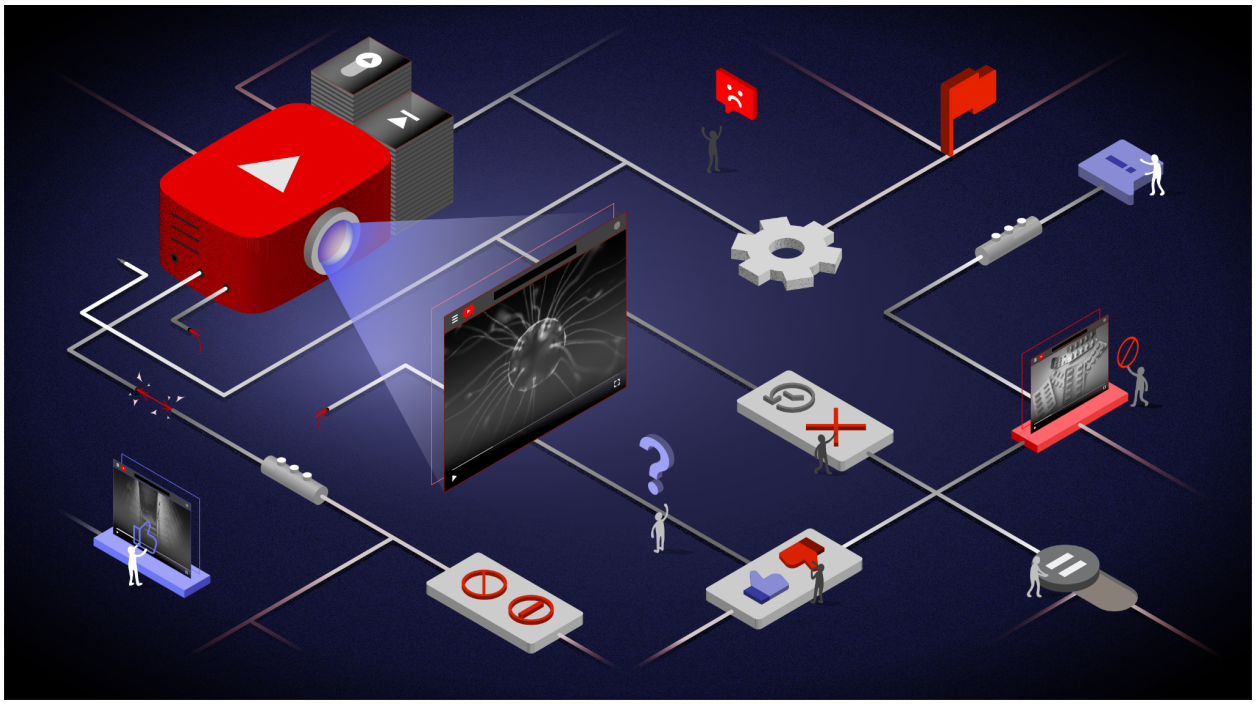
moz://a

# Does This Button Work? Investigating YouTube's ineffective user controls



Authors: Becca Ricks and Jesse McCrosky

September 2022

# Table of Contents

## About Mozilla

Mozilla's mission is to ensure the internet is a global public resource, open and accessible to all. An internet that truly puts people first, where individuals can shape their own experience and are empowered, safe and independent.

Founded as a community open source project in 1998, Mozilla currently consists of two organizations: the non-profit Mozilla Foundation, which leads our movement building work; and its wholly owned subsidiary, the Mozilla Corporation, which leads our market-based work, including the development of the Firefox web browser. The two organizations work in close concert with each other and a global community of tens of thousands of volunteers under the single banner: Mozilla.

foundation.mozilla.org

# Executive Summary

YouTube is the second most visited website in the world, and its algorithm drives most of the video views on YouTube. Previous Mozilla research determined that people are routinely recommended videos they don't want to see, including violent content, hate speech, and political misinformation.

YouTube says that people can manage their recommendations and search results through the feedback tools the platform offers, but we heard from people that they do not feel in control over their experience with the YouTube algorithm. We surveyed 2,757 participants about their feelings of control in relation to the platform and we learned that many people feel their actions don't have any effect on YouTube recommendations.

To test whether these experiences are backed by data, we evaluated the effectiveness of these controls for real users of the platform. Powered by Mozilla's research tool RegretsReporter, 22,722 people donated data about their interactions with YouTube. This study represents the largest experimental audit of YouTube by independent researchers, powered by crowdsourced data.

We looked at what happened over time to people's recommended videos after they had used one of YouTube's feedback tools – buttons like "Dislike" and "Don't Recommend Channel." From Dec 2021 to June 2022, RegretsReporter participants shared 567,880,195 video recommendations with us. In collaboration with researchers from the University of Exeter, we used a machine learning model we built to analyze video similarity. Through this approach, we were able to study what kind of effect YouTube's tools have on video recommendations for real users of the platform.

**In this report, we describe what we learned from our research using RegretsReporter data.** Through complementary qualitative and quantitative studies, we determined that:

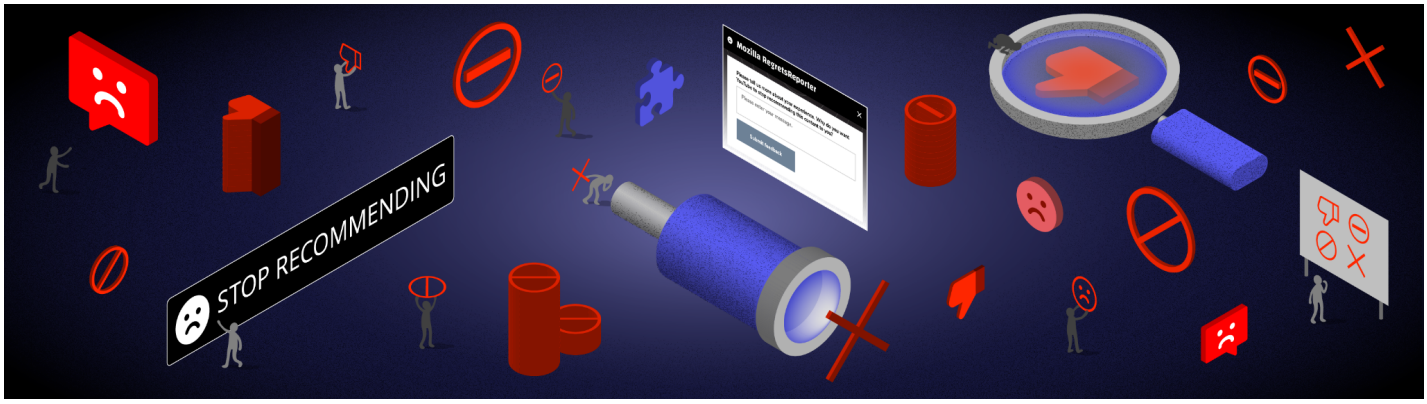
1. **People feel that using YouTube's user controls does not change their recommendations at all.** We learned that many people take a trial-and-error approach to controlling their recommendations, with limited success.
2. **YouTube's user control mechanisms are inadequate for preventing unwanted recommendations.** We determined that YouTube's user controls influence what is recommended, but this effect is negligible and most unwanted videos still slip through.

In the report, we provide some examples of videos that were recommended after RegretsReporter participants used YouTube's feedback tools. For example, one participant asked YouTube to stop recommending firearm videos — but was shortly after recommended more gun content. Another person asked YouTube to stop recommending cryptocurrency get-rich-quick videos, but then was recommended another crypto video.

**In this report, we also provide a set of recommendations to both YouTube and policymakers.** This guidance includes:

1. **YouTube's user controls should be easy to understand and access.** People should be provided with clear information about the steps they can take to influence their recommendations, and should be empowered to use those tools.
2. **YouTube should design its feedback tools in a way that puts people in the driver's seat.** Feedback tools should enable people to proactively shape their experience, with user feedback given more weight in determining what videos are recommended.
3. **YouTube should enhance its data access tools.** YouTube should provide researchers with access to better tools that allow them to assess the signals that impact YouTube's algorithm.
4. **Policymakers should protect public interest researchers.** Policymakers should pass and/or clarify laws that provide legal protections for public interest research.

This report also includes details about our research questions, methodology, and analysis for both our qualitative and quantitative studies.




## Introduction

YouTube is the second most visited website in the world, and its algorithm drives most of the video views on YouTube, more than channel subscriptions or search. The problems with YouTube's video recommendation algorithm have been well-documented, with claims that the platform promotes incremental radicalization ("rabbit holes")[1], political extremism[2], and ideological bias.[3]

In its communications, the platform claims that it optimizes its recommendation system for metrics like "user satisfaction" and "time well spent" – not just watchtime. YouTube CEO Susan Wojcicki has emphasized this point in public events. However, evidence suggests that the platform continues to prioritize engagement over people's well-being. A 2019 paper written by Google engineers[4] describes how the algorithm is designed to balance user engagement with user satisfaction: a tradeoff between promoting content that leads people to spend more time on the site and content the algorithm thinks people will like.

---

[1] Zeynep Tufekci, "YouTube's Recommendation Algorithm Has a Dark Side," *Scientific American* (April 1, 2019), https://www.scientificamerican.com/article/youtubes-recommendation-algorithm-has-a-dark-side/.
[2] Becca Lewis, "Alternative Influence," *Data & Society* (Data & Society Research Institute, September 18, 2018), https://datasociety.net/library/alternative-influence/.
[3] Muhammad Haroon et al., "YouTube, The Great Radicalizer? Auditing and Mitigating Ideological Biases in YouTube Recommendations" (*arXiv*, March 24, 2022), https://doi.org/10.48550/arXiv.2203.10666.
[4] Zhe Zhao et al., "Recommending What Video to Watch next: A Multitask Ranking System," in *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19 (New York, NY, USA: Association for Computing Machinery, 2019), 43–51, https://doi.org/10.1145/3298689.3346997.

YouTube says that people can control their recommendations and search results through the feedback tools the platform offers. However, many of the stories surfaced through Mozilla's 2019 YouTube Regrets campaign suggest that people continue to see unwanted videos despite having followed the steps prescribed by YouTube to avoid them. In our own 2021 investigation into YouTube's recommender system, we heard from people that they do not feel in control over their experience on YouTube, nor do they have clear information about how to change their recommendations.

Mozilla's vision for Trustworthy AI (such as recommendation algorithms) is that people have meaningful control over these technologies. When a platform has poorly designed controls — or controls that do not perform at all — people feel disempowered and helpless. In collaboration with Mozilla, the organization Simply Secure mapped and analyzed YouTube's controls in 2021 to understand if the platform's design supported user experience principles of control, freedom, and transparency. Their analysis of those controls, "Dark Patterns in User Controls: Exploring YouTube's Recommendation Settings," determined that YouTube's user controls do not appear to be designed with people's well-being in mind. They concluded that:

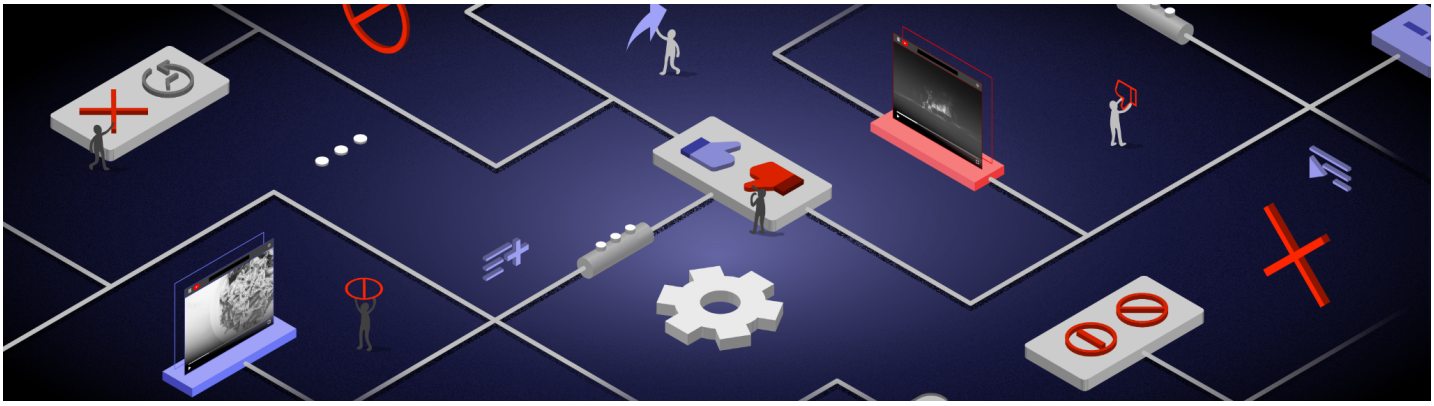
- YouTube's existing user controls are reactive and not proactive, leaving people to catch-up to the recommendation engine rather than designing what they want to see.
- Options to "teach" the YouTube algorithm are few and limited in scope.

Ultimately, the controls on YouTube are reactive tools: They don't empower people to actively shape their experience on the platform.

To evaluate the effectiveness of YouTube's controls for people who use the platform, we carried out a study that leverages Mozilla's large community of RegretsReporter volunteers. 22,722 people donated their data to Mozilla, generating a dataset of the 567,880,195 videos they were recommended. This study represents the largest experimental audit of YouTube by independent researchers, powered by crowdsourced data.

To understand whether people feel in control, Mozilla surveyed 2,757 RegretsReporter participants to better understand their experiences with YouTube's recommendation algorithm.

Ultimately we wanted to learn whether people feel in control on YouTube – and whether those experiences are actually validated by our RegretsReporter data. By combining quantitative and qualitative insights in this research project, we aim to paint a more complete picture of how YouTube's recommendation algorithm handles user feedback.



# ‘Nothing changed’: A qualitative analysis of YouTube’s user controls

## Overview

People test out many different strategies to control their experience with YouTube’s recommendation algorithm. Many people use the controls that YouTube explicitly suggests like the “Don’t Recommend Channel” button, but many also engage in different behaviors that YouTube may not anticipate, taking a trial-and-error approach to controlling their experience.

The beliefs that people hold about social media algorithms shape the way they behave online. People develop their own theories about why certain content does or doesn’t show up in their feed[5] and they may change how they present themselves on the platform in response.[6] They may use different strategies to “game” or control the algorithm, depending on their objectives. In the past, communities of content creators on YouTube have used various tactics to manage how much they show up on the platform,[7] including strategically

[5] Taina Bucher, “The Algorithmic Imaginary: Exploring the Ordinary Affects of Facebook Algorithms,” *Information, Communication & Society* 20, no. 1 (January 2, 2017): 30–44, https://doi.org/10.1080/1369118X.2016.1154086.
[6] Michael Ann DeVito, “Adaptive Folk Theorization as a Path to Algorithmic Literacy on Changing Platforms,” *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2 (October 18, 2021): 339:1-339:38, https://doi.org/10.1145/3476080.
[7] Sophie Bishop, “Anxiety, Panic and Self-Optimization: Inequalities and the YouTube Algorithm,” *Convergence* 24, no. 1 (February 1, 2018): 69–84, https://doi.org/10.1177/1354856517736978.

using thumbnail images to boost views.[8] Many researchers treat such actions as legitimate attempts by people to exercise control in the face of information or power asymmetry.[9] These tactics are ways to wrest back control from the algorithm, which can be characterized as a form of user resistance.[10]

In the qualitative research we conducted, we looked at how people talk about their experiences with YouTube's recommender system. We surveyed 2,757 YouTube users and conducted user interviews in order to better understand people's feelings of control and autonomy in relation to the platform.

The people who chose to participate in this study are not a representative sample of YouTube's user base. They are people who voluntarily downloaded the RegretsReporter browser extension and agreed to complete a survey and/or interview. We assume that many of them came to this experiment with a desire to express their feelings about the platform, which prevents us from drawing insights about all YouTube users.

Nevertheless, in this study we take seriously the complaints that people express about their experience with YouTube. Referencing Sara Ahmed's thinking on refusal, resignation and complaint, Burrell and others state that "the act of complaint itself can be a way for people to record their grievances and build solidarity in the face of limited recognition."[11] Building on their insight, we believe it's crucial to understand people's grievances with algorithmic systems in order to understand their expectations for how such systems should behave. In the case of YouTube, such complaints can help us better understand the frustrations that users face in relation to the platform's recommendation algorithm.

### Research questions

We carried out a qualitative study to better understand whether people felt they were in control of their YouTube experience. Going into this study, we had three primary research questions:

---

[8] Taina Bucher, "Cleavage-Control: Stories of Algorithmic Culture and Power in the Case of the YouTube 'Reply Girls,'" in *A Networked Self and Platforms, Stories, Connections* (Routledge, 2018).
[9] Jenna Burrell, Zoe Kahn, Anne Jonas, and Daniel Griffin, "When Users Control the Algorithms: Values Expressed in Practices on Twitter," *Proceedings of the ACM on Human-Computer Interaction* 3, no. CSCW (November 7, 2019): 138:1-138:20, https://doi.org/10.1145/3359240.
[10] Michael A. DeVito, et al., "'Algorithms Ruin Everything': #RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17 (New York, NY, USA: Association for Computing Machinery, 2017), 3163–74, https://doi.org/10.1145/3025453.3025659.
[11] Burrell et al.

1. **What steps do YouTube users take to control their video recommendations?**
2. **Do they feel that their recommendations change when they use YouTube's user controls?**
3. **Do they feel that they have control over their recommendations?**

We also had some secondary questions, including: How easy is it for YouTube users to find information about how the algorithm works? What information do they want to know? What kinds of alternative features or tools do they say they want? Note that we won't cover our analysis of answers to these questions in this report.

## Research setup

### Data collection

To answer our research questions, we ran a survey that invited people to reflect on a specific experience they'd had when they had tried to curate or control their video recommendations: What steps did they take? Why? What kind of effect did they think it had on their experience? Did they feel like they were in control?

Our RegretsReporter community was invited to fill out the optional survey upon download of the browser extension. The survey ran for four months from Dec 2021 – April 2022, with a total of 2,757 responses received. We did not collect demographic data so we do not have detailed information about the sample.

We also conducted semi-structured interviews in May 2022 with five people who had filled out our survey. Interview participants were selected at random from a pool of people who indicated they were willing to be interviewed by Mozilla researchers, and questions were tailored to their specific survey responses. The goal of the interviews was to contextualize and deepen our understanding of some of the themes that emerged in the survey responses.

### Analysis

Once we collected the survey data, we performed a content analysis[12] of the responses, categorizing and tagging them according to a coding schema that combined inductive and deductive approaches. We began with a set of codes that reflect the kinds of responses we anticipated receiving, but ultimately codes were developed based on the kinds of

[12] Johnny Saldaña, "An Introduction to Codes and Coding," in *The Coding Manual for Qualitative Researchers*, 3rd edition (Sage, 2016).

responses we received. We did not code the interviews, but we did rely on them as sources for rich contextual information that informed our coding of the survey responses.

Through the coding process, we identified the most prominent codes or themes. Our findings were structured around these themes, which looked at:

- The control strategies and behaviors people used;
- People's impressions of whether their control strategies were successful; and
- People's impressions about how the algorithm behaved in response to their control strategies.

## Findings

### 1. People use a broad range of tactics and behaviors in an attempt to control their YouTube recommendations.

Survey respondents were asked to walk us through a specific experience or set of experiences in which they had taken steps to control their YouTube recommendations. From those responses, we identified several tactics, behaviors, and actions that emerged:

| Control strategy | Description |
|---|---|
| Used YouTube's feedback tools | *Clicked "Not Interested" or "Do Not Recommend Channel", "like"/"dislike", blocked channels and videos, unsubscribed or subscribed to channels, made a playlist, etc.* |
| Changed YouTube settings | *Turned off recommendations, disabled personalization, erased preferences, removed video from watch history, switched from autoplay to random play, etc.* |
| Adjusted viewing behaviors | *Avoided watching unwanted videos, re-watched desired videos, avoided clicking recommendations, only used search, only viewed unwanted videos in private browsing mode, etc.* |

| | |
|---|---|
| Used a different account | *Logged in from another computer or YouTube account, created a YouTube account for a specific purpose or for their kids, etc.* |
| Used non-YouTube privacy tools | *Erased browser history, cleared cookies, used a privacy browser extension or other privacy tools, etc.* |

### Used YouTube's feedback tools

Survey participants mentioned using one or more of YouTube's feedback tools, often in combination with other behaviors. Of participants who said they took some steps to control their recommendations, 78.3% mentioned using YouTube's existing feedback tools and/or changing YouTube settings.

### Used privacy tools and behaviors

It's not surprising to us that RegretsReporter volunteers engage in privacy-conscious behaviors. Participants talked about using VPNs (15 mentions) and privacy browser extensions (17 mentions) to manage recommendations. They also mentioned routinely deleting cookies (31 mentions) and clearing their browser history. One participant put it this way:

> *"Stop YouTube from keeping track of my history. Turn off autoplay. Maxed out my YouTube privacy settings. Most successful and satisfactory, though, has been logging out of Google and staying logged out. YouTube still gives me recommendations (I guess it's tracking my usage based on my IP address) but it is a lot less bad."* (Survey ID2319)

Survey participants also changed their browser settings in an attempt to avoid unwanted recommendations and protect their privacy. People said that they watched certain videos in private browsing or incognito mode (43 mentions) or by creating a new account just to watch certain videos (73 mentions).

Participants emphasized that there are situations in which they may want to watch a video on YouTube but don't want it to affect their recommendations. In those cases, people said they viewed the video in private browsing mode, by logging out of their account, or by removing the video from their watch history. One interviewee describes their experience watching Superbowl commercials on YouTube:

> *"When the Superbowl came around… if someone recommended a particular commercial, I used to log out of YouTube, watch the commercial, and then log back in. I have sometimes even gone to a different device just because… I don't want that clouding my recommendations."* (Interview ID4)

### Adjusted viewing behaviors

Participants told us that they intentionally only clicked on and watched those videos they wanted to be recommended. Some said that they proactively rewatched videos they liked in an effort to "teach" the algorithm about their interests. Several participants said that they only watched videos from their subscriptions (65 mentions) or specific topics. For instance:

> *"I made a conscious effort to only click on videos that pertained to a 'target' category, such as political videos."* (Survey ID1615)

Other people actively used features like the search bar (184 mentions) as a tool to influence their recommendations.

> *"Searched for known credible sources to reset recommendation algorithm"* (Survey ID2301)

> *"Search for some good content or topics and pray the recommendations would adjust after watching them."* (Survey ID1558)

Similarly, a number of people said that they intentionally avoided or ignored certain videos (224 mentions) in order to skirt bad recommendations and to "teach" the algorithm about their interests. Some participants said they even avoided "hovering" over unwanted videos (6 mentions) as a way to prevent unwanted recommendations.

> *"I usually just avoid watching questionable materials so as not to 'feed' the algorithm"* (Survey ID1762)

> *"I avoided watching videos I knew were designed to be clickbait, to avoid filling the recommended stream with clickbait videos."* (Survey ID2347)

Interestingly, not all of the videos people avoided were completely "unwanted" — some people said they were interested in a video or a topic, but they still avoided it because they were worried that the YouTube algorithm would over-recommend similar content in the future. Some of our participants described their experiences this way:

*"I avoided videos on subjects that I found interesting, but which the algorithm was giving too much emphasis to."* (Survey ID2589)

*"I avoided clicking a video I would like to watch, only because I was worried that doing so would lead me to get politically extreme recommendations."* (Survey ID113)

*"I actively avoid content that I may want to watch as a guilty pleasure because I don't want recommendations for that kind of content in the main."* (Survey ID1064)

Others said they closed the browser window or tab (25 mentions) where they saw the video, with hopes that the platform would interpret it as a negative feedback signal. Others simply stopped using YouTube altogether (5 mentions).

### 2. More than a third of people said that using YouTube's controls did not change their recommendations at all.

We learned that people are generally not satisfied with YouTube's user controls. A significant minority (39.3%) of people who used YouTube's controls did not feel that doing so impacted their recommendations at all. Just over a quarter (27.6%) felt that their recommendations did change in response, and fewer (23.0%) felt the system had an ambivalent or mixed response. Responses that did not mention engaging with YouTube's user controls were exempt from this particular analysis (596 responses).

After you took steps to control your recommendations, did they change? If so, how did they change?

| *Category* | **No** | **Yes** | **Mixed** | **Don't know** |
|---|---|---|---|---|
| *Responses (n=2161)* | 850 | 597 | 498 | 216 |
| *Percent* | 39.3% | 27.6% | 23.0% | 10.0% |

One participant who felt the algorithm had no response explained it:

> *"Nothing changed. Sometimes I would report things as misleading and spam and the next day it was back in. It almost feels like the more negative feedback I provide to their suggestions the higher bullshit mountain gets. Even when you block certain sources they eventually return."* (Survey ID915)

A participant who had a more positive experience said:

> *"The channels I asked not to be shown were not shown anymore. Also, lots of times instead of getting relevant recommendations I get a section called 'watch again' with some videos from creators I am already subscribed or that I've seen already."* (Survey ID927)

Many participants had ambivalent or conflicted feelings about how the algorithm behaved in response to their actions. One participant put it this way:

> *"They did sort of change, although I feel that 'she likes music theory' info packet that the algorithm had didn't get removed entirely, just lowered in priority. I still get the recommendations, and I still shoot them down; the recommendations are no longer half of my home page, maybe like 1/6th of my page."* (Survey ID811)

Another participant expressed that the algorithm changed in response, but not for the better:

> *"Yes they did change, but in a bad way. In a way, I feel punished for proactively trying to change the algorithm's behavior. In some ways, less interaction provides less data on which to base the recommendations."* (Survey ID112)

### 3. Of those people who had mixed experiences, common themes included unwanted videos popping up, controls that don't work as anticipated, and significant effort.

Many of our participants were conflicted about whether or not using YouTube's tools impacted their recommendations and their overall user experience. Among people who had ambivalent or mixed responses to this question, several common experiences emerged. During the inductive coding process, we categorized responses with mixed experiences as follows:

| | Sub-category | Percent with theme (n=498) |
|---|---|---|
| a | At first recommendations change, but eventually unwanted videos slip back in. | **23.7%** |
| b | Clicking "Don't Recommend Channel" effectively blocks a specific channel, but continue to get recommended similar videos from different channels. | **12.3%** |
| c | Recommendations change, but it takes significant time and sustained effort. | **9.3%** |

**(a) People said that at first their recommendations changed, but eventually unwanted videos "crept" back into their recommendations over time.**

This theme was present in 23.7% of coded responses from our "mixed" group. This group of users noticed some positive changes after they used the controls, but said that over time unwanted recommendations would return. Some blamed themselves for accidentally clicking a clickbait video and ruining their recommendations, while others blamed the algorithm for giving too much weight to clickbait videos.

> *"They change for a time, but reappear later on again. Some recommendations seem to be driven by trend [sic] created by larger audiences…The algorithm favours these trends and overwrites individual selections. I do not think it creates a general profile of individual users, or if so ignores it after a while."* (Survey ID265)

> *"Eventually it always comes back. The algorithm seems incapable of remembering a lesson for very long."* (Survey ID187)

**(b) Many people said that clicking the "Don't Recommend Channel" button seemed to be most effective at blocking a specific creator or channel, but they said that they continued to get recommended similar videos from different channels.**

This theme was present in 12.3% of coded responses from our "mixed" group. This group of users observed that when they clicked "Don't Recommend Channel" it seemed to actually work most of the time — videos from a particular channel were mostly blocked. However,

they said that they would continue to get recommended videos from different channels on similar topics. This experience highlights an expectation gap we encountered throughout our research: People simply do not have clear information about what each of these controls are designed to do.

> *"They do not re-recommend the precise channels I say an outright 'no' to but still recommend alternative channels of almost identical content in their place."* (Survey ID112)

> *"That specific channel was hidden, but other related videos popped up. Takes several other 'hide channels' plus watching videos on a completely different topic to fade those out. Sometimes they still come back."* (Survey ID112)

**(c) Some people said that gradually their recommendations changed, but it took a significant amount of time and sustained effort on their part.**

This theme was present in 9.3% of coded responses from our "mixed" group. People emphasized that changing their recommendations required vigilance: Repeatedly sending feedback signals to YouTube, curating their subscriptions and watch history, and avoiding certain videos. Many expressed frustration about the effort required, or how slowly the algorithm changed.

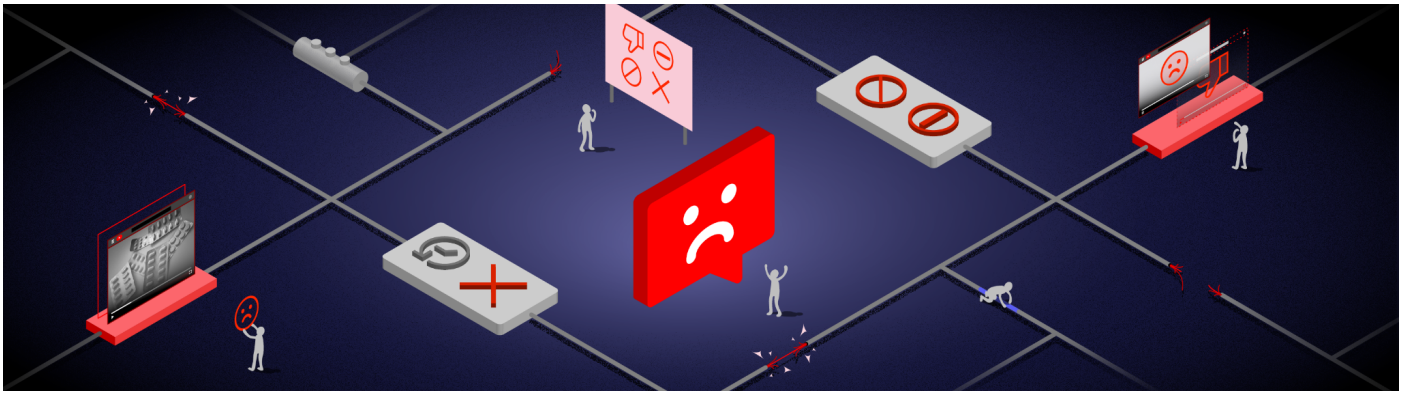
> *"I feel like I have to constantly curate the videos YT thinks I will like, or wants me to like. It's an ongoing battle."* (Survey ID112)

> *"It seems like you routinely have to prune things away or it will keep shoving them in your face until you tell it otherwise."* (Survey ID112)

## Takeaway

YouTube's user controls aren't designed in a way that allows people to actively design their experience on the platform. Our research demonstrates that YouTube's current controls leave people feeling confused and frustrated. Participants say they don't understand how YouTube's feedback tools impact what they are recommended, and often do not feel in control of their experience on the platform. Many resort to a trial-and-error approach that mixes tools, behaviors, and tactics, with limited success.

In the next section of the report, we examine whether people's feelings that they are not in control were validated when we analyzed the interaction data we collected from RegretsReporter users.



# Meager and inadequate: A quantitative analysis of YouTube's user controls

## Overview

Our qualitative research revealed that YouTube's current feedback tools leave people feeling frustrated, unable to control what they see. We wanted to learn whether the kinds of experiences our survey participants described were backed up with data: How does using these user controls impact the kinds of recommendations people get?

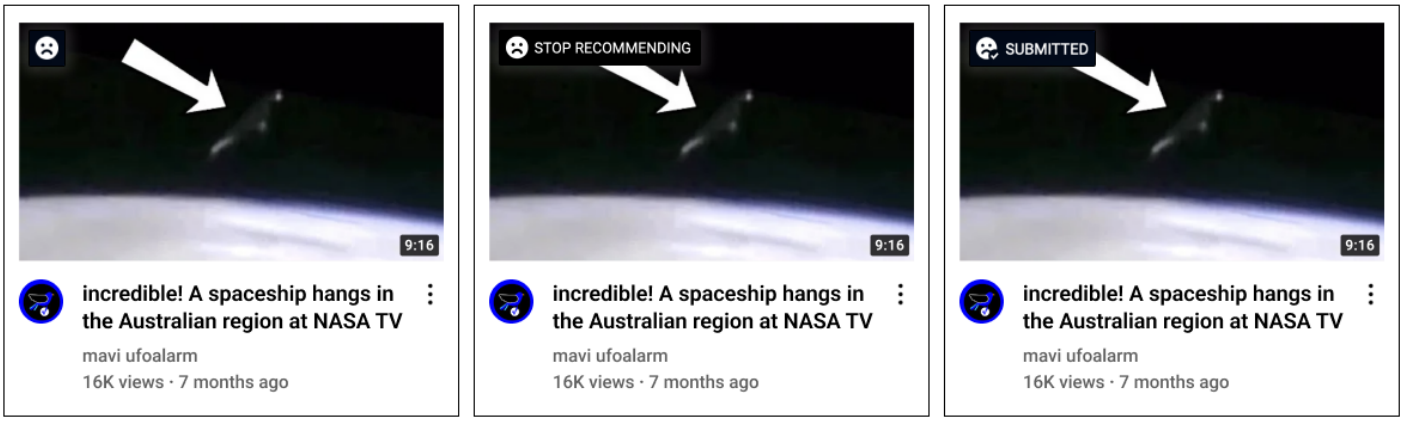
To answer this question, we ran a randomized controlled experiment across our community of RegretsReporter participants that could directly test the impact of the user control options that YouTube offers. This experiment allowed us to test how YouTube's user control features impact what videos people are recommended.

### The extension

The current version of our RegretsReporter extension is designed to answer the quantitative research questions posed in this study:

1. **How effective are YouTube's user controls at preventing unwanted recommendations?**
2. **Can adding a more convenient button for user feedback increase the rate at which feedback controls are used?**

After installing the extension, a “Stop Recommending” button is added to every video player or recommendation on participants’ YouTube.




Pressing that button sends a signal to YouTube that the participant doesn't want recommendations similar to that video. Depending on which experiment group the participant is part of, clicking the button will send one of many different types of feedback to YouTube (e.g. “Do Not Recommend Channel”, “dislike”, etc.), or it will send no feedback at all if the participant is in the control group.

For these participants who’ve opted into our research, the extension keeps track of which videos the “Stop Recommending” button is pressed for and what videos YouTube subsequently recommends.

Throughout this report, we use the following terms that describe aspects of the study:

| Terms used in the study | |
|---|---|
| **Rejected video** | In our study, to reject a video is to press the “Stop Recommending” button on it. This allows the participant to express that they do not want to see recommendations like this in the future and will (except for the control group) send a user control signal to YouTube to express this. |
| **Video pair** | A video pair is made up of a rejected video and a video that YouTube subsequently recommended. After a video is rejected, all following recommended videos will be paired with that rejected video for analysis. For example, if a participant rejects one vaccine skepticism video, and is later recommended a cat video, a music video, and another vaccine skepticism video, each of these recommendations will represent a pair with the rejected vaccine skepticism video. |

| | |
|---|---|
| **Bad recommendation** | A bad recommendation is a video pair for which the rejected and recommended videos are too similar according to our policy. Whether a video pair is a bad recommendation or not may be assessed by one of our research assistants, or by our machine learning video similarity model. |
| **Bad recommendation rates** | Our analysis is based on a metric that we refer to as a "bad recommendation rate". When we analyze the video pairs for each experiment group, the proportion of those pairs that are classified as bad recommendations is referred to as the bad recommendation rate for that group. |

## Findings

### 1. YouTube's user controls are inadequate tools for preventing unwanted recommendations.

Our study found that YouTube's user controls do have a measurable impact on subsequent recommendations. But contrary to what YouTube suggests, this effect is small and inadequate to prevent unwanted recommendations, leaving people at the mercy of YouTube's recommender system.

#### Research setup

In order to understand what kinds of recommendations people see after using YouTube's controls, we designed the experiment so that there were different experiment groups to compare: a control group (with users for whom no feedback was sent) and four treatment groups (groups of users for which the "Stop Recommending" button sent different types of feedback signals to YouTube). People who signed up for RegretsReporter were randomly assigned to one of these five groups.

Our control group helped us set the baseline. People who were part of our control group had the option to reject videos by clicking the "Stop Recommending" button, but no feedback would be sent to YouTube. Using the data collected from this group, we were able to calculate the baseline "bad recommendation rate" — YouTube's normal recommendation behavior without user feedback. By comparing the results of other experiment arms against this baseline rate, we were able to measure the effectiveness of YouTube's user controls.

Our treatment groups were then compared against that baseline. For them, clicking the button sent one of four different types of feedback to YouTube, either “Dislike”, “Don’t recommend channel”, “Not interested”, or “Remove from watch history.”


When we compared the bad recommendation rates for each of these four groups to the baseline rate from the control group, we found that each control does slightly reduce the bad recommendation rate relative to the baseline. However, participants are still served many bad recommendations. This demonstrates that YouTube’s user controls are not very effective at doing what people might expect them to do.

### Data analysis

In order to compare the experiment arms against one another, our research assistants reviewed about 40,000 pairs of videos and labeled them according to similarity. The goal was to determine whether the videos participants were being recommended were similar to videos they had rejected in the past, so that we could calculate the “bad recommendation” rate. We also used this data to train a machine learning model to analyze similarity for the rest of the video pairs.

What does a “bad recommendation” look like in practice? Below are some examples of videos that participants rejected (on the left), alongside videos that were subsequently recommended (on the right).[13] These examples demonstrate that YouTube continues to recommend videos that people have clearly signaled they do not want to see, including disturbing content like war footage and gruesome horror clips. *Trigger warning: Gruesome and disturbing images appear on the following page.*

[13] View a selection of video pairs here: https://drive.google.com/file/d/19bqoM6YIJttNt_4x14ZW-DixcAYtC1cl/view or download them from our JSON endpoint: https://public-data.telemetry.mozilla.org/api/v1/tables/telemetry_derived/regrets_reporter_study/v1/files/000000000000.json.

REGRETTED VIDEO


**Tucker: Justin Trudeau is attacking human rights**

Fox News

RECOMMENDED VIDEO


**Tucker: why is trans community running everything?**

Fox News

REGRETTED VIDEO

**Live Webcams From Around Ukraine | Conflict Zones ⚠️ | Kiev, Sumy, Slovyansk, Percomaisk**

Audionix

RECOMMENDED VIDEO


**How dead Russian soldiers are taken out of Gomel - Ukraine war**

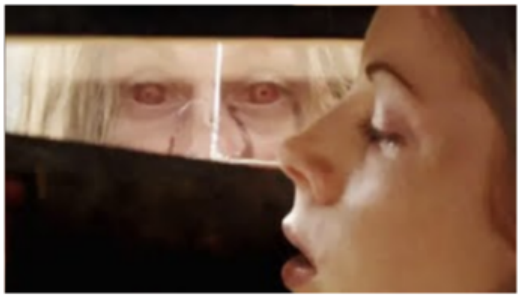
War Shock

REGRETTED VIDEO

**7 Scariest Horror Movie Twist Endings That Keep You Up At Night**

WhatCulture Horror

RECOMMENDED VIDEO



**10 Scariest Opening Horror Movie Scenes Ever**

WhatCulture Horror

This is especially concerning in light of the findings from our previous RegretsReporter study. We reported countless stories from people whose lives were impacted by YouTube's recommendation algorithm. Stories included one participant who, exploring YouTube while coming out as transgender, was exposed to countless videos describing their transition as mental illness. In another, children's content about a cartoon train led to autoplay of a graphic compilation of train wrecks.

There were over 500 million videos collected from RegretsReporter users to analyze. Since manually labeling all of those video pairs would not have been possible, we trained a machine learning model to effectively analyze the similarity between any two videos. Our model enabled us to calculate the actual bad recommendation rates by estimating similarity for all pairs and not just those that the research assistants reviewed.

The model is highly accurate for establishing our baseline and gives us a good sense of actual bad recommendation rates overall. For people in the control group, the bad recommendation rate was about 2.3%. We used an early version of the model to identify video pairs in our dataset that our RAs could manually evaluate in order to get a more accurate reading of the bad recommendation rates. Our analysis finds that even the most effective user controls prevent less than half of bad recommendations.

Percentage of unwanted recommendations prevented

| ~43% | ~11% | ~12% | ~29% |
|---|---|---|---|
| DON'T RECOMMEND CHANNEL | NOT INTERESTED | DISLIKE | REMOVE FROM WATCH HISTORY |

**Interpretation of findings**

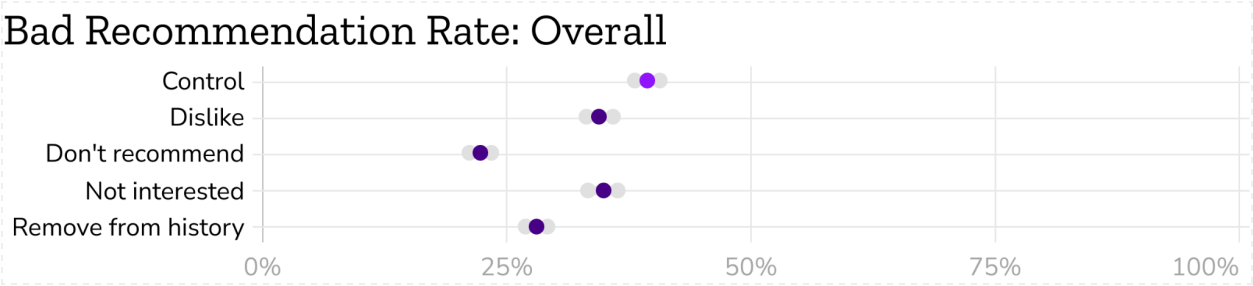
In our analysis of the data, we determined that YouTube's user control mechanisms are inadequate as tools to prevent unwanted recommendations.

To illustrate how we came to that conclusion, we'll walk through the various analyses we carried out and how we interpreted the data. Specifically, we'll look at how recommendations are impacted not just by the type of feedback signal (e.g. "dislike" versus "do not recommend channel"), but also things like channel, recommendation types (homepage versus sidebar), and time since feedback (1 week versus 4 weeks). Overall, a

consistent theme is that some of these tools have a small effect on improving recommendations but are inadequate as tools for exercising meaningful control.

### Analysis: Type of feedback signal

*Summary: YouTube's "don't recommend channel" and "remove from history" controls work better than others — but still don't work very well at all.*




*Bad recommendation rates among RA-labeled pairs by experiment arm, with 95% confidence interval.*

For this analysis, we looked at the different feedback signals people can send YouTube. In the graph above we see bad recommendation rates for the five different experiment groups: control, "dislike," "don't recommend," "not interested," and "remove from history." After calculating the bad recommendations rate[14] for each group, we compared them against one another and determined that:

- Using the "dislike" and "not interested" buttons seemed to slightly decrease the bad recommendation rate, so we can say they are marginally effective. But the impact on bad recommendations was very small.
- The "don't recommend channel" and "remove from history" buttons had slightly greater effectiveness, but our data still showed that the tools are inadequate and that people were still being served many bad recommendations after using these tools.

We don't know how YouTube handles this feedback internally, but it is interesting to note that the more "effective" methods might be interpreted by users as specific instructions, whereas the "less effective" controls might be interpreted as expressions of user preferences:

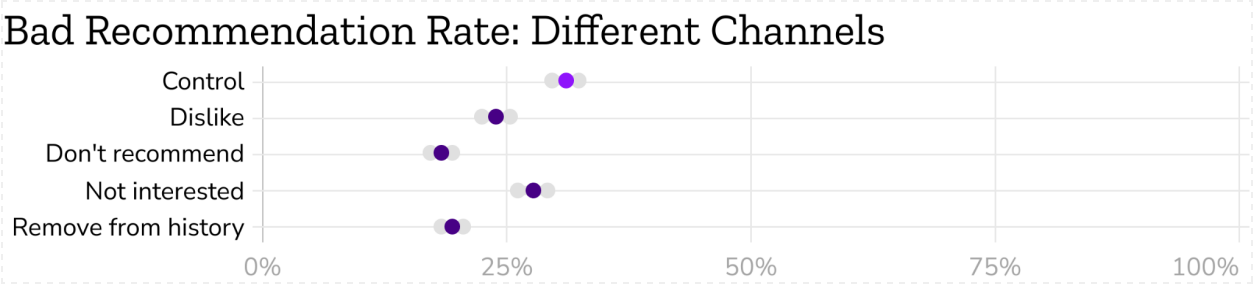
[14] Note that these rates were calculated over video pairs assessed by our RAs and thus the absolute rates are not representative of all recommendations. However, the comparisons that our analysis is based on are still valid.

- “Don’t recommend channel” might be interpreted by users as a relatively clear instruction: don’t show me this channel (although people might also consider it an expression of preference). We can attempt to confirm or disprove how effective the control is by comparing how it performs against those expectations.
- “Remove from history” might also be interpreted as a  fairly clear instruction, and can be confirmed by the user (using YouTube’s history browser). However, it’s not completely clear how this signal should influence future recommendations.
- “Dislike” and “not interested” might be interpreted as expressions of user preferences. These signals are less clear about how that preference will be accommodated.

As “don’t recommend channel” and “remove from history” are more effective, our assumption is that they send a stronger signal, whereas “not interested” and “dislike” send a weaker signal to YouTube. However, we do not know exactly how YouTube’s algorithm interprets various feedback signals because the platform does not make specific information available about the recommendation system’s parameters, inputs, and how people can adjust them.

### Analysis: Channel

*Summary: The “don’t recommend channel” control does have some impact even on similar videos from other channels, but does not consistently prevent recommendations from the unwanted channel.*




*Bad recommendation rates among RA-labeled pairs in which rejected and recommended videos come from different channels, by experiment arm, with 95% confidence interval.*

For this analysis, we looked at only those video pairs where the two videos came from different channels. For instance, someone clicked “don’t recommend” on a video from Jordan Peterson’s channel and then got recommended a video from the Fox News channel.

In the graph above, we visualized the differences in bad recommendation rates across those video pairs. As you can see, the “don’t recommend channel” button is still the most effective tool, even when it’s different channels that are being recommended. There could be many reasons for this: Perhaps people are seeing fewer videos from that channel so there are fewer clicks on those types of videos, changing YouTube’s understanding of the user’s interests over time. Or perhaps YouTube interprets “don’t recommend channel” as more generalized negative feedback.

### Handling of “Don’t recommend channel”

Compared to the other user controls we tested, people might have a clear idea about what the “don’t recommend channel” button is meant to do — block a channel from recommendations. We can analyze whether the same channel continues to pop up after a participant clicks this button.

For people in our control group, about 0.4% of subsequent recommendations after a rejected video were from the same channel as the rejected video. Meanwhile in the “don’t recommend channel” group, we see this rate drop to about 0.1%. In other words, telling YouTube to stop recommending a channel seems to have the impact you might expect in most cases: fewer videos from that channel.

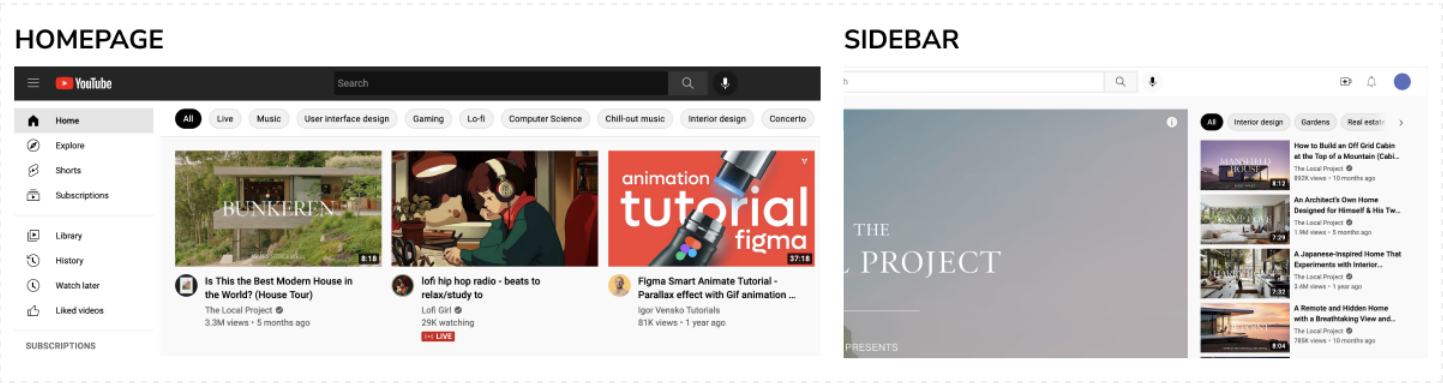
However, this feedback doesn’t appear to be consistently respected. In about 0.1% of the video pairs we analyzed, a recommendation was made from the same channel. Even when we limited our analysis to just a one-week time period between rejected video and recommendation, we saw the same pattern persist.

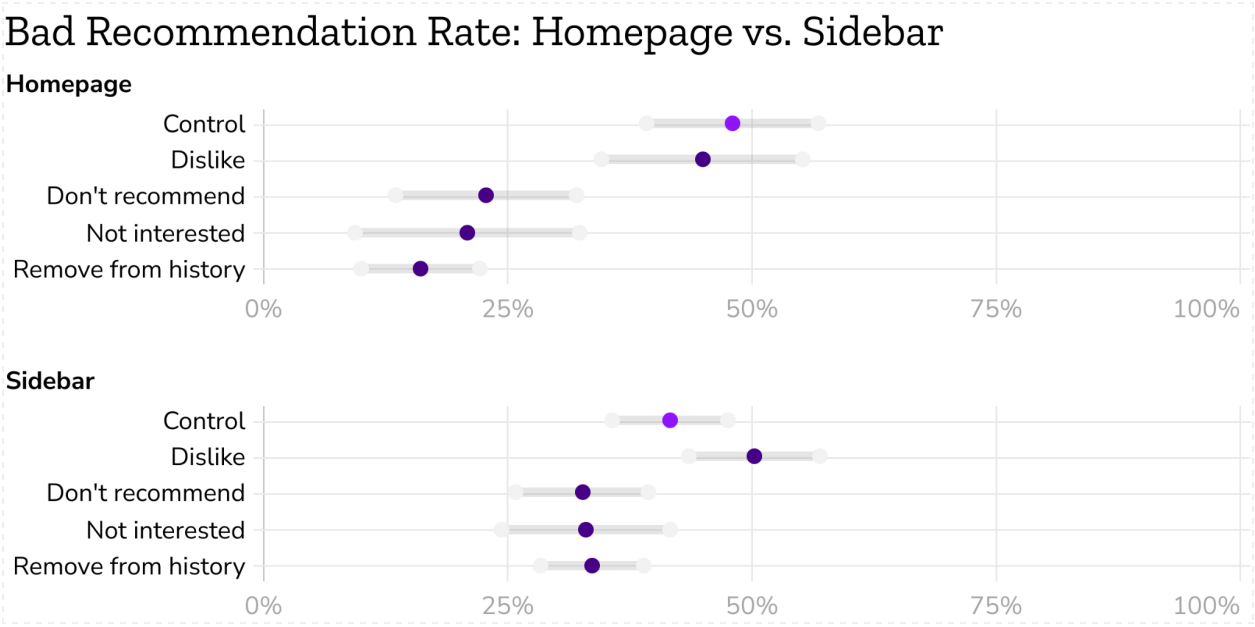
Our data does not allow us to confirm that the participant has not given YouTube a reason to ignore the “don’t recommend channel” feedback, perhaps watching a video from that channel that shows up in a search result, but it seems unlikely that this explains all the cases that we see. It appears very likely that “don’t recommend channel” doesn’t always work. This echoes one of the themes that emerged in our qualitative research: Many people said that they continued to get recommended similar videos from different channels even after clicking the button. Our participants feel that they don’t have much control over their recommendations, and our data backs up these experiences.

## Analysis: Recommendation types

*Summary: YouTube’s controls are slightly more effective for homepage recommendations and slightly less effective for sidebar recommendations.*

For this analysis, we looked at the impact on recommendations in different locations: the sidebar and the homepage.



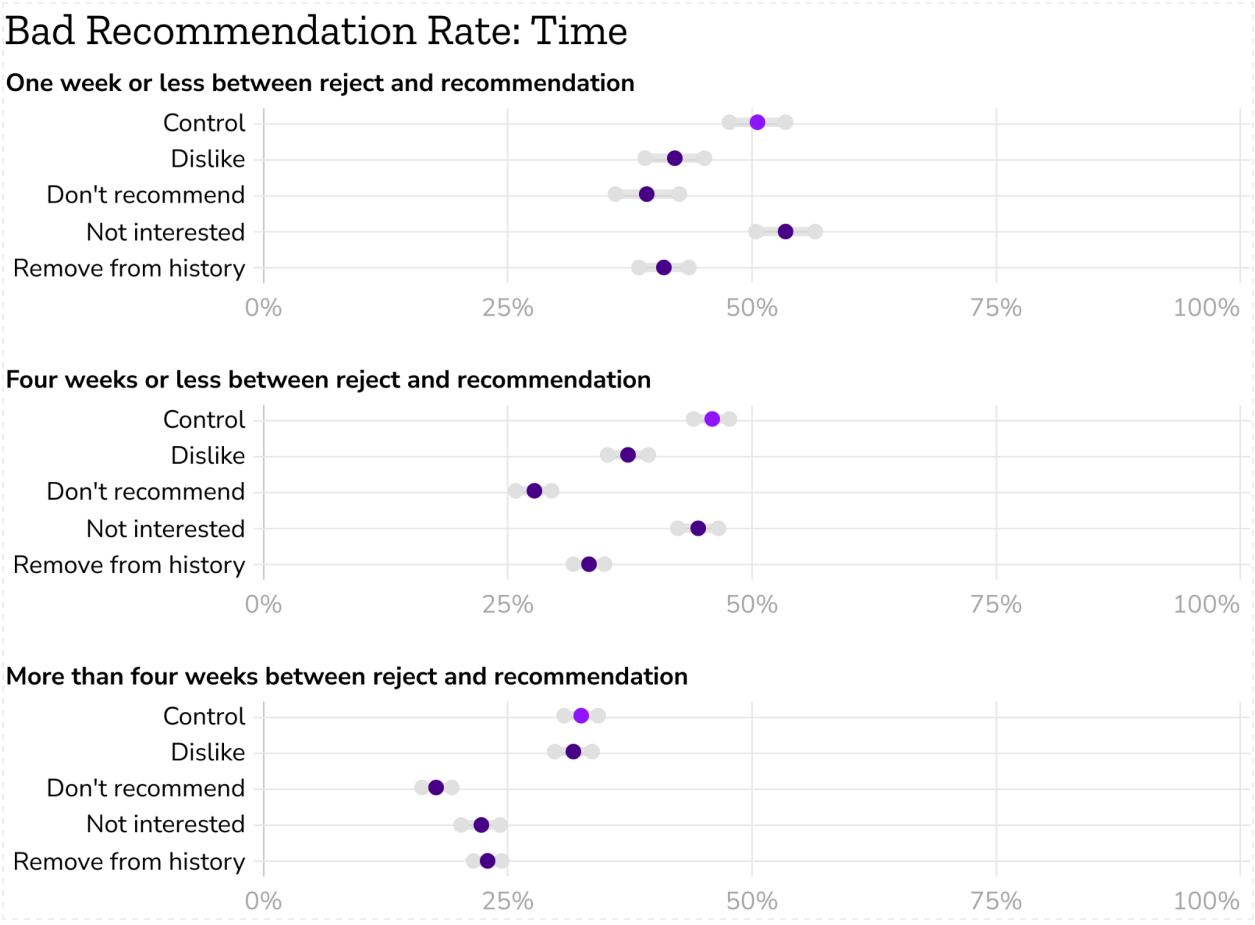

*Bad recommendation rates among RA-labeled pairs by experiment arm, broken down by source of recommendation, with 95% confidence interval.*

As illustrated in the graph above, user feedback seems more effective for homepage recommendations than sidebar recommendations. One possible interpretation could be that sidebar recommendations are made in a higher-information context (YouTube knows what video you are currently watching) and so it is easier to optimize for engagement and there is less weight given to user feedback. For homepage recommendations, YouTube has

less ability to know what might be engaging at the moment and so considers user feedback slightly more.

### Analysis: Time between reject and recommendation

*Summary: It does not appear that the effectiveness of YouTube's controls changes over time. However, we do see a drop in bad recommendation rates as time passes after a video is rejected.*




*Bad recommendation rates among RA-labeled pairs by experiment arm, broken down by time between rejection and recommendation, with 95% confidence interval.*

For this analysis, we looked at the amount of time that had passed between when a video was rejected and when another video was recommended. There were three time periods we looked at: within one week, within four weeks, and anything more than four weeks.

Although the impact of user feedback doesn't significantly change based on the time between rejection and recommendation, we do see that the overall bad recommendation

rates do drop as time between rejected video and recommendation increases. This might be explained by a general shift over time in the kind of content YouTube recommends, as well as YouTube using an interest model that decays over time: If YouTube's algorithm determines that a user is interested in a topic, but that user doesn't watch many videos on that topic, YouTube may slowly decrease its estimate of the user's interest in that topic as time passes.

## 2. An alternative UX can double the rate of user feedback on YouTube.

One of the problems identified in Simply Secure's usability audit of YouTube was that the platform's tools are not easy to use. There are very few options available for people to "teach" the algorithm beyond the handful of tools we've discussed, which are all reactive. Ideally, there should be tools made available to YouTube users that would allow them to define their interests, express their preferences for recommendations, and more actively shape their overall experience.

Even when people try to use YouTube's user controls, there are very basic obstacles to giving feedback. In this next section, we'll talk about how the UI of RegretsReporter was designed as an experiment to test how a different interface might encourage more feedback.

YouTube offers its feedback tools through a couple different user interfaces:

- "Don't recommend channel" and "Not interested" are available through the three dots menu on a recommendation.
- "Dislike" is available only in the video player screen.
- Removing a video from watch history requires navigating to the history tab on YouTube, finding the appropriate video, and then clicking on the "X" icon next to it.

Our extension makes submitting feedback easier. Feedback can be provided with a single click on a recommendation — YouTube's normal methods all require at least two clicks from a recommendation. In our study we investigated the degree to which this improved design increased the rate of user feedback submission.

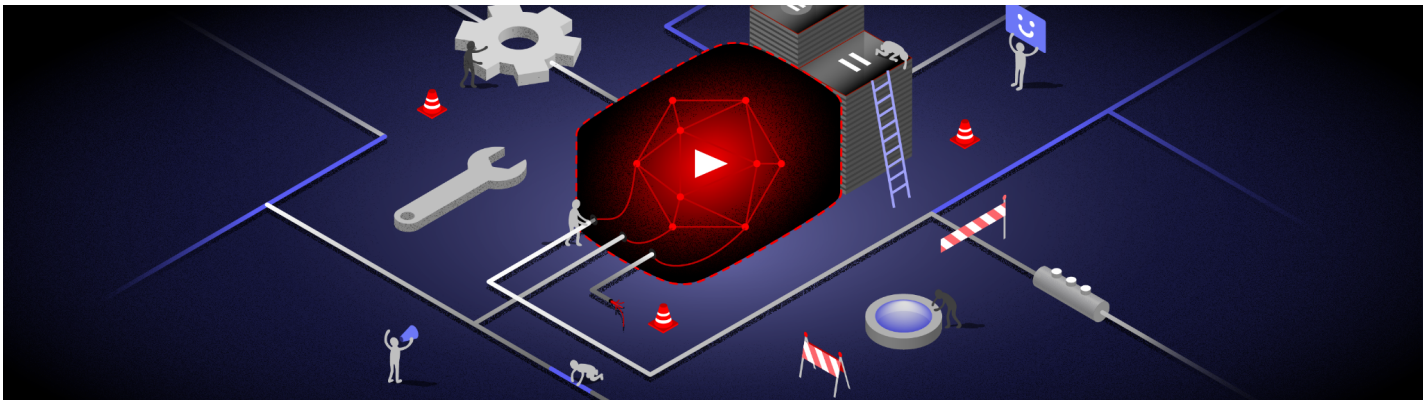
To analyze this, we included a special UX-control group. For participants in the group we do not show our "Stop Recommending" button and they are thus unable to reject videos – but they can still use YouTube's native user controls. For these participants, the extension has no apparent effect, but we still collect standard data. This allows comparisons of the frequency of user feedback submitted between those that do and do not see the "Stop Recommending" button.

We find that participants that see our button submit about 80 pieces of feedback per 1,000 videos watched, while those that do not see the button submit only about 37. This difference is not statistically significant due to enormous variation between participants, but if we restrict the analysis to participants with 50 or fewer feedback submissions to reduce variance, we still see a similar relationship (50 and 23 feedback submissions per 1,000 videos watched respectively) and a strong statistical significance. It is clear that adding our button more than doubles the rate of feedback submission.

In order to anticipate arguments that our button might be obtrusive and reduce video watch rate on YouTube, we also analyzed the video watch rates per participant in these two groups and found no statistically significant difference. Actually, the watch rate was higher among participants that saw the button.

## Takeaway

Through our controlled experiment, we were able to measure the effectiveness of YouTube's user control tools for preventing unwanted recommendations. While some effectiveness was observed for each tool, even the most effective tools were inadequate for preventing unwanted recommendations. Our research suggests that YouTube is not really that interested in hearing what its users really want, preferring to rely on opaque methods that drive engagement regardless of the best interests of its users.



## Recommendations

Many of our survey participants said they took a trial-and-error approach to curating their recommendations, with limited success. When people feel that they need to log out, create new accounts, or use privacy tools just to manage their YouTube recommendations, it not only ruins the overall user experience, it also erodes people's confidence in their ability to make meaningful choices on the platform.

YouTube should respect the feedback users share about their experience, treating them as meaningful signals about how people want to spend their time on the platform. **Here are our recommendations for YouTube and policymakers:**

| Recommendations | |
|---|---|
| **1** | **YouTube's user controls should be easy to understand and access.** People should be provided with clear information about the steps they can take to influence their recommendations, and should be empowered to use those tools. |
| **2** | **YouTube should design its feedback tools in a way that puts people in the driver's seat.** Feedback tools should enable people to proactively shape their experience, with user feedback given more weight in determining what videos are recommended. |

| | |
|---|---|
| 3 | **YouTube should enhance its data access tools.** YouTube should provide researchers with access to better tools that allow them to assess the signals that impact YouTube's algorithm. |
| 4 | **Policymakers should protect public interest researchers.** Policymakers should pass and/or clarify laws that provide legal protections for public interest research. |

## 1. YouTube's user controls should be easy to understand and access.

**People should be provided with detailed, accurate information about the steps they can take to influence their recommendations.**

In our study, we learned that YouTube's user controls have varying levels of effectiveness, but people don't always understand how those controls influence the recommendations they see. YouTube already provides a general overview of information, but it should provide more detailed information about what data points or signals influence recommendations, including what specific third-party data YouTube uses to inform a person's recommendations.

Importantly, YouTube should also help people understand what effect each control will have on their recommendations within the product, not just in an overview page. The tools themselves should use plain language that tells you what will happen when you use the control (e.g. "Block future recommendations on this topic") rather than the signal the control will send ("I don't like this recommendation").

YouTube should help people understand specifically what each of these controls do by providing clear, in-product explanations. For instance, through our study we learned that clicking "Don't Recommend Channel" sends a stronger signal than "dislike," but neither was completely effective at preventing unwanted recommendations. In order to help people understand what each of these controls do, for instance, YouTube could explain:

> Pressing 'Don't Recommend Channel' will reliably reduce recommendations from this channel, but content from similar channels might still appear in your recommendations.

Under the terms of the EU Digital Services Act (DSA),[15] enhanced user transparency will soon become a legal obligation for YouTube,[16] requiring the platform to explain how the recommendation algorithm prioritizes and displays content, including information about the algorithm's parameters.

But this level of transparency should be the bare minimum: General policy pages like "How YouTube Works" are useful to the general public but need to be accompanied by in-product tools and explanations that help people understand what kind of control they can exercise. These kinds of product changes would require rigorous user research and testing to ensure explanations are meaningful and useful to people, but would ultimately empower people to make more informed choices about their recommendations.

**People should be directed to documentation and tools through more intuitive user pathways.**

Simply Secure's mapping of user controls determined that managing recommendations on YouTube is a mess: false settings that don't do much, convoluted and confusing pathways, with multiple pages and pop-ups to consult for guidance. In some cases, the user pathways are circular, leading people back to pages they'd previously visited without providing additional clarity. YouTube has since made some improvements: a centralized hub where users can take action and manage their account settings and privacy.

YouTube should continue to assess the pathways users take to access controls and settings and redesign the user experience accordingly. These paths should be designed in a way that centers people's lived experiences, aimed at enhancing their autonomy and control. People should also be able to provide more detailed feedback about why they don't want to see certain videos that goes beyond just clicking a button.

## 2. YouTube should design its feedback tools in a way that puts people in the driver's seat.

As YouTube works towards giving users better information about each control, it should also ensure that those controls actually shape recommendations. We suggest YouTube

[15] European Parliament legislative resolution of 5 July 2022 on the proposal for a regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, https://www.europarl.europa.eu/doceo/document/TA-9-2022-0269_EN.pdf

[16] See Recital (70) DSA in the latest seen version of the text, previously Recital 52c.) DSA.

overhaul its existing controls in favor of better feedback tools that enable people to make informed, active choices about their experience on the platform.

### User controls should give people more control over what they see.

Many people we surveyed said they simply wanted to block certain kinds of channels and videos from future recommendations, but our study determined the current controls do not effectively do this. Concerningly, our study demonstrated that even after using YouTube's user controls, people continued to see videos that may violate the platform's community guidelines, or videos potentially considered "borderline" – videos that don't violate YouTube's policies, but that a broad audience might not want to see. In our dataset of recommended videos,[17] we observed that RegretsReporter users were recommended a number of videos that could fall under YouTube's policies on firearms, graphic content, or hate speech. Not only does this suggest that YouTube is struggling to enforce its own content policies, but that the platform is actively recommending this content to users even after they've sent negative feedback.

At a minimum, user controls should be reasonably effective at preventing recommendations of videos on a particular topic or by a particular channel. People should be able to exclude specific keywords, types of content, specific accounts, or other criteria from their recommendations. YouTube should explain how each of these categories of content is defined so that people can make informed choices.

### User feedback should be given more weight in determining how videos are recommended.

Over the course of doing this research, we learned that people's intentions for how they want to spend time on the platform might not line up with their behavior. There are many reasons people watch videos they do not want to be recommended in the future: Participants described watching videos that were "guilty pleasures," being drawn into "clickbait" videos, or accidentally clicking on a video with millions of views. People should have the ability to actively shape their recommendations in line with their interests and/or wellbeing.

YouTube should value and respect the feedback users send about their experience on the platform, treating them as meaningful signals about how people *want* to spend their time on YouTube. For instance, YouTube could design a more collaborative recommendation

[17] View a selection of video pairs here: https://drive.google.com/file/d/19bqoM6YIJttNt_4x14ZW-DixcAYtC1cl/view.

system in which user feedback is treated as the most important signal, above watch time or engagement. YouTube should consider overhauling its existing user controls in favor of feedback tools that actually put people in the driver's seat.

**YouTube's feedback tools should enable people to more proactively design their YouTube experience.**

Research into recommender systems has shown that people feel more satisfied when controls are paired with the ability to make meaningful choices about the recommendation algorithm.[18] The platform should overhaul its current user controls, and test alternative feedback tools that would allow people to more proactively design their YouTube experience. For instance, YouTube could design a feedback model that enables people to specify their interests and disinterests, including their preferences for subject matter, format, or diversity of recommendations.

Designing better feedback mechanisms could be a win-win for both users and YouTube: Rather than interacting with a paternalistic platform that makes choices on their behalf, people would have a greater say over what they see, and YouTube wouldn't have to rely as heavily on passive data collection in order to infer what people want to watch (for instance, by using engagement data to approximate a user's preferences). Simply put, YouTube could be "asking people what they want instead of just watching what they do."

YouTube may soon be legally required to make some of these changes. According to a DSA citation,[19] online platforms will need to take steps to ensure its users can influence how information is presented to them. But our recommendation goes further: YouTube should reimagine how user feedback is given and interpreted on the platform. Feedback tools should empower people to take an active role in designing and curating their YouTube experience – rather than relying on reactive tools like the "dislike" button and hoping it sends the right signal.

[18] Jaron Harambam et al., "Designing for the Better by Taking Users into Account: A Qualitative Evaluation of User Control Mechanisms in (News) Recommender Systems," in *Proceedings of the 13th ACM Conference on Recommender Systems* (RecSys '19: Thirteenth ACM Conference on Recommender Systems, Copenhagen Denmark: ACM, 2019), 69–77, https://doi.org/10.1145/3298689.3347014.
[19] See Recital 70 (previously 52c) DSA.

## 3. YouTube should enhance its data access tools.

Mozilla built RegretsReporter as a crowdsourced research platform because it is very difficult to access the kind of data needed to study YouTube's algorithm. To run this study, we relied primarily on crowdsourced data collected from RegretsReporter users as well as automated web requests to acquire video metadata from YouTube. The reality is that YouTube's API has its limits: Critical classes of data are not made available, and rate limits prevent large scale data collection. What's more, YouTube does not provide other kinds of audit tools for researchers to assess harm on the platform.

**YouTube should give researchers access to platform data so that they can adequately scrutinize the platform.**

YouTube should ensure that the tools and systems it builds enable large-scale analysis of platform content. At a minimum, YouTube should improve its public-facing API so that independent researchers, journalists, and the general public can have greater access to platform data, in a way that's privacy-protecting.

Researchers need to be able to access the content hosted on YouTube in order to adequately diagnose problems on the platform. Vetted researchers should be given access to platform data that goes beyond what is offered in the public API, such as public platform content, users, or pages on the platform, in a way that's in line with data protection rules and principles. The European Digital Media Observatory (EDMO) has proposed a draft code of conduct on how platforms can give researchers access to data while protecting user privacy, in compliance with the General Data Protection Regulation (GDPR).

Mandating researcher access to platform data was one key outcome of the EU Digital Services Act (DSA), legislation which was adopted by the European Parliament in July 2022. Under the DSA's article on "Data Access and Scrutiny",[20] vetted independent researchers must be given the tools they need to investigate and run experiments on recommender systems as they relate to platform harms or "systemic risks". While the details of who will get access and what data will be made available are still being worked out,[21] platforms will need to expand and improve their data access tools.

On the heels of the DSA's adoption, YouTube announced its release of a new researcher-facing API. The researcher API seems to provide the exact same data that was already available through its public API, but with increased rate limits on a case-by-case

[20] See Article 40 of the latest seen text of the DSA (previously Article 31).
[21] European Commission, "2022 Strengthened Code of Practice on Disinformation", 2022, https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation.

basis. In the context of our study, access to this API would have saved a lot of time and trouble, since we had to rely on alternative tools for acquiring video metadata due to the API's current rate limit.

However, the researcher API would not have helped with much else in this study. The data we analyzed — what videos are being watched, what is being recommended, and how people interact with the platform — is not data that would be made available with this new researcher API. In addition, under the terms of the researcher API program, only researchers at accredited and vetted academic institutions will be eligible for accessing the API. This means that journalists and independent researchers like Mozilla's would not be able to use this researcher API.

As YouTube works towards enhancing its data access tools, it's important that it ensures both the researcher-facing and public API are designed to help civil society actors assess the risks and harms on the platform.

**YouTube should build tools that enable researchers to carry out audits and experiments on the platform.**

In addition to ensuring researcher access to data, YouTube should build tools that enable researchers to audit and run tests on the platform. Independent audits and experiments are essential to holding platforms accountable. Such efforts may be voluntary on YouTube's part, but policymakers should consider mandating the release of such tools.

YouTube should provide tools that simulate user pathways or recommendation pathways. One of the challenges researchers face when studying YouTube is that personalization is very difficult to study without access to real user data. Simulation tools would allow researchers to tweak inputs and parameters in order to run experiments and tests on YouTube's recommendation algorithm. In the context of this study, simulation tools could have helped us test how different user controls impact what videos the recommendation algorithm serves, looking at a range of different variables.

## 4. Policymakers should protect public interest researchers.

Mozilla researchers relied on a crowdsourced approach to studying YouTube for this study, working with tens of thousands of volunteers who offered up their data to support the research. Without access to data from real people about what they are experiencing, it's impossible to analyze the spread and impact of algorithmic harms on the platform. As previously discussed, we recommend YouTube provide greater access to platform data for

researchers – whether such moves are voluntary or mandated by regulators, under legislation like the DSA.

However, there are not clear legal protections in place for independent researchers who are conducting good-faith research in the public interest about tech platforms. Institutes like New York University's Ad Observatory that are carrying out critical research into ad targeting on Facebook have had their access to Facebook data revoked, in spite of the fact that their crowdsourced approach does not violate user privacy.

Policymakers should act to pass laws that provide legal protections for public interest research. Such protections should be written into platforms' terms of service agreements and encoded in law. They should protect independent researchers, civil society organizations, and journalists conducting research, as long as the research complies with data privacy laws and research ethics standards.

These protections should complement data access tools by protecting research that doesn't rely only on platform access tools. As we've demonstrated with this study, there are research questions that the data from YouTube's API doesn't currently answer. If we had waited for YouTube to release the data we need, we would never have been able to carry out this audit. As problems rapidly emerge and scale on platforms, it's critical that researchers have legal protections to carry out research, as long as it complies with privacy principles.

Such amendments and/or research exceptions could stipulate that platforms cannot actively block research tools or impose rate limits that are designed to prevent researchers from carrying out their work. Such protections should apply to researchers who are collecting crowdsourced data from platform users, scraping data from platforms, or conducting sock puppet audits – methods that in many cases violate platforms' terms of service.

In any case, clarity around the legality of public interest research is urgently needed. The current lack of legal protections stifles legitimate public interest research: Researchers worry that platforms may take legal action against them, or they may avoid research projects altogether that violate platforms' terms of service. Worryingly, researchers are deterred from conducting research when they don't have institutional support, or can otherwise assume legal risk. The current uncertainty around platform research further entrenches platform power and stifles critical public interest research.

# Conclusion

In our research, we determined that YouTube's user controls do not work for many people. Our analysis of RegretsReporter data revealed that these user controls don't seem to have a major effect on how videos are recommended. We heard from several people who expressed frustration with the user controls, and who said they wanted better tools that simply work the way they'd expect them to. Importantly, the tools themselves are largely reactive and they don't empower people to actively design how videos are recommended.

YouTube should make major changes to how people can shape and control their recommendations on the platform. YouTube should respect the feedback users share about their experience, treating them as meaningful signals about how people want to spend their time on the platform. YouTube should overhaul its ineffective user controls and replace them with a system in which people's satisfaction and well-being are treated as the most important signals.

# About the Methodology

## Survey

### Survey questions

Our survey ran for four months, from Dec 2021 to April 2022. Survey participants responded to the following questions:

1. Think about a time you took steps to curate or control your YouTube recommendations. What did you do?
2. What kinds of videos were you seeing that prompted you to take these steps? Why do you think YouTube was recommending those videos to you?
3. After you took steps to control your recommendations, did they change? If so, how did they change?
4. In this scenario, did you feel like you had meaningful control over your video recommendations? Why or why not?
5. What do you wish you had been able to do in this scenario? Are there other platforms you know of that allow you to do this? Provide examples if so.
6. Overall, what information do you wish YouTube provided users about how its recommendation algorithm works? What would you do with that information?
7. Are you interested in being interviewed by Mozilla researchers about your responses to these questions? If so, please provide your email address.

## Quantitative study

### The extension

The study was carried out by participants that installed our web extension and then opted into our experiment and data collection. For those that did not opt in, the reject functionality sent a “dislike” signal and no data was collected.

For those that did opt in, a random assignment was made, with equal probability, to each of our experiment arms:

- Control (or “placebo”): Pressing the "Stop Recommending" button will have no effect.
- Dislike: Pressing the button will send YouTube a "dislike" message.
- History: Pressing the button will send a message to remove the video from your watch history.
- Not interested: Pressing the button will send YouTube a "Not interested" message.

- Don’t recommend: Pressing the button will send YouTube a "Don't recommend channel" message.
- No button: The "Stop Recommending" button will not be shown, but standard data will still be collected.

The extension also collected data (for those that opted in) which was sent to Mozilla servers using Firefox telemetry. Data collected included:

- A unique installation ID
- Experiment arm assigned
- A record of all uses of the “Stop recommending” button with timestamp and video ID on which the button was pressed.
- A record of all recommendations made by YouTube including timestamp, video ID, and type of recommendation, i.e. sidebar or homepage.
- A record of all interactions with native YouTube user control features.

## Data

### Participant data

We analyzed data collected from Dec 2, 2021 to June 26, 2022. In this time period there were 22,838 participants that opted into our experiment and data collection. After removing exceptional behavior (either bots, scripts, or very unusual people), we have 22,722 participants for analysis. Of these, 6,147 of them rejected at least one video. There were a total of 30,314 rejected videos. Our participants were recommended 567,880,195 videos. In total there are 162,983,496 video pairs that we were able to analyze. This was limited by the quantity of YouTube metadata collected.

### Research Assistant Data

We contracted a set of 24 research assistants from Exeter university supervised by Dr. Chico Camargo to classify video pairs according to our policy. Between April 22 and June 26, 2022, they classified 44,434 pairs (after removal of data by a handful of RAs that were found to have produced high rates of incorrect classifications).

The classifications were:

- Acceptable Recommendation 75%
- Bad recommendation 22%
- Unsure 3%

Video language was automatically classified using the gcld3 model applied to the video description. Based on this, the classified videos included 102 different languages. We did seek out research assistants with varied language skills, although we also allowed them to classify pairs in languages they didn't understand by using translation tools or in cases where language understanding was not necessary to make a classification.

For the initial days of classification, we asked the research assistants to classify random pairs from our data. For the majority of the classification period, however, we asked them to classify pairs selected to have a wide range of different predicted (by our model) probabilities of being bad. This was partly to employ a method known as active learning to improve our classification model and partly to ensure that we could effectively calibrate that model.

### YouTube data

The extension reports video IDs and, when possible, includes metadata such as video title, description, and channel. However, we needed this data consistently and also needed video transcripts. This data was obtained by automatic web requests to YouTube's servers. We obtained information including title, transcript, channel, and description for over 6 million videos.

Additionally, the channel that was obtained was not always a canonical representation. As such we also obtained a map from all the channels we observed to their canonicalized forms, also through automated web requests to YouTube.

## Analysis

### Bad recommendation rates

As we mentioned before, a "bad recommendation rate" is defined as how often YouTube recommends videos that are similar to a video that a participant has rejected. The analysis of bad recommendation rates allowed us to measure the effectiveness of YouTube's user control mechanisms in preventing unwanted recommendations. The underlying belief is that, since YouTube recommends using user control mechanisms to manage recommendations, that using such mechanisms to express what types of recommendations are unwanted should be effective in preventing such recommendations.

As such, if negative feedback is submitted for a particular "rejected" video, we consider that any subsequent recommendation similar to the rejected video is "bad". Of course, there is a lot of nuance here, and it is not clear that a single negative feedback submission should

suppress all future recommendations about a topic. It is also possible the user behavior on YouTube after feedback submission may justify recommendations on topics similar to the video for which feedback was submitted. Regardless, as these are the only tools YouTube offers to prevent unwanted recommendations, we do expect them to be effective for that purpose.

*Metrics*

We calculate the bad recommendation rate by considering all video pairs in the segment in question (for example, those contributed by participants in a particular experiment arm) and determining the proportion of them that are bad, whether as classified by our research assistants (in which case we must restrict to the subset of pairs that they have assessed) or as classified by our model.

There are alternative possible metrics. For example, we can calculate a bad recommendation proportion for each rejected video, and then aggregate those proportions among all rejected videos in the segment in question. A similar approach can be taken at the level of the participant. We tested various metrics, but found no meaningful differences in the findings, and so used the simplest approach of simply aggregating across all pairs in the segment to calculate a rate, or proportion.

### Interaction Rates

We calculate interaction rate for participants as the sum of user control interactions made, divided by number of videos watched. A user control interaction may be a press of the “Stop recommending” button, or a native YouTube control (dislike, not interested, remove from watch history, or don’t recommend channel). For the interaction rate analysis, we divide participants into two groups, those in the UX Control group (which have no “Stop recommending” button) and others (which do). We calculate the average interaction rate for each of these two groups. In the first group, the rate includes only native interactions, as they are the only options available, while the latter group includes “Stop recommending” button interactions.

### Semantic Similarity Model

Our research assistants classified 44,434 pairs but we analyzed a total of 162,983,496. For the pairs that were not classified by the research assistants we applied a machine learning model that estimates semantic similarity. Details on the model are available in our recent blog post.

# References


Ahmed, Sara. “Refusal, Resignation and Complaint.” *feministkilljoys* (blog), June 28, 2018. https://feministkilljoys.com/2018/06/28/refusal-resignation-and-complaint/.

Bishop, Sophie. “Anxiety, Panic and Self-Optimization: Inequalities and the YouTube Algorithm.” *Convergence* 24, no. 1 (February 1, 2018): 69–84. https://doi.org/10.1177/1354856517736978.

Bucher, Taina. “Cleavage-Control: Stories of Algorithmic Culture and Power in the Case of the YouTube ‘Reply Girls.’” In *A Networked Self and Platforms, Stories, Connections*. Routledge, 2018.

———. “The Algorithmic Imaginary: Exploring the Ordinary Affects of Facebook Algorithms.” *Information, Communication & Society* 20, no. 1 (January 2, 2017): 30–44. https://doi.org/10.1080/1369118X.2016.1154086.

Burrell, Jenna, Zoe Kahn, Anne Jonas, and Daniel Griffin. “When Users Control the Algorithms: Values Expressed in Practices on Twitter.” *Proceedings of the ACM on Human-Computer Interaction* 3, no. CSCW (November 7, 2019): 138:1-138:20. https://doi.org/10.1145/3359240.

DeVito, Michael A., Darren Gergle, and Jeremy Birnholtz. “‘Algorithms Ruin Everything’: #RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media.” In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3163–74. CHI ’17. New York, NY, USA: Association for Computing Machinery, 2017. https://doi.org/10.1145/3025453.3025659.

DeVito, Michael Ann. “Adaptive Folk Theorization as a Path to Algorithmic Literacy on Changing Platforms.” *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2 (October 18, 2021): 339:1-339:38. https://doi.org/10.1145/3476080.

Eslami, Motahhare, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. “‘I Always Assumed That I Wasn’t Really That Close to [Her]’: Reasoning about Invisible Algorithms in News Feeds.”

In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 153–62. CHI '15. New York, NY, USA: Association for Computing Machinery, 2015. https://doi.org/10.1145/2702123.2702556.

Harambam, Jaron, Dimitrios Bountouridis, Mykola Makhortykh, and Joris van Hoboken. "Designing for the Better by Taking Users into Account: A Qualitative Evaluation of User Control Mechanisms in (News) Recommender Systems." In *Proceedings of the 13th ACM Conference on Recommender Systems*, 69–77. Copenhagen Denmark: ACM, 2019. https://doi.org/10.1145/3298689.3347014.

Haroon, Muhammad, Anshuman Chhabra, Xin Liu, Prasant Mohapatra, Zubair Shafiq, and Magdalena Wojcieszak. "YouTube, The Great Radicalizer? Auditing and Mitigating Ideological Biases in YouTube Recommendations." arXiv, March 24, 2022. https://doi.org/10.48550/arXiv.2203.10666.

Hirsch, Tad, Kritzia Merced, Shrikanth Narayanan, Zac E. Imel, and David C. Atkins. "Designing Contestability: Interaction Design, Machine Learning, and Mental Health." In *Proceedings of the 2017 Conference on Designing Interactive Systems*, 95–99. DIS '17. New York, NY, USA: Association for Computing Machinery, 2017. https://doi.org/10.1145/3064663.3064703.

Lewis, Becca. "Alternative Influence." Data & Society. Data & Society Research Institute, September 18, 2018. https://datasociety.net/library/alternative-influence/.

Nagel, Emily van der. "'Networks That Work Too Well': Intervening in Algorithmic Connections." *Media International Australia* 168, no. 1 (August 1, 2018): 81–92. https://doi.org/10.1177/1329878X18783002.

Saldaña, Johnny. "An Introduction to Codes and Coding" In *The Coding Manual for Qualitative Researchers*, 3rd edition. Sage, 2016.

Thorburn, Luke. "What Does It Mean to Give Someone What They Want? The Nature of Preferences in Recommender Systems." *Understanding Recommenders* (blog), May 11, 2022. https://medium.com/understanding-recommenders/what-does-it-mean-to-give-someo

ne-what-they-want-the-nature-of-preferences-in-recommender-systems-82b5a1559157.

Tufekci, Zeynep. “YouTube’s Recommendation Algorithm Has a Dark Side.” Scientific American, April 1, 2019. https://www.scientificamerican.com/article/youtubes-recommendation-algorithm-has-a-dark-side/.

Witzenberger, Kevin. “The Hyperdodge: How Users Resist Algorithmic Objects in Everyday Life.” *Media Theory* 2, no. 2 (December 17, 2018): 29–51.

Zhao, Zhe, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, and Ed Chi. “Recommending What Video to Watch next: A Multitask Ranking System.” In *Proceedings of the 13th ACM Conference on Recommender Systems*, 43–51. RecSys ’19. New York, NY, USA: Association for Computing Machinery, 2019. https://doi.org/10.1145/3298689.3346997.

## Acknowledgements

This report was written by Becca Ricks and Jesse McCrosky.

Mozilla's design team created the graphics and illustrations in this report: Rebecca Lam, Sabrina Ng, Kristina Shu, and Nancy Tran.

We are grateful to Brandi Geurkink for her leadership in the development and design of this research project and RegretsReporter, as well as her contributions to Mozilla's approach to crowdsourced research.

Thank you to our Mozilla colleagues who gave feedback on and contributed to this report, including: Carys Afoko, Stefan Baack, Christian Bock, Ashley Boyd, Maximilian Gahntz, Brandi Geurkink, Jenn Hodges, Tracy Kariuki, Glenda Leonard, Claire Jenifer Pershan, Udbhav Tiwari, D'Andre Walker, and Kevin Zawacki.

We appreciate Georgia Bullen and Ame Elliott from Simply Secure for their design research that informed this work. Thank you to Semyon Bondarenko and Adel Salakh from Zetico for their work developing the RegretsReporter extension and Aapo Tanskanen from ThoughtWorks Finland for assisting with the development of the machine learning model.

We're grateful to Chico Camargo and Ranadheer Malla from the University of Exeter for leading the analysis of RegretsReporter data. Thank you to the research assistants at the University of Exeter for analyzing the video data: Josh Adebayo, Sharon Choi, Henry Cook, Alex Craig, Bee Dally, Seb Dixon, Aditi Dutta, Ana Lucia Estrada Jaramillo, Jamie Falla, Alice Gallagher Boyden, Adriano Giunta, Lisa Greghi, Keanu Hambali, Clare Keeton Graddol, Kien Khuong, Mitran Malarvannan, Zachary Marre, Inês Mendes de Sousa, Dario Notarangelo, Izzy Sebire, Tawhid Shahrior, Shambhavi Shivam, Marti Toneva, Anthime Valin, and Ned Westwood.

Finally, we're so grateful for the 22,722 RegretsReporter participants who contributed their data, the 2,757 people who responded to our survey, and the five people who agreed to be interviewed. Without all of you, this research would not have been possible.


## Annex: Examples of video pairs recommended

Link to Addendum Report PDF

Link to public JSON endpoint