Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

How to Train Your YouTube Recommender to Avoid Unwanted Videos

Alexander Liu, Siqi Wu, Paul Resnick

University of Michigan, Ann Arbor
{avliu, siqiwu, presnick}@umich.edu

Abstract

YouTube provides features for users to indicate disinterest when presented with unwanted recommendations, such as the "Not interested" and "Don't recommend channel" buttons. These buttons are intended to allow the user to correct "mistakes" made by the recommendation system. Yet, relatively little is known about the empirical efficacy of these buttons. Neither is much known about users' awareness of and confidence in them. To address these gaps, we simulated YouTube users with sock puppet agents. Each agent first executed a "stain phase", where it watched many videos of an assigned topic; it then executed a "scrub phase", where it tried to remove recommendations from the assigned topic. Each agent repeatedly applied a single scrubbing strategy, either indicating disinterest in one of the videos visited in the stain phase (disliking it or deleting it from the watch history), or indicating disinterest in a video recommended on the homepage (clicking the "not interested" or "don't recommend channel" button or opening the video and clicking the dislike button). We found that the stain phase significantly increased the fraction of the recommended videos dedicated to the assigned topic on the user's homepage. For the scrub phase, using the "Not interested" button worked best, significantly reducing such recommendations in all topics tested, on average removing 88% of them. Neither the stain phase nor the scrub phase, however, had much effect on videopage recommendations (those given to users while they watch a video). We also ran a survey (N = 300) asking adult YouTube users in the US whether they were aware of and used these buttons before, as well as how effective they found these buttons to be. We found that 44% of participants were not aware that the "Not interested" button existed. Those who were aware of it often used it to remove unwanted recommendations (82.8%) and found it to be modestly effective (3.42 out of 5).

1 Introduction

YouTube is the world's largest long-form video sharing platform, with users watching a billion hours of YouTube's content every day (YouTube 2017). In recent years, the YouTube recommendation algorithm has come under increased scrutiny for its role in promoting conspiracy theories (Tomlein et al. 2021), radical content (Hosseinmardi et al. 2021, 2024), and Alt-Right ideology (Lewis 2018; Ribeiro et al. 2020). Studies have found that watching such content can lead to its continual and sometimes increased promotion (Ledwich and Zaitsev 2020; Hussein, Juneja, and Mitra 2020). Besides societally-harmful information, YouTube has also been known to make unwanted recommendations to individuals, sometimes in the form of offensive, triggering, or outrageous videos (Mozilla 2019; Haroon et al. 2023).

In the context of both individual and societal reasons for users to better tailor their personalized content on YouTube, the platform provides several buttons, such as "Not interested" and "Don't recommend channel", which allow users to express disinterest in a specific video or channel and alter their recommendation feeds accordingly (Burch 2019; Cooper 2023). These buttons exist among a variety of platform features, such as "Disliking" videos and deleting videos from one's watch history. All of those buttons may help users eliminate certain content from their feeds.

However, relatively little is known about the efficacy of these buttons in practice, nor about users' awareness of and confidence in them. To address these gaps, this work investigates how well simulated YouTube users (agents) can populate their recommendation feeds with content from a certain topic (the "stain" phase), as well as the ability for the recommendations of that topic to be removed by using a strategy to indicate disinterest (the "scrub" phase). We selected four topics: *Alt-Right*, *Antitheism*, *Political Left*, and *Political Right*. These topics are particularly interesting because, while plenty of literature exists on how one can get recommended more of these topics, little is known about removing them. Also, each topic is a realistic one that some users would no longer want to see. We examined six strategies (*Watch neutral*, *Dislike*, *Delete*, *Not interested*, *No channel*, and *Dislike recommended*), as well as a *None* control strategy. Lastly, we conducted a complementary survey study to understand real users' awareness of and experience with those scrubbing strategies on YouTube.

The main findings are as follows:

- Watching a topic increased its presence on the homepage, though the stain never covers more than half of the recommendations. Watching a topic had less effect on the recommendations shown on videopages.
- For scrubbing a topic from the homepage, the most effective action was clicking the "Not interested" button on a recommended video. In contrast, none of the scrubbing

actions significantly reduced the number of recommendations of videos on the topic shown on videopages.

- Nearly half of survey respondents were not aware of the most effective feature (pressing "Not interested"). Those who were aware of it used it frequently, and perceived it to be less effective than the "Don't recommend channel" button, contrary to our findings from the audit study.

2 Background and Related Work

2.1 Locations of Relevant YouTube Features

There are three main YouTube pages that are relevant to our study: the homepage, videopage, and watch history page. The *homepage* is the landing page for users upon entering the platform.[1] It presents recommendations in a grid format. If a user is logged in, the content will be personalized to their account. Each recommendation contains a dropdown menu, where users are presented with two relevant buttons: the "Not interested" and "Don't recommend channel" buttons. These allow the users to (ostensibly) indicate disinterest with respect to specific recommendations.

The *videopage* is the page users see while watching a video. Recommendations are given in a right-hand sidebar. The feature on this page that is relevant to our study is the "Dislike" button, indicated by a thumbs-down symbol.

Finally, the *watch history page*[2] (or *watch history* for short) is the page that displays a log of the user's previously-watched videos. This page is only available to users who are logged in. Relevant to our study is the option for users to delete specific videos from their watch history. Specifically, the "Delete" button is "X" located on the upper-right hand corner of each video in the log.

We note that these pages and features were accurately described as of the time of data collection (August 2022) and the paper writing (January 2023). Since the platform often undergoes changes to its user experience and interface design, they may be outdated.

2.2 Sock Puppet Algorithm Audits

Our study takes a sock puppet algorithm audit approach. Several prior works quantifying the YouTube recommender system's role in promoting and removing unwanted content also use the algorithm auditing approach. So we begin with a review of this method.

An algorithm audit "is a method of repeatedly and systematically querying an algorithm with inputs and observing the corresponding outputs in order to draw inferences about its opaque inner workings" (Metaxa et al. 2021). Algorithm audits are a tool for researchers to investigate algorithms whose code and data are shielded from the public.

One type of algorithm audit is the sock puppet approach (Tomlein et al. 2021; Haroon et al. 2023; Hosseinmardi et al. 2024). Sock puppet audits use code scripts to create simulated users. These fake users – also called "agents" – interact with the platform or algorithm of interest as if they were real users. In the meantime, researchers record and compare the recommendations that the agents receive.

2.3 Recommender Systems' Role in Promoting Problematic and Unwanted Content

YouTube is one of the most popular video sharing platforms. In recent years, it has received increasing public scrutiny from journalists and academics alike in assessing its recommendations of problematic content.

The center of the platform's content dissemination is the recommendation engine, which plays an important role in helping users decide what to watch (Covington, Adams, and Sargin 2016; Wu, Rizoiu, and Xie 2019). From a vast, growing pool of videos on the platform, users are suggested what to watch next based on their previous interactions with YouTube and what content they will most likely engage with next (Zhao et al. 2019).

YouTube's recommendation engine has been theorized to promote problematic recommendations, which can broadly be split into two categories. The first is its propensity to suggest content that violates political and societal democratic ideals, such as promoting extremism and conspiracy theories (Lewis 2018; Tomlein et al. 2021) or creating political filter bubbles and causing radicalization (Tufekci 2018; Hosseinmardi et al. 2021). The second category of problematic recommendations are those that conflict with individual preferences. Many users have found recommendations on video sharing platforms to be personally offensive, triggering, violent, and outrageous (Mozilla 2019; Haroon et al. 2023), as well as conflicting with their sense of identity (Karizat et al. 2021), even if the video is completely legal and enjoyable for others.

Several YouTube algorithm audits have investigated the role of recommender systems in promoting such content, largely focusing on political ideologies and conspiracy theories (Tomlein et al. 2021; Haroon et al. 2023; Hosseinmardi et al. 2024). While previous studies often refer to this phenomenon as "filter bubbles", we choose to use the term "stain". This is because previous studies (as well as ours) find that topical recommendations rarely take up more than half of one's feed and never reach 100% despite watching many videos of that topic, and we would like to avoid the misleading interpretation of the term "bubble" as being completely surrounded by (i.e., having 100% of) recommendations of a certain topic.

Regardless of the term, studies have agreed that continued consumption of videos of a certain topic will lead to further (and sometimes increased) recommendation of that topic, on both the homepage and the videopage (Hussein, Juneja, and Mitra 2020; Papadamou et al. 2022; Haroon et al. 2023). Recommendations in search results, on the other hand, do not experience such personalization effects (Tomlein et al. 2021). Despite these findings, it is still unknown if the stain is comprised largely or completely of videos from previously-watched channels, or if YouTube introduces new channels yet to be watched by the user that contain videos with content on the same topic.

[1] https://www.youtube.com

[2] https://www.youtube.com/feed/history

2.4 User Controls to Remove Unwanted Content

Combined calls from academics and journalists alike to mitigate the YouTube recommender's role in problematic content consumption have contributed to recent platform changes. These include features that are intended to give users more control in tailoring their recommendations (Burch 2019; Cooper 2023). Other social media platforms such as TikTok and Instagram have also released and experimented with similar user controls (Ariano 2021; Meta 2022). Such features may improve user satisfaction in online spaces mediated by recommender systems, especially on systems (such as YouTube) that receive much of their content viewership from users watching recommender-suggested content.

However, compared to what is known about the prevalence of unwanted content on YouTube, relatively little is known about these features that remove them or their efficacy. We review this literature here.

Algorithm audits on how to reduce recommendations Two experiments used an intuitive strategy to try to reduce content of a given topic: watching videos of a *different* topic. Tomlein et al. (2021)'s sock puppet audit found that agents were recommended less conspiratorial content after they watched many videos debunking conspiracy theories. Haroon et al. (2023)'s sock puppet audit found that a politically-biased recommendation feed could be "debiased" – or achieve similar amounts of left and right-leaning videos – by watching a diet of videos heavily featuring the ideology that was less prevalent originally.

These studies suggest that it is possible to remove some unwanted content from one's feed. However, the degree to which it can be done varies and is never 100%.

Further, we find it necessary to expand recommendation-reduction strategies beyond video-watching and towards platform-provided buttons. First, many such buttons are designed for the explicit purpose of removing unwanted content (e.g., the "Not Interested" button). Second, they may be much faster to perform: studies suggest a minimum watch time of 10 minutes is required to register significant changes to recommendations (Papadamou et al. 2022); meanwhile, pressing a button takes just seconds. Lastly, using buttons may avoid the side effect of infusing too much content from another topic to replace the unwanted topic.

Ricks and McCrosky (2022) provide the first quantitative study of such buttons on YouTube. They supplied users with a browser extension with a custom "Stop Recommending" button displayed on each video recommendation. Then, whenever it was clicked, it caused a press of a native platform button in the background, with users randomly assigned to different native platform buttons. Their results show that the native "Don't recommend this channel" button, which appears on recommendations, produced suggestions least similar to them.

Ricks and McCrosky (2022)'s study benefits from a large sample. Their field experiment design also presents distinct advantages, particularly external validity that a sock puppet audit cannot achieve. At the same time, we still find it valuable to perform a sock puppet experiment with a more controlled environment for two reasons. First, because users could press the custom "Stop Recommending" on any recommendation from any topic, the study was not able to identify the effects of the buttons for well-defined topics. Second, there are possible confounds from uncontrolled user behavior, such as users watching videos similar to the ones they pressed "Stop Recommending" on, or cross-contamination between conditions where users clicked on native YouTube-provided buttons in addition to the custom Mozilla-provided one.

Users' relationship with user controls A few studies also used qualitative methods to understand users' experiences and perceptions of different strategies to remove unwanted content from their personal feeds.

Ricks and McCrosky (2022) surveyed and interviewed a subset of their participants from the quantitative arm of their study. They find that users take a variety of strategies to combat unwanted recommendations, generally find platform-provided features to be ineffective, and that achieving effective results requires sustained time and effort.

While these surveys and interviews solicit the breadth of strategies that users have to combat unwanted recommendations, the degree to which general YouTube users are aware of each platform-provided feature is still unknown. It is also unknown whether they use these features, even if they are aware of them. Such data is important because an effective feature may be moot if they are unknown or unused.

Smith, Bullen, and Huerta (2021) also examined YouTube user controls for altering recommendations. They found that the actions performed by such buttons were reactive (i.e., only useful *after* a user received an unwanted recommendation) and that the feedback provided to the user after clicking them was often unclear and vague. They also found that navigating to some of these features was difficult, which could limit users' ability to use them.

3 Research Questions

Previous studies of "filter bubbles" on YouTube found that recommendations of a given topic can increase as a result of watching videos of that topic, but we do not know whether these recommendations are from channels the user has watched before, or whether they are new channels that YouTube finds similar. Such a breakdown would add to the knowledge of YouTube's role in promoting unwanted content by quantifying how much YouTube is "inferring" this content rather than simply suggesting content from previously-watched channels. Also, confirming the general result of increasing topical recommendations motivates our next study phase, which attempts to remove them.

Thus, we first address the question, **how responsive are YouTube recommendations to watching many videos of the same topic?** (RQ1) In particular, do they recommend more videos of the same topic, and if so are they from channels that users watched up to that point or are they new ones? Do the results vary for different topics? We study four topics whose prevalence on YouTube has been previously studied: *Alt-Right*, *Antitheism*, *Political Left*, and *Political Right* (motivated and described in Section 4.1).

We are also interested in the effects of platform features in removing unwanted recommendations. While a previous study investigated their usage "in the wild", the effects of each feature, uncontaminated by usage of other features, on topics that are well-defined, is still unknown. Such questions are worth answering because YouTube users in general could benefit from knowing what are the most effective strategies for removing unwanted content, specifically their effect on specific topics that they may dislike.

Thus, we ask, **how responsive are YouTube recommendations to repeatedly performing a particular strategy to try to remove unwanted videos of a topic?** (RQ2) Are they different between videopage and homepage? How much content is removed from similar channels that are not explicitly interacted with? Do they vary topic to topic? We identified six such strategies, such as pressing the "Not interested" button, and listed them in Section 4.1.

Finally, it is unknown how many YouTube users are aware of each platform feature, how many utilize them, and how effective the users find them to be. This information is important because effective strategies may be moot if users do not know about them, and because users should be both using effective strategies and finding them to be effective.

Therefore, we lastly ask, **what are real users' experiences with the platform features that we test in RQ2?** (RQ3) With respect to each platform feature, we designed a survey study to ask how many participants are aware of it, what percentage use it to remove unwanted recommendations (given they are aware), and how effective participants find it to be (given they are aware and have used the feature to try to amend the situation).

4 Sock Puppet Study

4.1 Sock Puppet Design

We take a sock puppet algorithm audit approach to examine how suggestions from certain topics can both be populated onto and removed from one's personal recommendation feed. Broadly, our agents first purposely populate their feed with videos from this unwanted topic ("stain phase"); Then, they take on one of a variety of strategies to try to eliminate such videos from being recommended ("scrub phase"). We collect data on how recommendations change throughout these phases in order to characterize the recommendation system's response to these various interactions.

Video topics Each topic is operationalized as a list of channels collected by previous researchers who have studied that topic on YouTube. They are used in our experiment in two ways. First, agents watch videos from the channel lists during the stain phase. Next, during the scrub phase, some strategies cross reference their homepage recommendations with the assigned topic's channel list to determine whether and which one to indicate disinterest on.

- *Alt-Right*: The most extreme group of the Alternative Influence Network, a loosely-defined community of YouTube channels that are defined by their opposition to mainstream media (Ricks and McCrosky 2022). The Alt-Right promotes white nationalism in the face of an increasingly diverse US population, and is often openly anti-semitic (ADL 2019). YouTube channels of the Alt-Right were first collected by Lewis (2018) through a snowball sampling method, and subsequently augmented by Ribeiro et al. (2020) and Chen et al. (2022).
- *Antitheism*: Collected by Ledwich and Zaitsev (2020). It is "the self-identified atheist who is also actively critical of religion".
- *Political Left*: Collected by Wu and Resnick (2021). They include local news, talk shows, and magazines. We use the US political left channels, which takes similar views among various issues of political significance, such as climate change.
- *Political Right*: Same as above, but with the US political right channels.

Scrubbing strategies The name and operation of each scrubbing strategy are listed below. Each agent is assigned one strategy, and performs it repeatedly during the "scrub phase" of the sock puppet run.

- *None* (control): Load the homepage, then do nothing except refresh the homepage.
- *Watch neutral*: Load and watch a video from mainstream, politically neutral news outlets as defined by the fact-checking organization Media Bias/Fact Check.[3]
- "History-based" strategies
  - *Dislike*: Load a previously-watched video from the stain phase and click the "Dislike" button.
  - *Delete*: Load the watch history and click "Delete" on the most recently-watched stain video.
- "Recommendation-based" strategies. Load the homepage. If there does *not* exist any recommended video on the homepage from a channel in the channel list, then just refresh. However, if such a video exists, do the following to the first such video:
  - *Not interested*: click the "Not interested" button and refresh the homepage.
  - *No channel*: click the "Don't recommend channel" button and refresh the homepage.
  - *Dislike recommended*: click on the video and dislike it (agents do not stay to watch the video), then return to the homepage.

The "watch neutral" strategy attempts to ignore the current issue by watching videos from a different topic, and most resembles the intervention strategies of related studies (Haroon et al. 2023; Tomlein et al. 2021). We call dislike and delete strategies "history-based" because they act on videos that the agents watched during the stain phase. We call the final three strategies "recommendation-based" because they are performed with respect to recommended videos.

[3] https://mediabiasfactcheck.com/

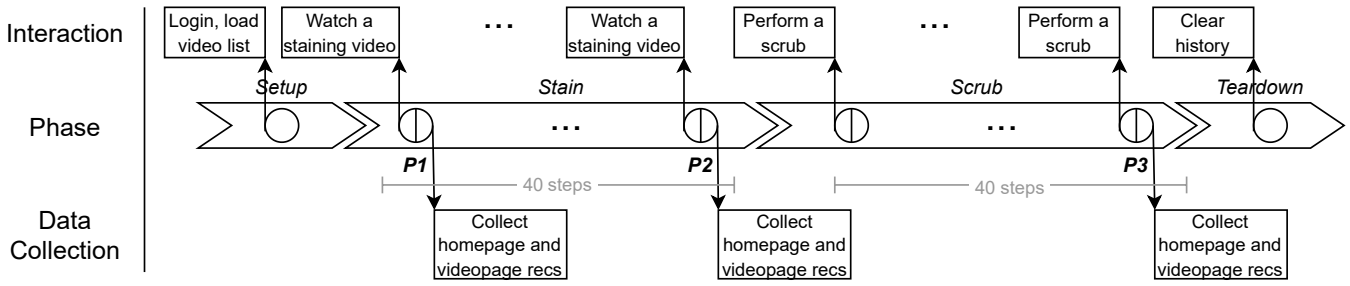

Figure 1: The interactions and data collection that a sock puppet agent performs on the YouTube platform, separated by phase.

Sock puppet phases and data collection A sock puppet agent follows the following process. After logging in to a YouTube account, an agent performs the "stain phase", where it watches 40 videos, for up to 30 minutes each,[4] from a "stain video list" which are sampled from the channel list belonging to its assigned topic. Next, it performs the "scrub phase", where it executes its assigned scrubbing strategy 40 times. Lastly, the agent clears its entire YouTube activity through Google's MyActivity page,[5] in order to leave a clean history for the next audit to start (Tomlein et al. 2021). This includes clearing all revertible actions made during its run, such as clicks of the "Dislike", "Not interested", "Don't recommend", and "Delete from watch history" buttons.

Our agents use web-scraping methods to collect the top 10 recommendations from the homepage and videopage at three strategic points:

- P1: The beginning of the stain phase.
- P2: The end of the stain phase.
- P3: The end of the scrub phase.

Because the video being watched during videopage collection may itself have an effect on recommendations, each agent always loads the *same* video at all three collection points (P1, P2, P3). This video is from the stain video list.

Algorithm 1 provides an overview of a sock puppet agent's interactions with the platform. A visual flowchart of this process is given in Figure 1. The code implementation is available on the project webpage.[6]

We now describe the configurations of agents for the overall experiment. For each topic we tested seven strategies, each five times, resulting in (7 * 5 =) 35 agents. All 35 agents of a given topic were run in parallel in order to deal with recommendation noise that may arise from having agents make queries at different times.

For a given topic, we also drew five stain video lists, and assigned each list to exactly one agent within every strategy tested. Doing so assures that agents of the same strategy watch different sets of staining videos, boosting generalizability of the strategy effects, while simultaneously assuring that agents of different strategies watch, in total, the same videos, enabling comparability between strategies.

Additionally, each of the 35 agents have their own Google Accounts so that the platform can track their viewing habits and personalize content to each agent, and so that we can more closely simulate real users' experience with the platform. Logging into an account also grants access to buttons and features that are only available to users that are logged in (e.g., the "Not interested" button).

Our agents ran in a Google Chrome browser with adblocker installed. They had Google accounts with birthdays set at an arbitrary 5/5/1990, a gender selection of "Rather not say", and asexual names (e.g., "Tandy"). We also address the potential biases from location effects, which would occur if queries were made to the platform from different locations, or from (different accounts in) the same IP address. Thus, all agents are created and live in the same AWS Region of Ohio (US-East-2), but make queries from individual IP addresses.

Out of 140 sock puppets released over the course of five days in August 2022, 139 sock puppets ran successfully. Agents collected a total of 8330 recommendations.

4.2 Data Annotation

Our agents collected many recommendations during their runs. We would like to label them for whether they belong to the topics that the agents were assigned to (what we call "stain"), in order to quantify how well (1) the stain phase

Algorithm 1: Agent

```
Log into YouTube
Collect homepage recs                                   ▷ P1
Collect videopage recs from the first stain video       ▷ P1
for i ∈ [2 . . . 40] do                                 ▷ stain phase
    Watch a video from stain video list up to 30 minutes
end for
Collect homepage recs                                   ▷ P2
Collect videopage recs from the first stain video       ▷ P2
for i ∈ [1 . . . 40] do                                 ▷ scrub phase
    Perform assigned scrubbing strategy
end for
Collect homepage recs                                   ▷ P3
Collect videopage recs from the first stain video       ▷ P3
Clear YouTube activity (cancel all revertible actions)
```

[4] An alternative is to stay for the median watch time for videos with similar length, see Wu, Rizoiu, and Xie (2018)'s computation of relative engagement metric.

[5] https://myactivity.google.com/myactivity

[6] https://github.com/avliu-um/youtube-disinterest

		Alt-Right			*Antitheist*			*Political Left*			*Political Right*			Avg. relative change
		P1	P2	P3	P1	P2	P3	P1	P2	P3	P1	P2	P3	P2 to P3
Homepage	Total	2%	$\xrightarrow{*}$ 20%		11%	$\xrightarrow{*}$ 37%		6%	$\xrightarrow{*}$ 28%		9%	$\xrightarrow{*}$ 29%		
	None		20%	22%		32%	38%		20%	34%		32%	42%	+32%
	Watch neutral		12%	8%		34%	22%		32% $\xrightarrow{*}$	18%		26% $\xrightarrow{*}$	16%	-38%
	Delete		20% $\xrightarrow{*}$	0%		32%	24%		30% $\xrightarrow{*}$	0%		24% $\xrightarrow{*}$	4%	-77%
	Dislike		20%	4%		30% $\xrightarrow{*}$	10%		24%	28%		32% $\xrightarrow{*}$	16%	-45%
	Not interested		18% $\xrightarrow{*}$	0%		42% $\xrightarrow{*}$	2%		28% $\xrightarrow{*}$	10%		26% $\xrightarrow{*}$	2%	-88%
	No channel		28% $\xrightarrow{*}$	10%		44% $\xrightarrow{*}$	22%		30% $\xrightarrow{*}$	16%		36% $\xrightarrow{*}$	24%	-49%
	Dislike rec.		20% $\xrightarrow{*}$	8%		42% $\xrightarrow{*}$	14%		32%	24%		26%	34%	-30%
Videopage	Total	6%	5%		26%	$\xrightarrow{*}$ 37%		41%	42%		12%	$\xrightarrow{*}$ 28%		
	None		6%	4%		40%	34%		38%	34%		20%	30%	-2%
	Watch neutral		4%	6%		28%	6%		42%	46%		30%	34%	-1%
	Delete		2%	2%		38%	34%		42%	34%		32%	16%	-20%
	Dislike		12%	14%		40%	34%		44%	46%		22%	42%	+24%
	Not interested		4%	2%		40%	48%		44%	44%		32%	30%	-9%
	No channel		0%	0%		32%	30%		44%	40%		34%	36%	-2%
	Dislike rec.		6%	2%		42%	34%		40%	50%		24%	22%	-17%

Table 1: Stain percentage values, and Wilcoxon signed-rank tests, for the homepage (top) and videopage (bottom), at data collection points P1, P2, P3, for each strategy (row) and topic (column). Values in the "Total" row combine stain for all agents of that topic. The asterisked arrow ($\xrightarrow{*}$) between two values indicates significant changes at the 0.05 significance level.

worked to populate agent' recommendations with stain, and (2) how well the scrub phase worked to remove it.

We adopted an iterative strategy in labeling the recommended channels. We first developed an initial annotation codebook by surveying prior research. Next, we randomly sampled 50 channels for each topic. Two authors who had extensive experience in studying political polarization and YouTube platforms independently labeled those channels by following the codebook. The preliminary inter-rater reliability (IRR), measured by Cohen's kappa, was 0.648, 0.728, 0.634, 0.563 for *Alt-Right*, *Antitheism*, *Political Left*, *Political Right*, respectively, demonstrating substantial agreement. The two raters discussed every disagreed case to reach consensus and updated the codebook whenever needed.

The two raters then each labeled all the remaining channels. The final IRR kappa scores were 0.660, 0.822, 0.854, and 0.945, respectively. The raters also discussed all disagreed cases and resolved disagreement. The final annotation codebook, annotation results, and IRR calculation, along with sock puppet implementation are publicly available on the project webpage.[7]

4.3 Result 1: Stain Phase

In this subsection, we answer our questions posed in RQ1.

Effects of stain phase We wanted to know whether our agents experienced a significant change in stain (the percentage of recommendations of their assigned topic) after the stain phase. To address this question, we compared the stain of our agents at P1 with those at P2 for each topic, in both the homepage and the videopage. To determine whether changes were significant, we chose the Wilcoxon signed-rank test because (a) the data was non-normal; (b) the comparison before and after the "stain phase" treatment was a paired test. Results are given in Table 1 (P1 to P2).

[7]https://github.com/avliu-um/youtube-disinterest

On the homepage, we find that all topics experienced significant increases in stain as a result of the stain phase. *Antitheism* received the most stain at P2 (37%), while *Alt-Right* received the least (20%). In contrast, the videopage demonstrated significant changes in stain only on *Antitheist* and *Political Right*. *Alt-Right* actually showed a slight decrease from P1 to P2. Despite this, a non-zero stain still existed in the videopage – absolute percentages at P1 varied between 5% for *Alt-Right* and 42% for *Political Left*. We lastly remark that, across both homepage and videopage, and across topics and strategies, stain never reached more than half.

These findings set us up well for the scrub phase, because it assures that our agents will indeed have stain to remove when they perform their scrubbing strategies.

Stain from watched channels vs. new channels We wanted to examine the stain in P2 with more granularity. Specifically, how much of it was from channels the agent had explicitly watched before, and how much was from channels that the agents had never explicitly watched before?

To answer this question, we categorized all recommendations at point P2 as "off-topic", "on-topic watched-channel" (i.e., the agent that collected this recommendation had already watched a video from the same channel during the stain phase), or "on-topic new-channel" (i.e., the agent had not watched any videos from the same channel up to that point). Then, for each topic, we found the percentage of recommendations belonging to each category. We report these ratios for the homepage and videopage in Table 2.

On the homepage, we find that *Political Left* had the most recommendations from new channels (26%), as well as the highest ratio of recommendations from new channels to those from watched channels (13:1). On the other hand, the *Alt-Right* had the smallest new-channel percentage, both in absolute terms (5%) and as a ratio of watched-channel percentage ($\sim$1:3).

	Alt-Right			*Antitheism*			*Political Left*			*Political Right*		
	Off-topic	On-topic watched	On-topic new	Off-topic	On-topic watched	On-topic new	Off-topic	On-topic watched	On-topic new	Off-topic	On-topic watched	On-topic new
Homepage	79%	16%	5%	60%	18%	23%	72%	2%	26%	70%	9%	21%
Videopage	95%	5%	0%	63%	9%	29%	58%	13%	29%	72%	19%	9%

Table 2: Stain at P2 split into categories of off-topic, on-topic watched-channel, and on-topic new-channel (we omit the word "channel" to save space), for each of homepage and videopage (row) and each topic (column).

	Alt-Right			*Antitheism*			*Political Left*			*Political Right*		
	Off-topic	On-topic scrubbed	On-topic new	Off-topic	On-topic scrubbed	On-topic new	Off-topic	On-topic scrubbed	On-topic new	Off-topic	On-topic scrubbed	On-topic new
Watch neutral	90%	8%	2%	74%	14%	12%	78%	2%	20%	82%	6%	12%
Delete	96%	0%	4%	74%	18%	8%	100%	0%	0%	94%	2%	4%
Dislike	92%	2%	6%	84%	10%	6%	72%	0%	28%	84%	2%	14%
Not interested	100%	0%	0%	96%	0%	4%	90%	0%	10%	94%	0%	6%
No channel	84%	0%	16%	76%	2%	22%	84%	0%	16%	76%	0%	24%
Dislike rec.	90%	2%	8%	80%	0%	20%	76%	0%	24%	64%	4%	32%

Table 3: Stain at P3 on the homepage split into categories of off-topic, on-topic scrubbed-channel, and on-topic new-channel (we omit the word "channel" to save space), for each of strategy (row) and each topic (column).

On the videopage, *Political Left* and *Antitheist* had the highest absolute percentage of recommendations from new channels (29%), and *Antitheist* had the highest ratio of new channel recommendations to those from watched channels ($\sim$3:1). By contrast, the *Alt-Right* agents received no recommendations from new channels (0%) while all other topics received at least 9%.

These findings suggest that the YouTube recommendation system sometimes plays a role in providing stain to the user by not only suggesting content that is from the same channel, but rather by inferring and providing that from different but similar channels.

4.4 Result 2: Scrub Phase

In this subsection, we answer our questions posed in RQ2.

Effects of scrub phase We wanted to know whether our agents could remove stain after the scrub phase. To address this question, we compared the stain of our agents at P2 with those at P3, for each topic, in both homepage and videopage. Again, our data was non-normal and paired, so we ran Wilcoxon signed-rank tests to see whether stain decreased significantly. Results are again in Table 1 (P2 to P3).

On the homepage, *Not interested* and *No channel* were the only strategies that significantly reduced the amount of stain across all topics. Comparing average relative changes between P2 and P3, *Not interested* wins out (-88%). One strategy successfully scrubbed three out of four topics (*Delete*), while two strategies were successful in two out of four topics (*Dislike recommendation* and *Watch neutral*). On the other end, the *None* strategy did not produce any significant effect, which was expected because it was our control strategy. On the videopage, we did not find any significant scrub phase effects.

Stain from scrubbed channels vs. new channels We wanted to know whether scrubbing strategies removed stain in general, or if they only removed the subset of channels the agent had explicitly scrubbed up to that point.

To answer this question, for all recommendations at point P3, we categorized them as either "off-topic", "on-topic scrubbed-channel" (i.e., the agent that collected this recommendation had already scrubbed a video from the same channel during the scrub phase), or "on-topic new-channel" (i.e., the agent had not scrubbed a video from the same channel up to that point). Notice that these categories are analogous to watched/new categories made for P2 in Section 4.3.

Then, for each topic/strategy pairing, we found the ratio of recommendations at P3 that belonged to each category. We report these ratios for the homepage in Table 3. We did not examine the videopage because it did not experience any significant changes in this phase. The *None* strategy was excluded because no videos were scrubbed.

Categorizing recommendations this way reveals that scrubbing strategies behaved differently in removing unwanted recommendations. For instance, in three out of four topics, at least half of the *Watch neutral* strategy's remaining stain at P3 was from scrubbed channels. On the other hand, *No channel* rarely left recommendations from scrubbed channels (0-2%); most stain remaining after using this strategy was from new channels. The behavior of *No channel* agrees with many user perceptions of the "Don't recommend channel" button (Ricks and McCrosky 2022), and matches an intuitive interpretation of the button name.

5 Survey Study

5.1 Survey Design

In the sock puppet section of our work, we collected data from simulated users' interactions with YouTube to quantify how platform features may help remove unwanted recommendations. In this section, we want to understand better the relationship between real users and these features. We ran a survey to determine this.

Survey overview We first asked whether respondents had experienced getting unwanted recommendations before. Respondents were specifically asked whether they have experienced this scenario before: "*You are browsing YouTube, and notice videos recommended to you that you would rather not have recommended (because they are offensive to you, triggering, not safe for work, or some other reason)*". The buttons we consider are "Delete" (delete a video from watch history), "Dislike", "Not interested", and "Don't recommend channel". We asked for respondents' experiences with these buttons, with respect to three constructs:

- *Awareness*: Before taking this survey, were you aware this button existed?
- *Usage*: Have you used this button to remove unwanted recommendations?
- *Belief in efficacy*: Recall the times when you used this button to remove unwanted recommendations. How effective do you think it was? Please rate from 1 (not at all effective) to 5 (completely effective).

Only those who have experienced unwanted recommendations before and were aware of the buttons were asked to report on their real usage, while others were given a hypothetical question ("*If* you had known this button existed, would you have used it?"). Furthermore, only those who were (1) experienced, (2) aware of button, and (3) have used the button were asked to report their belief in its efficacy in removing recommendations, while others were given a hypothetical question ("*If* you had used this button for this scenario, how effective do you think would you find it?").

Survey implementation We recruited 300 participants from the survey recruitment platform Prolific, and ran the survey on the survey delivery platform Qualtrics. We selected participants that had used YouTube before, were adults (18+), and resided in the US. Participants were paid $15 an hour. The University of Michigan Health Sciences and Behavioral Sciences Institutional Review Board has determined that this research is exempt from IRB oversight (Study ID: HUM00224551).

Surveys were pretested with colleagues. We emphasized honest rather than "right" answers so that respondents would not be tempted to "please" us by saying they knew about a button when in reality they did not (Paolacci and Chandler 2014). We included screenshots of buttons so that they didn't have to know them by name. Since attention checks are important to maintaining experimental validity, we also implemented three of them throughout the survey to make sure the respondents were focusing on and comprehending the survey questions. At three separate points, we gave them a question whose format was identical to that of others and instructed them to select a specific choice. For example, *"Please select 'Dislike' in the choices below"*.

We filtered responses by eliminating those from people who failed any of the three attention checks. Respondents were paid regardless of whether they passed or failed attention checks. Then, we eliminated responses from anybody who answered "not sure" to our questions.

5.2 Survey Analysis Methods and Results

In this subsection, we answer the questions posed in RQ3.

In total, our survey received 274 responses from those who passed all three attention checks. However, our respondent sample did not immediately generalize to a more general population. Thus we used post-stratification, a popular statistical method that adjusts estimates on non-probability samples (Salganik 2019), to generalize our results to the adult YouTube-using population in the US.

To perform post-stratification, we divided our respondents into binary genders and age buckets (roughly 20 years apart), making a total of eight subgroups. We then made estimates of each subgroup's prevalence in the target population by combining Census data on age/gender subgroups (Duffin 2022) and PEW data on the percentage of each subgroup that use YouTube (Auxier and Anderon 2021). Comparing our survey sample's distribution among subgroups with that of the target population revealed that our sample skewed young: Among usable responses, we routinely over-sampled the 18-45 subgroups and under-sampled 65+ ones. Fortunately, post-stratification corrects this bias.

We report the answers in Table 4. *Awareness* percentages were calculated by aggregating answers from all respondents that passed our attention checks. For *Usage*, we restricted our calculation to those who were both aware that the feature existed, and had experienced having unwanted recommendations. *Belief in efficacy* was the most restricted because we only wanted ratings from those who would be well-informed of its effects from personal usage: Only those who experienced unwanted recommendations, were aware the button existed, and used that button to try to resolve the issue, were considered. These population restrictions are applied to both the table and our discussion of results.

Moving onto results, we find for *Awareness* that survey respondents were most aware of the "Dislike" button's existence (93.94%). "Don't recommend channel" was the least well-known (35.37%). As for *Usage*, they favored "Not interested" (82.83%) and "No channel" (80.53%) to remove unwanted recommendations when they experienced it. "Dislike" was the least used button (37.75%). Looking at *Belief in efficacy*, users found "Delete" (3.76), "Not interested" (3.42), and "No channel" (4.10) all more effective than the "Dislike" button (2.52).

These findings suggest that users do not use the "Dislike" button to remove unwanted recommendations, despite most knowing about its existence. Respondents' intuitions about this button match our empirical findings: We saw in Section 4.3 that the *Dislike* and *Dislike recommendation* strategies both reduced less stain compared to other scrubbing strategies (*Delete*, *Not interested*, *No channel*). Meanwhile,

	Awareness	*Usage*	*Belief in efficacy*
Delete	51.41% ± 7.55% [248]	53.64% ± 13.52% [110]	3.76 ± 0.25 [48]
Dislike	93.94% ± 3.90% [263]	37.75% ± 7.90% [226]	2.52 ± 0.25 [62]
Not interested	56.03% ± 6.63% [258]	82.83% ± 8.23% [156]	3.42 ± 0.32 [122]
Don't recommend channel	35.37% ± 5.85% [255]	80.53% ± 9.58% [111]	4.10 ± 0.33 [88]

Table 4: Results from user survey. For each button (row), we report both the mean value and 95% confidence interval for constructs of interest (column), estimated from post-stratification. Sample sizes are given in brackets. Note that *Awareness* and *Usage* are in the range of 0 to 100% while *Belief in efficacy* is in a scale of 1 to 5.

the button for our (empirically) most effective scrubbing strategy – "Not interested" – was highly used by respondents who knew of it. However, awareness was not universal: almost 44% of survey respondents were unaware of its existence.

6 Discussion

We performed an algorithm audit of the YouTube recommendation system to test whether one could remove unwanted content from their feed. We paired our audit with a survey to understand whether users actually knew these buttons existed, used them, and believed them to be effective.

In our audit, we found that the stain phase produces a significant increase in stain in the homepage across all topics. We also saw that stain at P2 never reached more than half of recommendations in either the homepage or the videopage. These results confirm our initial suspicion that watching many videos from a given topic does not completely "surround" the user with topical recommendations, contrary to what the term "filter bubble" would suggest. This motivated our usage of the altenative term "stain".

Continuing to results from the stain phase, we broke down stain at P2 into those from channels watched before and those from channels not watched before. We found that their prevalence varied based on topic.

Both types of stain were present, but to varying degrees depending on topic. On the one hand, *Political Left* received the most stain from new channels for both the homepage and videopage, demonstrating that the platform had a notion of topical similarity by "inferring" other channels from the political left that the user may like. On the other hand, the *Alt-Right* received the least recommendations from new channels, for both the homepage and videopage. This finding is interesting given YouTube's recent public promises to curb misinformation and conspiracy theories (YouTube 2019), especially "harmful" ones such as Q-Anon (YouTube 2020), as well as a shift in company-wide attention towards stopping home-grown, right-wing extremism from spreading on its platform (Bergen 2022). While the lack of recommendations from new *Alt-Right* channels supplied to agents who watch that content could be evidence of YouTube operationalizing its promises, we cannot formally tell the difference between that and a general lack of *Alt-Right* videos remaining on the platform today.

Moving onto the scrub phase, we compared different scrubbing strategies and found that *Not interested* was the most effective one on the homepage: It produced significant decrease in stain across all topics, and using it resulted in the greatest average decrease in stain from P2 to P3 across topics (-88%). This strategy performed well in removing stain from both channels were explicitly scrubbed as well as similar ones the agent didn't interact with. Thus, users who would like to remove recommendations from any channels belonging to an unwanted topic should use this strategy.

In contrast to homepage findings, we found that the videopage never experienced significant effects from the scrub phase. At a cursory glance, it seems that our results disagree with Tomlein et al. (2021)'s finding. In their study, agents *could* significantly reduce conspiratorial recommendations on the videopage by watching many videos debunking the conspiracy theory. However, upon further inspection it should be noted that in fact we have two separate experiments. Whereas agents in our study collected videopage recommendations from a video at P3 that was the same as that used in P2, their study used a video at P3 that was the semantic *opposite* of that of P2. Specifically, Tomlein et al. (2021)'s bots collected them from a video *promoting* agents' assigned conspiracy theory at P2, and then collected them from a video *debunking* it at P3.

Combining our findings with those from Tomlein et al. (2021) suggests that videopage recommendations may be influenced more by the video that is playing at the time recommendations are shown, than by any prior interactions with the system. The implication for users is that they should not expect any scrubbing strategies to save them from further recommendations of an unwanted topic if they keep watching videos of that topic. Rather, they may want to stop watching content from that topic altogether.

Lastly, we wanted to know how users interacted with platform features in their daily YouTube usage. We found that US adult YouTube users were most aware of the "Dislike" button, yet more empirically effective strategies, such as "Not interested", were lesser known. Those who knew the "Not interested" button existed used it at a higher rate and saw it as more effective than those who knew of "Dislike".

Put together, our sock puppet and survey findings suggest that if YouTube wants to allow users to more effectively remove unwanted recommendations, it should make its effective platform-features for doing so more broadly known to the general YouTube population. Doing so would not only benefit users' experience; it would also be in the best interest of the platform because allowing users to have more agency to tailor algorithmic decisions to their preferences can build and maintain their trust in the system (Ekstrand et al. 2015), as well as increase overall satisfaction (Shin

2020). One implication for platform designers is that they should make buttons such as "Not interested" more widely known by increasing their discoverability on the website. To that end, Ricks and McCrosky (2022) provide a blueprint. In their experiment, they found that when their users were displayed "Don't recommend this" buttons prominently and clearly on recommendation title cards, instead of being hidden behind a menu or requiring navigation away from their current page, they were more than twice as likely to use it (Ricks and McCrosky 2022).

While this study demonstrates the benefits of YouTube's user controls, there still exist challenges to its uptake to remove unwanted recommendations. First, we note that these controls could be used to create digital media environments that run counter to democratic norms of diversity and breadth of perspectives. Thus, policy makers should pay attention to the potential for user agency to further limit their capacity for and consumption of cross-cutting content.

Second, as our survey highlights, knowledge of these buttons is still an issue. Much of the general YouTube-using public was not aware that the "Not interested" button exists, for example. Even more troubling was that even those who experienced unwanted recommendations recently – thus having ample motivation to discover content removal tactics – still had not become aware of the button.

Third, user interaction flows from these buttons may violate design principles in a way that limits users' ability to fully understand and anticipate the effects of different user controls (Smith, Bullen, and Huerta 2021). For instance, they found that users were not fully aware of their effects on recommendations and account settings, and so they shied away from using them at all. Further compounding users' hesitation to take up these features is the perception that some of their effects are irreversible.

Lastly, the actions that these user control buttons allow are responsive, rather than proactive. Users respond to a poor recommendation by eliminating it, rather than asking YouTube to tailor their recommendations before they see it. Thus, the worrisome effects of misinformation, toxicity, and offensiveness may have already taken their harmful course by the time the user decided to eliminate them. Therefore, these features cannot be seen as a substitute for diligent and thorough content moderation by the platform.

Ethical statement We now discuss the ethical concerns of our study. First, since our sock puppets are computer scripted, we do not risk making real users watch potentially harmful content, such as those from the *Alt-Right* channels. However, making the bots watch a lot of content from a given topic may still increase its prevalence on YouTube by boosting its general popularity. Also, pressing the "Dislike" button on channels may cause them to be demoted in recommendations, limiting YouTube creators' ability to generate advertising revenue.

While this is a possibility, we do not find these costs to outweigh the benefits of our study. First, we consider the potential cost to content creators of negative interactions with the system, such as pressing the "Dislike", "Not interested", "Don't recommend channel", and "Delete from watch history" buttons. Here we note that (a) our bots collectively injected up to 3 such interactions per video, which we expect is small compared to the number of "authentic" ones, (b) we cleared all the revertible actions caused by the audit when it exited its experimental runs, and (c) the average lifetime of an audit in this study is less than six hours, including both the stain phase and scrub phase. This means that not only is the effect of negative interactions small per video, it is also both short-lived and fully reversed.

Another potential cost of our study is that our audits would irreversibly alter two public metrics – total view count and total watch time. However, the costs are small because we do not expect to affect videos' and/or channels' overall prevalence by much because the number of views we are "artificially" introducing to the YouTube world are minuscule in relation to the number of "authentic" views that the videos have received.

Our study's findings are a benefit to all YouTube users alike, because they can inform users on how best to deal with and get rid of unwanted recommendations. We think these benefits outweigh the minimal harms.

As for the survey participants, we must make sure that they are not being put in the way of any harm. Because we did not ask them to view any actual content, but rather recall times in their life in which they had interacted with YouTube, we are only at the risk of having participants revisit potentially-triggering or traumatic events if they have had any. However, we include at our introduction of the survey a description of the survey, which describes the questions we wanted to ask them. Thus, if a participant expected such harms to occur to them, they would have not consented to the survey, and they would have been removed from the panel before seeing any questions.

7 Conclusion

With these results we conclude that different strategies to remove unwanted content on the YouTube platform work to different degrees, and from our tested strategies we found that using the "Not interested" button was a clear winner. However, this strategy has not seen widespread adoption among users. That is, while those who know about these effective buttons get to experience their effective behavior, 44% of adult YouTube users in the US are not aware that they exist. Thus, we join existing calls for YouTube to amplify more broadly the effective ways to remove unwanted recommendations on its platform.

Limitations Our findings add to a growing chorus of studies investigating problematic and unwanted recommendations on YouTube. However, our study is not without its limitations. First, we perform channel-level, rather than video-level labeling. Performing labeling at this level may result in labeling channels as a certain topic even if not all of its videos are of that topic (e.g. an *Alt-Right* channel sometimes posting music videos), or, conversely, labeling a channel as off-topic if just a few of its videos are topical (e.g. a science vlogger who occasionally talks about their journey to atheism). However, we are encouraged by the fact that other studies have taken this approach (Tomlein et al. 2021; Chen

et al. 2022). Analyzing channels as a whole is still important because they indicate a high number of videos of that topic, and users may be encouraged to subscribe to these channels even if not all videos in the channel are topical.

Another limitation is that of generalizing from the particular settings our sock puppets used. In particular, our sock puppets tested just four topics, using geolocation of the US East AWS center in August of 2021. While we found that certain strategies work in these conditions to remove recommendations, we cannot be sure it does in other conditions. New topics may be harder or easier to scrub due to more or less general interest in that topic in the broader YouTube ecosystem, respectively. Our code is publicly available so that we and other researchers may continue to understand more general effects of our scrubbing strategies.

Acknowledgments

This work was supported in part by the University of Michigan Center for Social Media Responsibility. We thank Rob Carleski for his efforts creating sock puppet infrastructure.

References

ADL. 2019. From Alt Right to Alt Lite: Naming the Hate. https://www.adl.org/resources/backgrounders/from-alt-right-to-alt-lite-naming-the-hate. Accessed: 2024-04-10.

Ariano, R. 2021. How to 'dislike' a TikTok video to make the app better understand what kind of content you want to view. https://www.businessinsider.com/guides/tech/how-to-dislike-a-tiktok. Accessed: 2024-04-10.

Auxier, B.; and Anderon, M. 2021. Social Media Use in 2021. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/. Accessed: 2024-04-10.

Bergen, M. 2022. YouTube Went to War Against Terrorists, Just Not White Nationalists. https://www.bloomberg.com/news/features/2022-08-30/youtube-s-video-purge-left-out-right-wing-extremism. Accessed: 2024-04-10.

Burch, S. 2019. YouTube Rolls Out New 'Don't Recommend' Feature. https://www.thewrap.com/youtube-rolls-out-new-dont-recommend-feature/. Accessed: 2024-04-10.

Chen, A. Y.; Nyhan, B.; Reifler, J.; Robertson, R. E.; and Wilson, C. 2022. Exposure to Alternative & Extremist Content on YouTube. https://www.adl.org/resources/reports/exposure-to-alternative-extremist-content-on-youtube. Accessed: 2024-04-10.

Cooper, P. 2023. How the YouTube Algorithm Works in 2023: The Complete Guide. https://blog.hootsuite.com/how-the-youtube-algorithm-works/. Accessed: 2024-04-10.

Covington, P.; Adams, J.; and Sargin, E. 2016. Deep neural networks for youtube recommendations. In *ACM RecSys*.

Duffin, E. 2022. Population of the U.S. by sex and age 2021. https://www.statista.com/statistics/241488/population-of-the-us-by-sex-and-age/. Accessed: 2024-04-10.

Ekstrand, M. D.; Kluver, D.; Harper, F. M.; and Konstan, J. A. 2015. Letting Users Choose Recommender Algorithms: An Experimental Study. In *ACM RecSys*.

Haroon, M.; Wojcieszak, M.; Chhabra, A.; Liu, X.; Mohapatra, P.; and Shafiq, Z. 2023. Auditing YouTube's recommendation system for ideologically congenial, extreme, and problematic recommendations. *PNAS*.

Hosseinmardi, H.; Ghasemian, A.; Clauset, A.; Mobius, M.; Rothschild, D. M.; and Watts, D. J. 2021. Examining the consumption of radical content on YouTube. *PNAS*.

Hosseinmardi, H.; Ghasemian, A.; Rivera-Lanas, M.; Horta Ribeiro, M.; West, R.; and Watts, D. J. 2024. Causally estimating the effect of YouTube's recommender system using counterfactual bots. *PNAS*.

Hussein, E.; Juneja, P.; and Mitra, T. 2020. Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube. *ACM CSCW*.

Karizat, N.; Delmonaco, D.; Eslami, M.; and Andalibi, N. 2021. Algorithmic Folk Theories and Identity: How TikTok Users Co-Produce Knowledge of Identity and Engage in Algorithmic Resistance. *ACM CSCW*.

Ledwich, M.; and Zaitsev, A. 2020. Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *First Monday*.

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

Meta. 2022. Testing More Ways to Control What You See on Instagram. https://about.fb.com/news/2022/08/testing-ways-to-control-what-you-see-on-instagram/. Accessed: 2024-04-10.

Metaxa, D.; Park, J. S.; Robertson, R. E.; Karahalios, K.; Wilson, C.; Hancock, J.; and Sandvig, C. 2021. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends® in Human–Computer Interaction*.

Mozilla. 2019. YouTube Regrets. https://foundation.mozilla.org/en/youtube/regrets/. Accessed: 2024-04-10.

Paolacci, G.; and Chandler, J. 2014. Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science*.

Papadamou, K.; Zannettou, S.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; and Sirivianos, M. 2022. "It is just a flu": assessing the effect of watch history on YouTube's pseudoscientific video recommendations. In *ICWSM*.

Ribeiro, M. H.; Ottoni, R.; West, R.; Almeida, V. A. F.; and Meira, W. 2020. Auditing radicalization pathways on YouTube. In *FAccT*.

Ricks, B.; and McCrosky, J. 2022. Does This Button Work? Investigating YouTube's ineffective user controls. https://foundation.mozilla.org/en/research/library/user-controls/report/. Accessed: 2024-04-10.

Salganik, M. J. 2019. *Bit by bit: Social research in the digital age*. Princeton University Press.

Shin, D. 2020. How do users interact with algorithm recommender systems? The interaction of users, algorithms, and performance. *Computers in Human Behavior*.

Smith, K.; Bullen, G.; and Huerta, M. 2021. Dark Patterns in User Controls: Exploring YouTube's Recommendation Settings – Simply Secure. https://simplysecure.org/blog/dark-patterns-in-user-controls-exploring-youtubes-recommendation-settings/. Accessed: 2024-04-10.

Tomlein, M.; Pecher, B.; Simko, J.; Srba, I.; Moro, R.; Stefancova, E.; Kompan, M.; Hrckova, A.; Podrouzek, J.; and Bielikova, M. 2021. An Audit of Misinformation Filter Bubbles on YouTube: Bubble Bursting and Recent Behavior Changes. In *ACM RecSys*.

Tufekci, Z. 2018. Opinion | YouTube, the Great Radicalizer. https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html. Accessed: 2024-04-10.

Wu, S.; and Resnick, P. 2021. Cross-Partisan Discussions on YouTube: Conservatives Talk to Liberals but Liberals Don't Talk to Conservatives. In *ICWSM*.

Wu, S.; Rizoiu, M.-A.; and Xie, L. 2018. Beyond Views: Measuring and Predicting Engagement in Online Videos. In *ICWSM*.

Wu, S.; Rizoiu, M.-A.; and Xie, L. 2019. Estimating Attention Flow in Online Video Networks. *ACM CSCW*.

YouTube. 2017. You know what's cool? A billion hours. https://blog.youtube/news-and-events/you-know-whats-cool-billion-hours/. Accessed: 2024-04-10.

YouTube. 2019. Continuing our work to improve recommendations on YouTube. https://blog.youtube/news-and-events/continuing-our-work-to-improve/. Accessed: 2024-04-10.

YouTube. 2020. Managing harmful conspiracy theories on YouTube. https://blog.youtube/news-and-events/harmful-conspiracy-theories-youtube/. Accessed: 2024-04-10.

Zhao, Z.; Hong, L.; Wei, L.; Chen, J.; Nath, A.; Andrews, S.; Kumthekar, A.; Sathiamoorthy, M.; Yi, X.; and Chi, E. 2019. Recommending what video to watch next: a multitask ranking system. In *ACM RecSys*.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? N/A
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes
   (g) Did you discuss any potential misuse of your work? Yes
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? All code used for this work is made public for finding reproduction.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? N/A
   (b) Have you provided justifications for all theoretical results? N/A
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A
   (e) Did you address potential biases or limitations in your theoretical framework? N/A
   (f) Have you related your theoretical results to the existing literature in social science? N/A
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? N/A
   (b) Did you include complete proofs of all theoretical results? N/A
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? N/A
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? N/A
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, we reported 95% CIs.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? N/A
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes

(b) Did you mention the license of the assets? We used several open datasets.

(c) Did you include any new assets in the supplemental material or as a URL? Yes

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Prolific participants must consent to participate in our survey study.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? N/A

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? N/A

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? N/A

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Yes, the annotation codebook is included in the appendix. Please find it at https://arxiv.org/abs/2307.14551

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, the data collection plan was submitted to the University of Michigan IRB and determined to be exempt (Study ID: HUM00224551).

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes. The first two authors annotated the data. Prolific participants were paid $15 an hour.

(d) Did you discuss how data is stored, shared, and deidentified? Yes