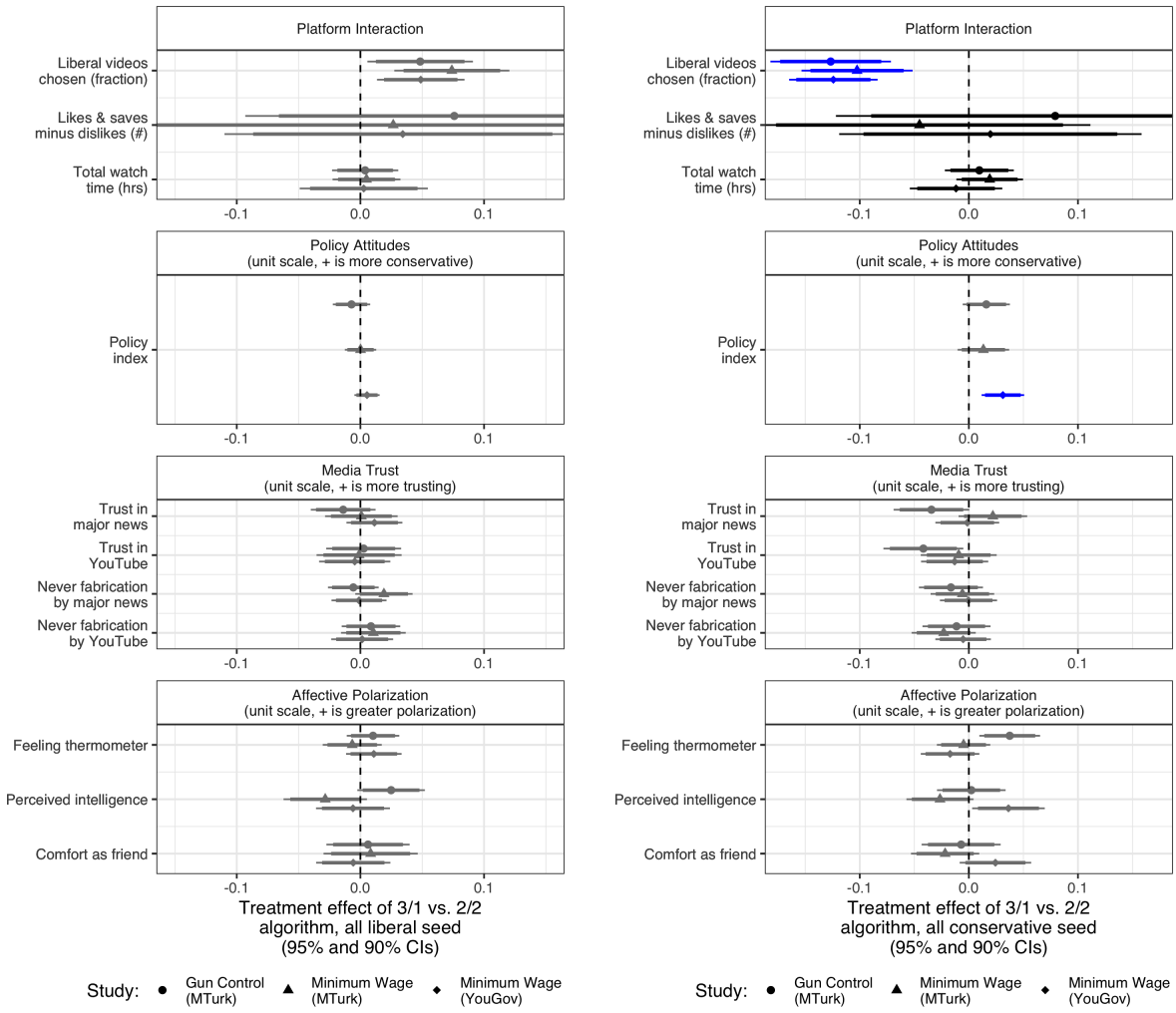


Figure 3: **Effects of recommendation algorithm among ideologues.** Both panels show the results of more algorithmic recommendation slant (vs. balance) on behaviors and attitudes among ideologues (those in the first and third tercile of pre-treatment policy attitudes). The left panel shows effects among 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.

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 attitude change. That is, we can rule out anything greater than these quite-modest effects on policy attitudes caused by more extreme recommendation algorithms.

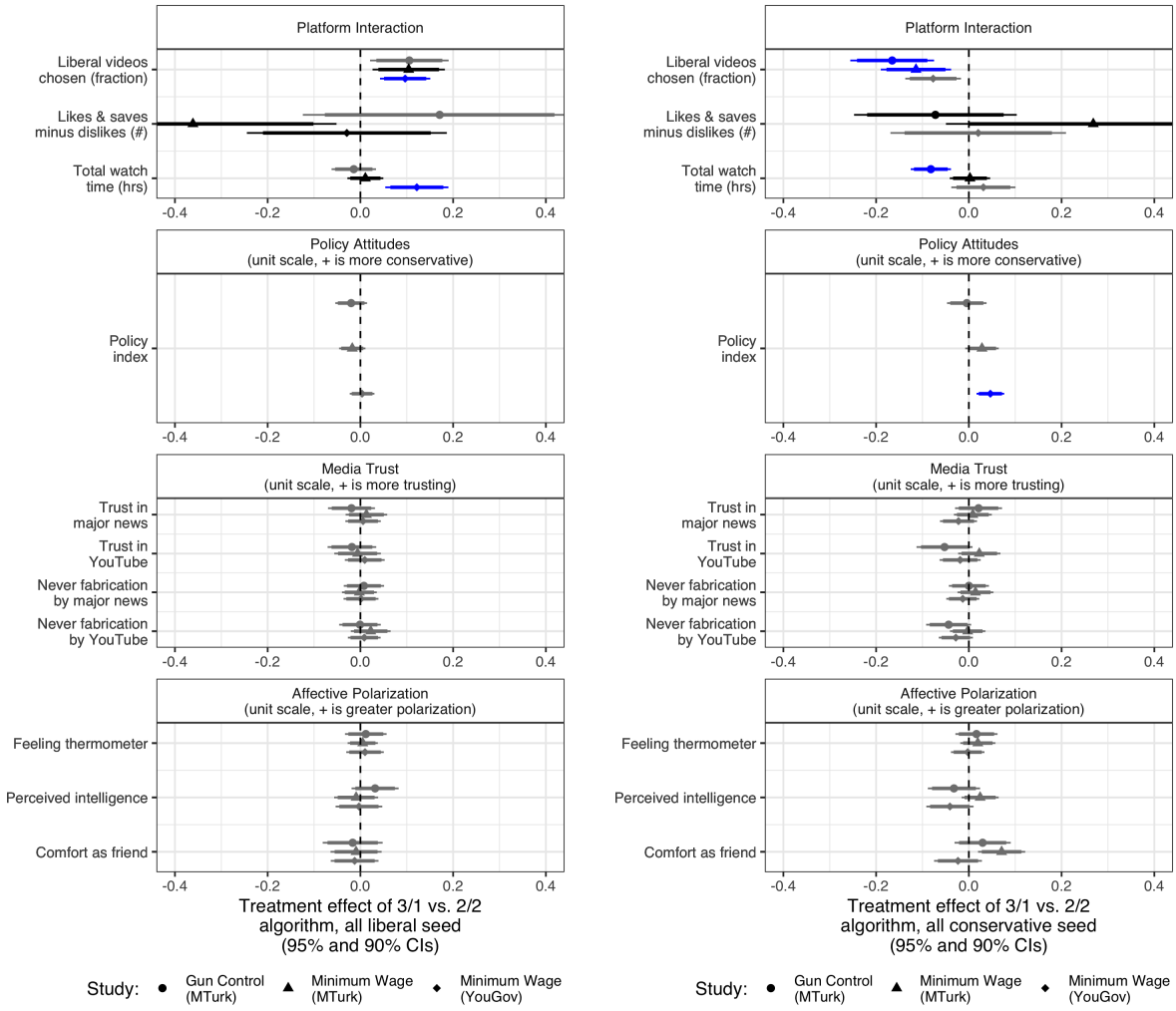


Figure 4: **Effects of recommendation algorithm among moderates.** Both panels shows the results of more algorithmic recommendation extremity (vs. balance) on behaviors and attitudes among moderates (those in the middle tercile of pre-treatment policy attitudes). The left panel shows effects among those 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.

4.3 Forced-Exposure Effects Among Moderate Respondents

Finally, 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 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 distinguishable from the null hypothesis.

5 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

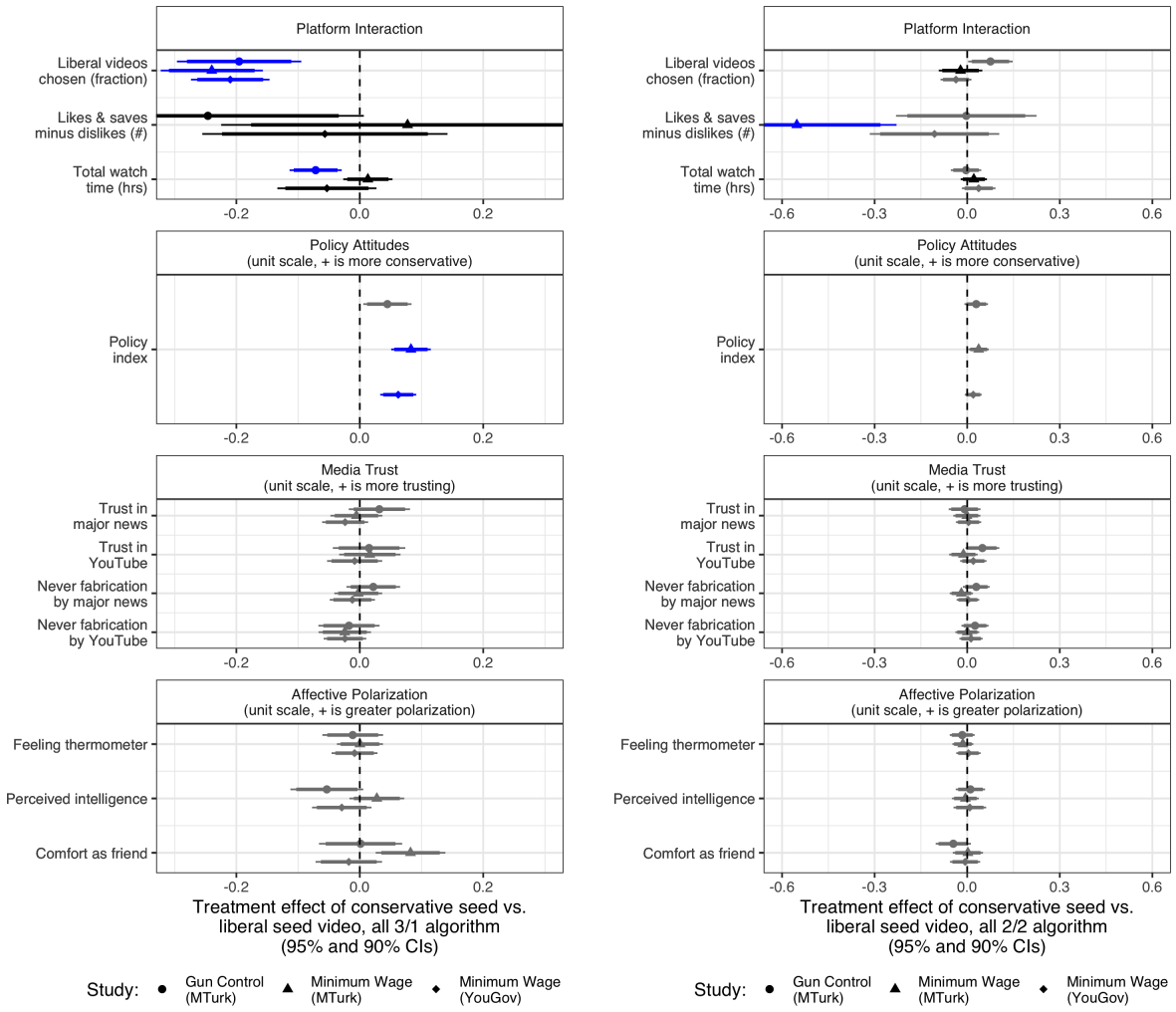


Figure 5: **Effects of seed video slant among moderates.** Both panels shows the results of a more conservative seed video on behaviors and attitudes among moderates (those in the middle tercile of pre-treatment attitudes). The left panel shows effects among those respondents assigned to a 3/1 recommendation 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.

> 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. (Tufekci 2018)

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 for YouTube, but for 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 (O'Neil 2017). Yet if this claim 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. Yet we find limited support for such hypotheses.

Of course, in many ways the situation we can test with our experimental design is not the entirety of the story that Tufekci and others describe. It remains possible that long-term exposure to personalized recommendation systems could lead to the conjectured radicalization. Work by Centola (2018) has shown that repeated exposure is important to behavior contagions. It also remains possible that there are heterogeneous effects—though we failed to detect such heterogeneity in preregistered exploratory analyses examining the moderating role of age, gender, political interest, and YouTube consumption. 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 a $N = 7851$ person randomized controlled trial, 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. Notably, the narrow confidence intervals on attitudinal effects show that even the maximum effect sizes consistent with our results are small, relative to recent experiments on media persuasion with approximately comparable stimuli.[12] 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 (Guess 2021; Wittenberg et al. forthcoming), which implies that only a limited set of people will consume highly imbalanced media when given the opportunity.

Although our study does not provide convincing evidence that the recommendation-

[12]While not completely analogous, our findings are substantially smaller than those in Guess and Coppock (2018), which looked in part at the effect of videos on minimum-wage policy with forced-choice experiments. Our largest attitudinal findings (the effect of seed video among 3/1-assigned moderates) are more consistent with the scale of the findings on persuasion in de Benedictis-Kessner et al. (2019) (see their Figure 3 and replications in the Appendix).

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 perturbations in exposure that we induce. Yet like users' behavior, the production of content is dynamic and subject to incentives. As Munger and Phillips (2019) elaborate, 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.[13] 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 three studies thus collectively suggest that extreme content served by algorithmic recommendation sys-

[13]In recruiting our experimental subjects, we used approval requirement qualifications and attempted to recruit a balanced set of political opinions on Mechanical Turk. We believe that the difficulty we had recruiting respondents that fit these criteria suggests 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.

tems has a limited radicalizing influence on political attitudes and behavior, if this influence even exists.

References

Arceneaux, K. and M. Johnson. 2013. *Changing Minds or Changing Channels?: Partisan News in an Age of Choice*. Chicago Studies in American Politics University of Chicago Press.
URL: *https://books.google.com/books?id=YZdyvQ9R_hEC*

Arceneaux, Kevin, Martin Johnson and Chad Murphy. 2012. "Polarized political communication, oppositional media hostility, and selective exposure." *The Journal of Politics* 74(1):174–186.

Bail, Christopher A, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout and Alexander Volfovsky. 2018. "Exposure to opposing views on social media can increase political polarization." *Proceedings of the National Academy of Sciences* 115(37):9216–9221.

Benjamini, Yoav and Yosef Hochberg. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal statistical society: series B (Methodological)* 57(1):289–300.

Bogomolov, Marina, Christine B Peterson, Yoav Benjamini and Chiara Sabatti. 2021. "Hypotheses on a tree: new error rates and testing strategies." *Biometrika* 108(3):575–590.

Brown, Megan A, James Bisbee, Angela Lai, Richard Bonneau, Jonathan Nagler and Joshua A Tucker. 2022. "Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users." *Available at SSRN 4114905* .

Centola, Damon. 2018. *How behavior spreads: The science of complex contagions*. Vol. 3 Princeton University Press Princeton, NJ.

Chaney, Allison J. B., Brandon M. Stewart and Barbara E. Engelhardt. 2018. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*. RecSys '18 New York, NY, USA: Association for Computing Machinery pp. 224–232.
URL: *https://doi.org/10.1145/3240323.3240370*

Chen, Annie Y, Brendan Nyhan, Jason Reifler, Ronald E Robertson and Christo Wilson. 2022. "Subscriptions and external links help drive resentful users to alternative and extremist YouTube videos." *arXiv preprint arXiv:2204.10921* .

Chong, Dennis and James N Druckman. 2007. "Framing theory." *Annu. Rev. Polit. Sci.* 10:103–126.

Covington, Paul, Jay Adams and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*. RecSys '16 New York, NY, USA: Association for Computing Machinery pp. 191–198.
URL: *https://doi.org/10.1145/2959100.2959190*

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.

de Benedictis-Kessner, Justin, Matthew A Baum, Adam J Berinsky and Teppei Yamamoto. 2019. "Persuading the Enemy: Estimating the Persuasive Effects of Partisan Media with the Preference-Incorporating Choice and Assignment Design." *American Political Science Review* 113(4):902–916.

Druckman, James N, Erik Peterson and Rune Slothuus. 2013. "How elite partisan polarization affects public opinion formation." *American Political Science Review* 107(1):57–79.

Druckman, James N and Matthew S Levendusky. 2019. "What do we measure when we measure affective polarization?" *Public Opinion Quarterly* 83(1):114–122.

Druckman, James N, SR Gubitz, Ashley M Lloyd and Matthew S Levendusky. 2019. "How Incivility on Partisan Media (De) Polarizes the Electorate." *The Journal of Politics* 81(1):291–295.

Gaines, Brian J., James H. Kuklinski, Paul J. Quirk, Buddy Peyton and Jay Verkuilen. 2007. "Same Facts, Different Interpretations: Partisan Motivation and Opinion on Iraq." *Journal of Politics* 69(4):957–974.

Guess, Andrew and Alexander Coppock. 2018. "Does Counter-Attitudinal Information Cause Backlash? Results from Three Large Survey Experiments." *British Journal of Political Science* pp. 1–19.
URL: *https://www.cambridge.org/core/article/does-counterattitudinal-information-cause-backlash-results-from-three-large-survey-experiments/526B71F3BB76A39C1101384D576208D4*

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

Guess, Andrew M, Pablo Barberá, Simon Munzert and JungHwan Yang. 2021. "The consequences of online partisan media." *Proceedings of the National Academy of Sciences* 118(14):e2013464118.

Haroon, Muhammad, Anshuman Chhabra, Xin Liu, Prasant Mohapatra, Zubair Shafiq and Magdalena Wojcieszak. 2022. "YouTube, the great radicalizer? Auditing and mitigating ideological biases in YouTube recommendations." *arXiv preprint arXiv:2203.10666* .

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).

Hosseinmardi, Homa, Amir Ghasemian, Miguel Rivera-Lanas, Manoel Horta Ribeiro, Robert West and Duncan J Watts. 2023. "Causally estimating the effect of YouTube's recommender system using counterfactual bots." *arXiv preprint arXiv:2308.10398* .

Hovland, Carl Iver, Irving Lester Janis and Harold H Kelley. 1953. "Communication and persuasion.".

Iyengar, Shanto and Donald R Kinder. 2010. *News that matters: Television and American opinion*. University of Chicago Press.

Iyengar, Shanto and Kyu S Hahn. 2009. "Red media, blue media: Evidence of ideological selectivity in media use." *Journal of communication* 59(1):19–39.

Jamieson, Kathleen Hall and Joseph N Cappella. 2008. *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press.

Kalla, Joshua L and David E Broockman. 2022. ""outside lobbying" over the airwaves: A randomized field experiment on televised issue ads." *American Political Science Review* 116(3):1126–1132.

Knox, Dean, Teppei Yamamoto, Matthew A. Baum and Adam J. Berinsky. 2019. "Design, Identification, and Sensitivity Analysis for Patient Preference Trials." *Journal of the American Statistical Association* 114(528):1532–1546. Publisher: Taylor & Francis.
URL: *https://amstat.tandfonline.com/doi/full/10.1080/01621459.2019.1585248*

Ledwich, Mark and Anna Zaitsev. 2019. "Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization." *arXiv:1912.11211 [cs]* . arXiv: 1912.11211.
URL: *http://arxiv.org/abs/1912.11211*

Levendusky, Matthew. 2013. *How Partisan Media Polarize America*. University of Chicago Press.

Levy, Ro'ee. 2021. "Social media, news consumption, and polarization: Evidence from a field experiment." *American economic review* 111(3):831–870.

Lewis, Rebecca. 2018. "Alternative influence: Broadcasting the reactionary right on YouTube." *Data & Society* 18.

Lin, Winston. 2013. "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique." *Annals of Applied Statistics* 7(1):295–318.

Munger, Kevin. 2019. "The limited value of non-replicable field experiments in contexts with low temporal validity." *Social Media+ Society* 5(3):2056305119859294.

Munger, Kevin and Joseph Phillips. 2019. "YouTube Politics.". Publisher: OSF.
URL: *https://osf.io/4wk63/*

O'Neil, Cathy. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Papadamou, Kostantinos, Antonis Papasavva, Savvas Zannettou, Jeremy Blackburn, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini and Michael Sirivianos. 2020. "Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children." *Proceedings of the International AAAI Conference on Web and Social Media* 14:522–533.
URL: *https://www.aaai.org/ojs/index.php/ICWSM/article/view/7320*

Pariser, Eli. 2011. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin. Google-Books-ID: wcalrOI1YbQC.

Peterson, Christine B, Marina Bogomolov, Yoav Benjamini and Chiara Sabatti. 2016. "Many phenotypes without many false discoveries: error controlling strategies for multitrait association studies." *Genetic epidemiology* 40(1):45–56.

Prior, Markus. 2007. *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*. Cambridge University Press.

Ribeiro, Manoel Horta, Raphael Ottoni, Robert West, Virgílio A. F. Almeida and Wagner Meira. 2019. "Auditing Radicalization Pathways on YouTube." *arXiv:1908.08313 [cs]* . arXiv: 1908.08313.
URL: *http://arxiv.org/abs/1908.08313*

Shaw, Aaron. 2023. "Social media, extremism, and radicalization." *Science Advances* 9(35):eadk2031.

Simes, R John. 1986. "An improved Bonferroni procedure for multiple tests of significance." *Biometrika* 73(3):751–754.

Stroud, Natalie Jomini. 2008. "Media use and political predispositions: Revisiting the concept of selective exposure." *Political Behavior* 30:341–366.

Sunstein, Cass R. 2017. *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.

Tufekci, Zeynep. 2018. "YouTube, the Great Radicalizer." *The New York Times* 10:2018.

Wittenberg, Chloe, Matthew A Baum, Adam J Berinsky, Justin de Benedictis-Kessner and Teppei Yamamoto. forthcoming. "Media Measurement Matters: Estimating the Persuasive Effects of Partisan Media with Survey and Behavioral Data." *Journal of Politics* .

Zhao, Zhe, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi and Ed Chi. 2019. Recommending what video to watch next: a multitask ranking system. In *Proceedings of the 13th ACM Conference on Recommender Systems*. pp. 43–51.

Supplementary Information

A Creating Recommendation Trees

A.1 Gun Policy

We collected two starting videos from YouTube about gun policy and used them to construct a recommendation network by querying the YouTube API. We use this directed network to construct recommendation trees representing the different recommendation systems discussed in the paper.

Using the YouTube Data API, we started from two roughly comparable videos (one gun-rights video and one gun-control video), we collected a recommendation network consisting of around 78,000 nodes (unique videos) and 350,000 directed edges (candidate recommendations). The starting videos were selected to ensure that they had a clear stance.[14] Up to 50 non-personalized recommendations were collected for each node, using the `Search` > `relatedToVideoId` functionality.

The videos vary in length from several minutes to several hours; the majority are shorter than 20 minutes.[15] For feasibility of the experiment, we use only videos up to 10 minutes long. We then coarsely screen for topicality by applying a regular-expression filter to their titles.[16] For videos passing this initial topicality screening, we extracted textual transcripts to classify for ideological valence.

A training set of roughly 2,000 videos was manually labeled as "anti-gun" policy videos, "pro-gun" policy videos, "gun enthusiast" videos, and "other" via workers on Mechanical Turk. A cross-validated (linear) support-vector machine was trained on the training-set transcripts using bag-of-words features, then used to label the full corpus of videos. Regularization term was selected by cross-validation using the training set. The SVM train-test split has a accuracy of 82%, using the 2,000 hand-labeled videos. We subset to videos categorized as "anti-gun" or "pro-gun" and subjected the most prominent 283 videos in the network (in terms of the number and position of placements in the recommendation trees described in the next section) to a manual evaluation by a subset of the authors. Corrections were made as necessary and the trees were regenerated. In the final trees, at least one of the authors had manually reviewed 100% of the seed videos, 93% of the first-level videos, 73% of the second-level videos, 46% of the third-level videos and 30% of the fourth-level videos.

For each of the 10 seed videos, we make 20 trees for each recommendation system condition. When a respondent is randomly assigned to a seed/system combination, we randomly chose one of the 20 unique trees to assign. We continually conduct checks to remove auto-generated trees that contain deleted videos.

[14] We used a video from Fox News and a video from *The Atlantic*.

[15] The 25th percentile in video length is 6 minutes, the median is 10 minutes, and the 75th percentile is 17.5 minutes.

[16] The filter was hand-tuned to retain both gun rights and gun control videos from a random sample of videos.

A.2 Minimum Wage

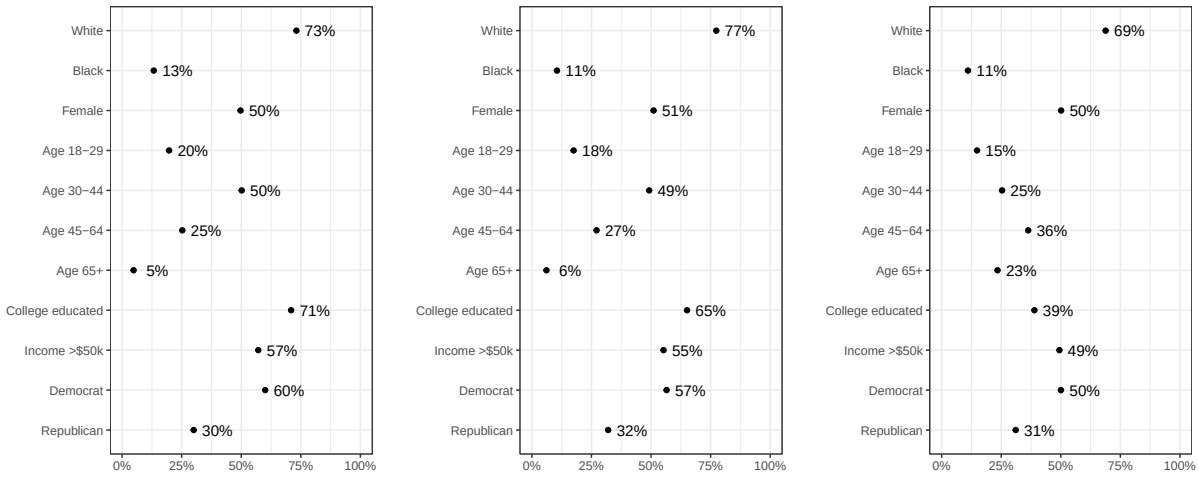
For the feasibility of these experiments, we use only videos up to 12 minutes long. We then coarsely screen for topicality by applying a regular-expression filter to their titles. For videos passing this initial topicality screening, we extracted textual transcripts to classify for ideological valence. MTurk workers manually coded all videos. For each video classification task, we assigned three workers and labeled the videos following the 2/3 majority opinion. We saw a very high inter-coder agreement rate (on average 80% to 85% for multiple rounds of classification). Then, we filtered out videos that did not have a binary label. Only videos that are either support or against raising the minimum wage appear in the final trees. Finally, authors conducted an additional round of classification on approximately 500 videos to validate the MTurk results. These steps resulted in a smaller sub-graph of around 1,090 unique videos with a binary label and are less than 12 minutes in length.

B Experimental Implementation and Preregistration Details

We preregistered all three of our experiments ahead of fielding each respective one. We preregistered study 1 on Tuesday, June 8, 2021 just before beginning to field Wave 1 of the survey. Wave 1 recruited 3,902 participants (with the last coming in on Tuesday, June 15) which was a smaller number of participants than initially intended. In order to increase participation, survey compensation was raised to $2 from $1.50 for later waves of participants and we lifted the quota on political views. We posted a revised pre-analysis plan on Thursday, June 17, 2021, immediately before inviting 2,862 respondents back for Wave 2. This was approximately two days later than initially intended. We posted Wave 3 on Friday, June 25, 2021 and closed on Friday, July 2, 2021.

Despite our attempts to recruit equal proportions of liberals and conservatives, our sample is somewhat skewed in terms of ideological self-placement (59% liberals and 30% conservatives including leaners) and partisan identification (63% Democrats and 27% Republicans including leaners). Well-known biases in terms of age distribution on MTurk are also present (20% under 30, 50% 30–44, and only 5% age 65 or older), though this arguably accords with the target population of frequent streaming video platform users.[17] Since partisanship is not completely predictive of gun attitudes, we still obtain substantial variation in our pre-treatment gun policy measure, though the distribution is still somewhat right-skewed (mean 0.41, median 0.35 on a 0–1 scale).

The demographics of our three survey samples were relatively similar and are shown in the three panels of Figure B-6. In addition, Tables B-1, B-2, and B-3 show these descriptive features of our data in tabular format.



(a) Study 1 (b) Study 2 (c) Study 3

Figure B-6: Respondent Demographics

[17] See `https://www.pewresearch.org/internet/fact-sheet/social-media/` for self-reported YouTube use by age category. See Figure B-6 and Tables B-1, B-2, and B-3.

Statistic	Mean	St. Dev.	Median	Min	Max	N
Female	0.50	0.50	0.00	0.00	1.00	3,904
White	0.73	0.44	1.00	0.00	1.00	3,903
Black	0.13	0.34	0.00	0.00	1.00	3,903
Age	39.94	12.25	37.00	18.00	84.00	3,903
College educated	0.71	0.45	1.00	0.00	1.00	3,904
Income >50k	0.57	0.49	1.00	0.00	1.00	3,902

Table B-1: Study 1 Survey Respondent Demographics (Wave 1)

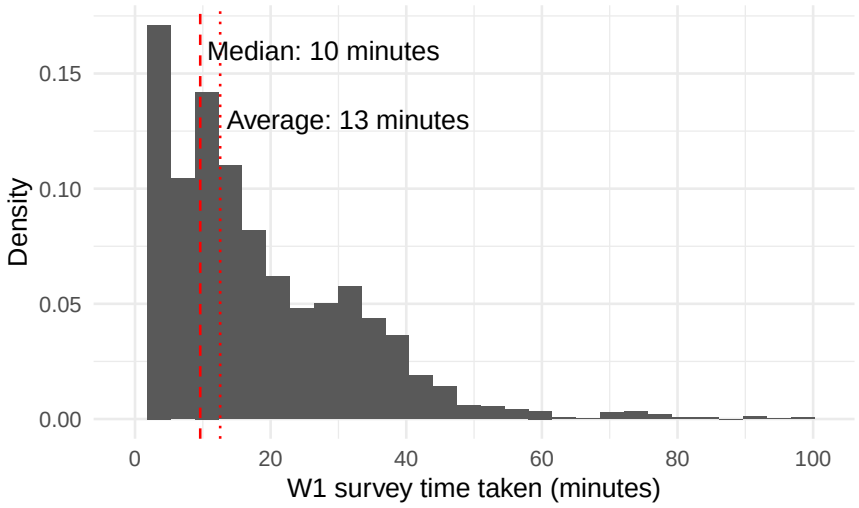
Statistic	Mean	St. Dev.	Median	Min	Max	N
Female	0.51	0.50	1.00	0.00	1.00	3,095
White	0.77	0.42	1.00	0.00	1.00	3,094
Black	0.11	0.31	0.00	0.00	1.00	3,094
Age	41.10	12.58	39.00	19.00	98.00	3,095
College educated	0.65	0.48	1.00	0.00	1.00	3,094
Income >50k	0.55	0.50	1.00	0.00	1.00	3,095

Table B-2: Study 2 Survey Respondent Demographics (Wave 1)

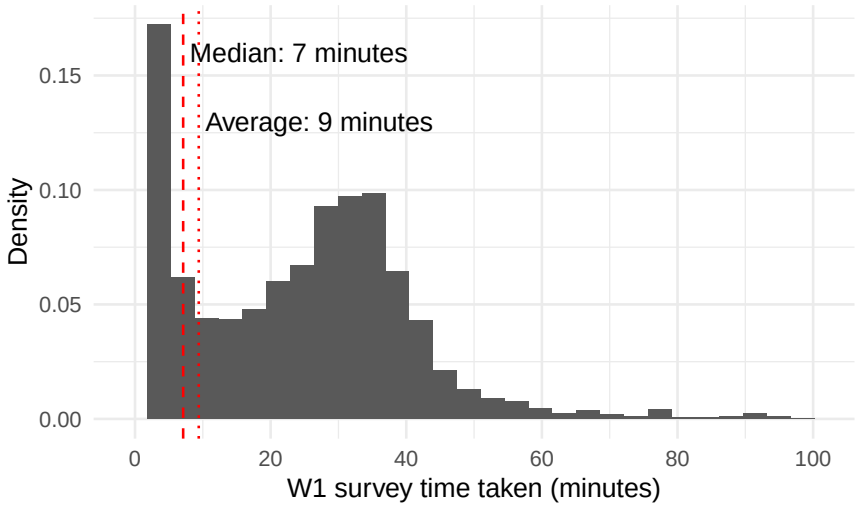
Statistic	Mean	St. Dev.	Median	Min	Max	N
Female	0.50	0.50	1	0	1	4,591
White	0.69	0.46	1	0	1	4,591
Black	0.11	0.31	0	0	1	4,591
Age	50.29	16.94	52	19	94	4,591
College educated	0.39	0.49	0	0	1	4,591
Income >50k	0.49	0.50	0	0	1	4,591

Table B-3: Study 3 Survey Respondent Demographics (Wave 1)

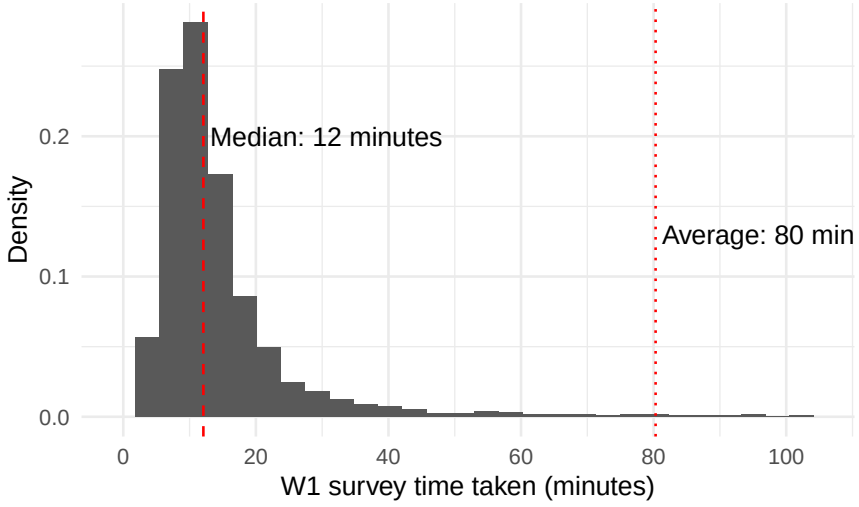
We also explore the amount of time respondents took on our survey. Looking at time spent during Wave 2, which included the video interface and main experiment, we find that participants outside the pure control group spent substantial time engaging with our stimuli and questions (study 1: median 18 minutes, mean 22 minutes); Figure B-7 and Figure **??** plot the full distributions of time taken on each wave of the survey for each study.



(a) Study 1

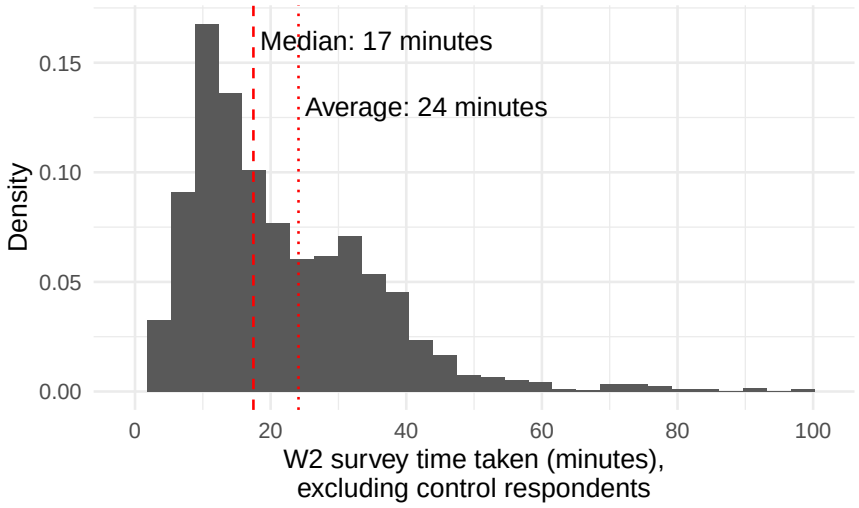


(b) Study 2

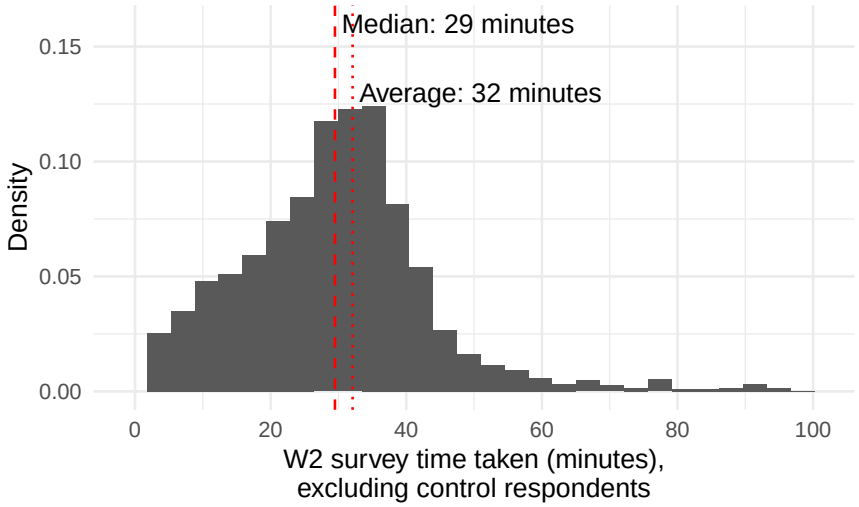


(c) Study 3

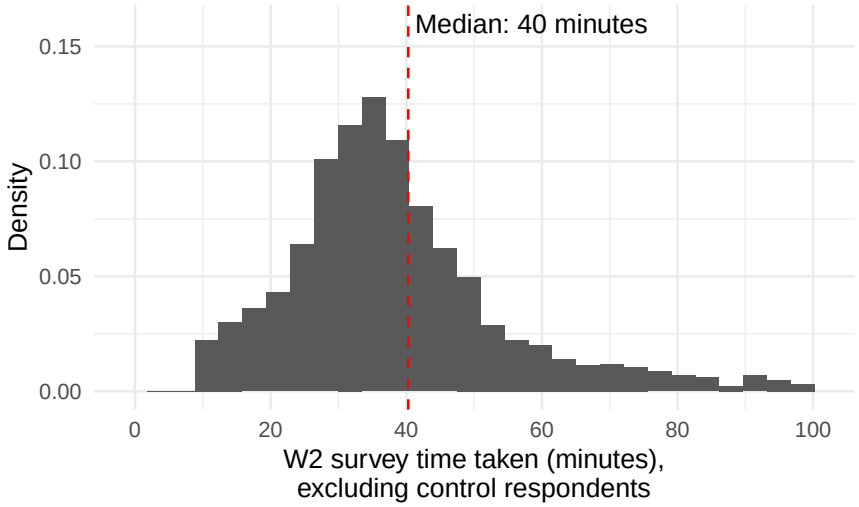
Figure B-7: Time taken by respondents on Wave 1 survey.



(a) Study 1

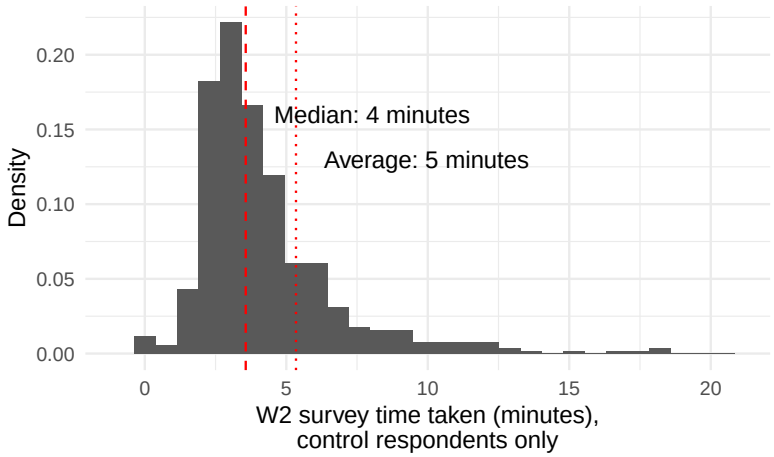


(b) Study 2

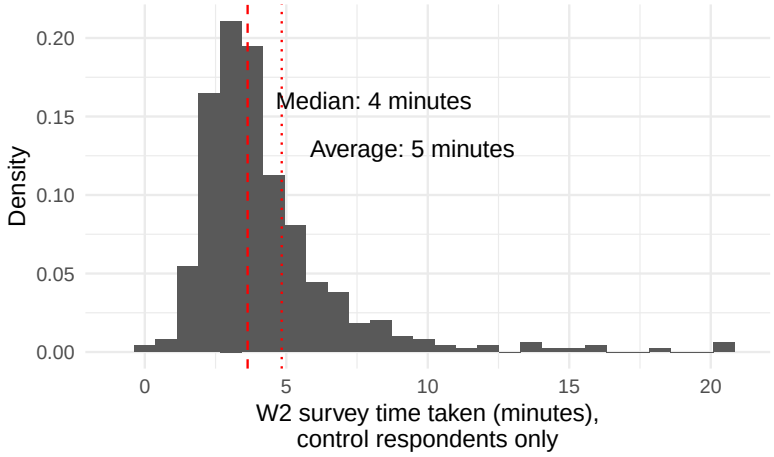


(c) Study 3

Figure B-8: Time taken by respondents on Wave 2 survey + video platform.



(a) Study 1



(b) Study 2

Figure B-9: Time taken by control condition respondents on Wave 2 survey.

C Similarity to Browser-Based Recommendations

To demonstrate the validity of our recommendation trees, which were created with the YouTube API, we create browser-based recommendation trees. To do so, we load a single seed URL in an automated anonymous browser. We record the 20 recommended videos for that seed video, then load each recommended video in separate anonymous browsers (to avoid dependencies created by the ordered history of viewing previous videos). We compare these recommendations to a tree concurrently created via the API, which is the method by which we created the trees used in our experiments (the trees used in the experiment were further filtered on topic according to the text of the video).

Both the API- and browser-based trees start with a single seed video: $15 minimum wage would cut 1.4 million jobs by 2025: CBO. We build the API-based tree by taking 3 steps, recording 50 recommendations for each video in each step. In other words, at the first step, we collect the 50 videos recommended from our single seed video. In the second step, we record the 50 recommended videos for each of those 50 videos, and so on.

However, when loading YouTube in a browser, 20 recommended videos are visible in the browser. It is possible to get additional recommendations by scrolling down, but doing so massively slows down data collection and increases the chances of connection errors. As a result, in our browser-based tree, we collect 20 recommended videos at each node in the tree. Additionally, instead of taking 3 steps, we take 5.

We compare these two trees and find that they are largely similar. In three steps,

To get a better sense of why some recommendations are in the natural tree and not in the API tree, we manually inspect 10 randomly selected recommendations. Figure C shows ten randomly selected branches of the tree. The column farthest on the right shows the origin video (the same for all branches), the second shows the first recommendation in that branch, and so on for five steps. The cell values are the video ID of the youtube video, and $***$ indicates that that particular video is *not* in the api tree.

This table highlights several features of this exercise. First, because each video was inspected in a history-less browser, if the browser-based tree branched off-topic, that branch never returns to videos related to the seed video, and so no subsequent recommendations are also found in the api-tree.

With this in mind, the most important nodes are those in which the recommendations deviate off-topic. To insure that that is in fact what is occurring (as opposed to the browser-based tree recommending on-topic videos that are simply different than those found in the api tree). They are as follows.

Two videos go off topic in the first step on video `Mqn4lYunTX4`. This is a 25-minute video titled "Bone in vs Boneless Steaks (How to be a Steak Expert) The Bearded Butchers."

One additional branch goes off topic in the second step: `wx_72QJTDUs`: "Chris Stapleton: The 60 Minutes Interview."

Three additional branches go off-topic in the third step: `OQ9zng2S810` ("Why North Korea is the Hardest Country to Escape"), `da1vvigy5tQ` ("Reversing Type 2 diabetes starts with ignoring the guidelines — Sarah Hallberg — TEDxPurdueU"), and `TLcw2xsQh68` ("Here's How Larger 34-Inch Off-Road Tires Affect My Ford F-150 Hybrid's MPG and 0-60 MPH Speed!").

In the fourth step, all but one of our branches is off topic. The newly off-topic branches

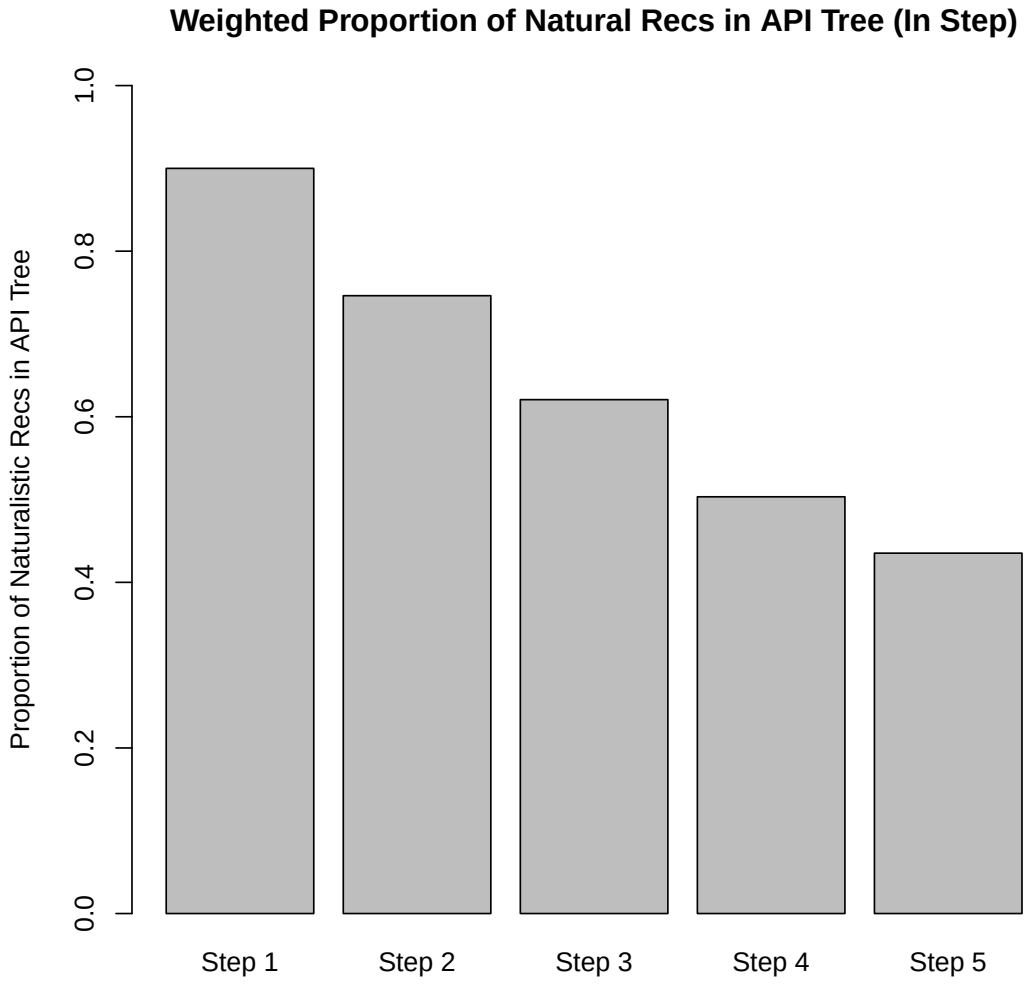


Figure C-10

are `wANiIPO9TiQ` ("Target packaging Tiktok compilation • part 1"), `fKME33GDFZI` ("What to expect next... out of underwriting & Closing Disclosures (CD)"), and `i2trJEIFIvY` ("Why does maths give humans the edge over machines? - with Junaid Mubeen").

In the final step, the last on-topic branch goes off topic with `e0LBJfCTCo8`: "CNBC's Courtney Reagan reports on the groundbreaking life of the late Queen Elizabeth."

Fifth Step	Fourth Step	Third Step	Second Step	First Step	Seed Video
ISaZduGmhEU***	TtzsU4WAJ-k***	phOyUhz-73U***	yomerhQkpSc***	Mqn4lYunTX4***	2voNlYS-8C0
e0LBJfCTCo8***	aqpr0uRsmcs	nLDtZN1dPHk	AtjaRuGkbgQ	RPwqBsc4Ffo	2voNlYS-8C0
7UAoT21eqXI***	RWQKa4qTbkE***	zRWvWe08HTA***	WZRvRbzTU_c***	Mqn4lYunTX4***	2voNlYS-8C0
e7Tao1t0i7E***	wANiIPO9TiQ***	oCmLhc1HNSI	aTVfbSeeS74	3-KMXng5Cp0	2voNlYS-8C0
wngB9_6Vqbc***	C0kWjEYMAfc***	OQ9zng2S810***	lwYOJLGw-Mw	zKyWRRJQbkM	2voNlYS-8C0
auw4Z6FfOT4***	fKME33GDFZI***	C7PfqazmSuQ	FPLcOOkFhP0	UdnkStBTG2k	2voNlYS-8C0
d5wfMNNr3ak***	4lzs5wpLkeA***	da1vvigy5tQ***	S1E8SQde5rk	Hatav_Rdnno	2voNlYS-8C0
63slKb4iGO8***	VUlRz2ih1uc***	TLcw2xsQh68***	-e55Vued028	Hatav_Rdnno	2voNlYS-8C0
wx_72QJTDUs***	GayEgDB1EZY***	kAE3F-35OP0***	wx_72QJTDUs***	wqKfL3z5yM4	2voNlYS-8C0
7dzoGb-jcW4***	i2trJEIFIvY***	ZuXzvjBYW8A	QaN6ibm5r-I	8H4yp8Fbi-Y	2voNlYS-8C0

D MTurk HIT Recruitment Language (Study 1, Wave 1)

Title: Participate in a Streaming Video Study (5–10 minutes)

Description: We're interested in learning how people like you respond to videos shown on an interactive interface. In this initial survey, we would like to learn more about your video habits and background. We may invite you to use a video platform in a future study.

What is this study about? We designed an interactive streaming video interface to present videos about a topic and adapt to your preferences. We are interested in learning how people like you respond to videos and what you remember from the experience.

What is the problem being solved by this study? How to discover and rank content (videos in this case) from the vast quantities available online is a difficult question. We would like to explore how best to present information that is both high quality and relevant to users' interests.

How might research in this area change society? What users demand and what is good collectively for society may not always align. We are broadly interested in understanding the consequences of different ranking approaches on key democratic outcomes. We hope our results will inform decisions by social platforms that increasingly structure our informational choices.

What does it involve? This initial task involves answering a few questions about yourself, including your video watching habits and preferences. Sound and video are required! We may follow up with you and invite you to use our streaming video platform and to answer another set of questions, for additional compensation.

To participate, please open the following survey (8-10 minutes) in a new tab or window.

As suggested, this initial survey will determine eligibility for a future study (with additional compensation) that will involve an interactive, streaming video interface.

E Survey Question Wording

Policy Attitudes

Study 1 - Gun Control

In study 1, our primary outcome of interest was an additive index ranging from 0 to 1 formed from a five-question battery of gun policy attitudinal questions. These questions were adapted from common question wordings placed on national surveys run by Pew, Gallup, the *Washington Post*, and other policy attitude surveys. We show these individual questions below:

1. What do you think is more important — to protect the right of Americans to own guns, or to regulate gun ownership?
    - Protect the right to own guns
    - Regulate gun ownership
2. Do you support or oppose a nationwide ban on the sale of assault weapons?
    - Strongly support
    - Somewhat support
    - Neither support nor oppose
    - Somewhat oppose
    - Strongly oppose
3. Do you support or oppose a nationwide ban on the possession of handguns?
    - Strongly support
    - Somewhat support
    - Neither support nor oppose
    - Somewhat oppose
    - Strongly oppose
4. Suppose more Americans were allowed to carry concealed weapons if they passed a criminal background check and training course. If more Americans carried concealed weapons, would the United States be safer or less safe?
    - Much safer
    - Somewhat safer
    - No difference
    - Somewhat less safe
    - Much less safe
5. Do you support or oppose stricter gun control laws in the United States?

- Strongly support
- Somewhat support
- Neither support nor oppose
- Somewhat oppose
- Strongly oppose

We rescaled each item to a unit scale, with 0 representing the most liberal of the response options and 1 representing the most conservative of the response options (i.e. reverse coding questions 1 and 4) for each question. Using principal components analysis, we found a Cronbach's α of 0.92 for the five-item scale, suggesting that all five items load on the same factor. In the appendix of our resulting manuscript we will report the results of an exploratory factor analysis with varimax rotation of these five attitudinal questions to verify that they load on the same underlying dimension. We then averaged the rescaled outcomes from all five questions to form the additive index such that the index has a range from 0 to 1.

Studies 2 & 3 – Minimum Wage

In studies 2 and 3, our primary outcomes of interest were an additive index ranging from 0 to 1 formed from a five-question battery of attitudinal questions about minimum wage policy. These questions were, similar to our questions from study 1, adapted from common question wordings placed on national surveys. Following an anchoring baseline page that stated "As you may know, the current federal minimum wage is $7.25 an hour," we asked the following individual questions:

1. What do you think the federal minimum wage should be? Please enter an amount between $0.00 and $25.00 in the text box below.

   - _____

2. Some people believe that raising the minimum wage would overly restrict the freedom of businesses to set their own employment policies. Imagine those people are all the way at one end of a scale, at 1. Other people might believe that raising the minimum wage protects workers from businesses exploiting workers. Imagine those people are at the other end of the scale, at 10. Of course, some people fall in between and believe that raising the minimum wage might or might not protect workers from businesses. Where would you place yourself on this scale?

   (a) Would restrict businesses' freedom
   (b)
   (c)
   (d)
   (e)
   (f)

(g)

(h)

(i)

(j)

(k) Would protect workers from exploitation

3. Some people believe that raising the minimum wage would help low-income workers get by. Imagine those people are all the way at one end of a scale, at 1. Other people might believe that raising the minimum wage would hurt low-income workers. Imagine those people are at the other end of the scale, at 10. Of course, some people fall in between and believe that raising the minimum wage might or might not hurt low-income workers. Where would you place yourself on this scale?

   (a) Would help low-income workers

   (b)

   (c)

   (d)

   (e)

   (f)

   (g)

   (h)

   (i)

   (j)

   (k) Would hurt low-income workers

4. How high do you think the federal minimum wage should be?

   - Much higher than the current level
   - Somewhat higher than the current level
   - About the current level
   - Somewhat lower than the current level
   - Much lower than the current level

5. Do you support or oppose raising the federal minimum wage?

   - Strongly support raising the minimum wage
   - Somewhat support raising the minimum wage
   - Neither support nor oppose raising the minimum wage
   - Somewhat oppose raising the minimum wage

- Strongly oppose raising the minimum wage

6. The Raise the Wage Act is a proposal to raise the minimum wage so that it would be increased to $15 per hour by 2025. Do you support or oppose the Raise the Wage Act?

    - Strongly support
    - Somewhat support
    - Neither support nor oppose
    - Somewhat oppose
    - Strongly oppose

7. The Raise the Wage Act is a proposal to gradually raise the minimum wage. The minimum wage would first be increased to $9.50 an hour in 2022. Then, it would be increased by $1.50 an hour or less every year through 2025. Do you support or oppose the Raise the Wage Act?

    - Strongly support
    - Somewhat support
    - Neither support nor oppose
    - Somewhat oppose
    - Strongly oppose

8. How strongly do you support or oppose a $15 minimum wage?

    - Strongly support
    - Somewhat support
    - Neither support nor oppose
    - Somewhat oppose
    - Strongly oppose

Similar to study 1, in studies 2 and 3 we rescaled each item to a unit scale, with 0 representing the most liberal of the response options and 1 representing the most conservative of the response options for each question. For question 1, we rescaled respondents' numeric entries such that $25/hour was the most liberal response option and $0 was the most conservative option.[18] Using principal components analysis, we found a Cronbach's α for the eight-item scale of 0.94 in study 2 and 0.94 for study 3, suggesting that all eight items load on the same factor. We then averaged the rescaled outcomes from all eight questions to form the additive index such that the index has a range from 0 to 1.

[18]We omit any answers that respondents gave that were over $25/hour.

Media Trust/Hostility

In order to measure effects on media trust/hostility, on all three studies we asked two questions about beliefs in fabricating news stories, both by major news organizations and YouTube channels, shown below.

1. Based on what you know, how often do you believe the nation's major news organizations fabricate news stories?
    - All the time
    - Most of the time
    - About half the time
    - Once in a while
    - Never
2. Based on what you know, how often do you believe YouTube channels fabricate news stories?
    - All the time
    - Most of the time
    - About half the time
    - Once in a while
    - Never

As an additional measure of media trust, we used a grid question which asked respondents to rate how much, if at all, they trust the information they get from several media sources. Specifically, this grid asked about trust in information from major news organizations, local news outlets, social media, and YouTube. Response options were: A lot, Some, Not too much, and Not at all. We examined effects on both trust in major news organizations and in YouTube.

Affective Polarization

Our fourth family of outcomes for all three studies measured respondents' affective polarization using several standard questions for this concept. First, we used a pair of questions (shown below) that asked respondents how smart people are who support the party the respondent prefers vs. the other party (1–5 where 5 indicates "extremely" smart for both). This outcome measure was calculated as the difference in perceptions between the ingroup question and the outgroup question. While the results were collected for respondents who did not indicate a preference for or lean towards a political party (i.e. "pure independents"), we did not use these responses.

1. In general, how smart are people who support Democrats?
    - Extremely

- Very
- Somewhat
- A little
- Not at all

2. In general, how smart are people who support Republicans?

    - Extremely
    - Very
    - Somewhat
    - A little
    - Not at all

Second, we looked at the difference between the feeling thermometer scores respondents assigned to the outparty vs. the inparty. Finally, we measured the difference between responses on two questions about comfort with having members of the inparty vs. outparty as close personal friends, shown below (same conditions on pure independents apply for these measures):

1. How comfortable are you having close personal friends who are Democrats?

    - Not at all comfortable
    - Not too comfortable
    - Somewhat comfortable
    - Extremely comfortable

2. How comfortable are you having close personal friends who are Republicans?

    - Not at all comfortable
    - Not too comfortable
    - Somewhat comfortable
    - Extremely comfortable