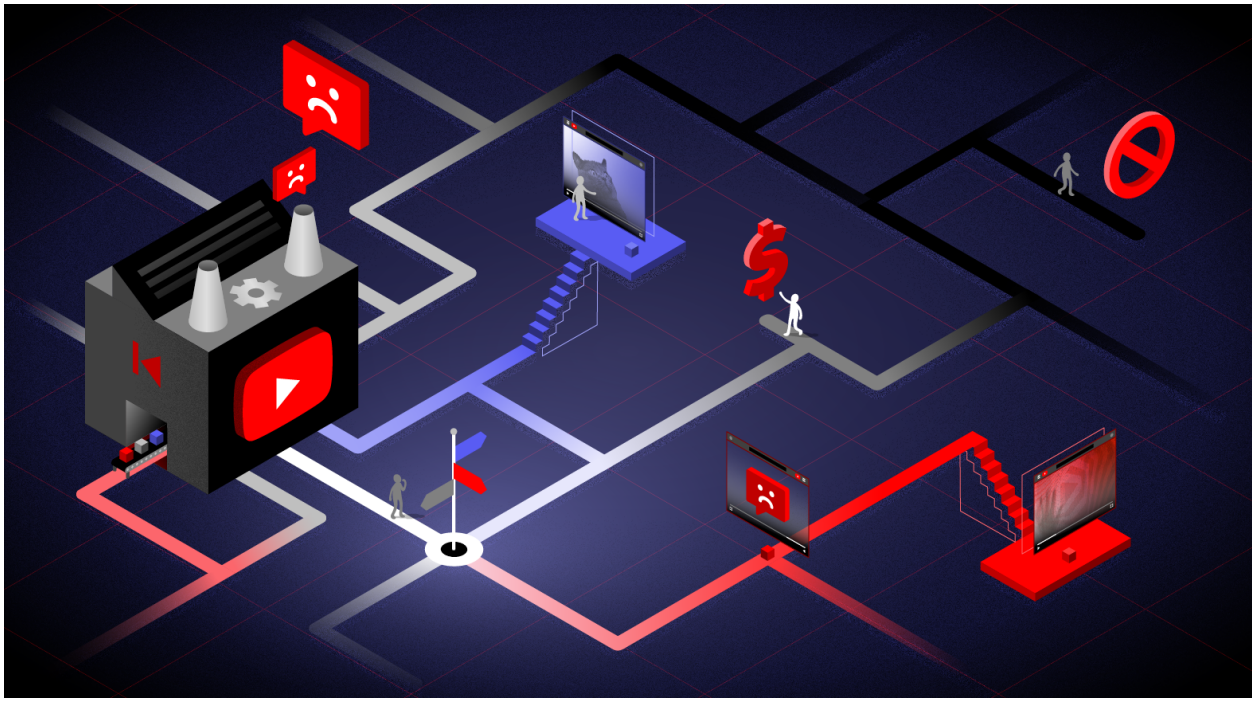
moz://a

YouTube Regrets

A crowdsourced investigation into YouTube's recommendation algorithm



July 2021

Table of Contents

Executive Summary 4

Introduction 6

What do we mean by 'regret'? 7

Findings 8

- YouTube Regrets are disparate and disturbing 8
- The algorithm is the problem 13
- Non-English speakers are hit the hardest 19

Recommendations 24

- Our recommendations to YouTube and other platforms 24
- Our recommendations to policymakers 27
- Our recommendations to people who use YouTube 28

Conclusion 29

Methodology 30

- Research Questions 30
- RegretsReporter Extension 30
- People-powered dataset 32
- Analysis methods 33
- Disclosures 34

References 35

Acknowledgements 38

Annex: Examples of YouTube Regrets by category 39

Executive Summary

YouTube is the second-most visited website in the world, and its algorithm drives 70% of watch time on the platform—an estimated 700 million hours[1] every single day. For years, that recommendation algorithm has helped spread health misinformation, political disinformation, hateful diatribes, and other regrettable content to people around the globe. YouTube's enormous influence means these films reach a huge audience, having a deep impact on countless lives, from radicalization to polarization.[2] And yet YouTube has met this criticism with inertia and opacity.

In 2020, after years of advocating for YouTube to be more transparent about their recommendation algorithm and allow researchers to study the platform, Mozilla responded to the platform's inaction by empowering its users to speak up instead. We launched RegretsReporter, a browser extension and crowdsourced research project to better understand the harms that YouTube's algorithm can inflict on people.

37,380 YouTube users stepped up as YouTube watch dogs, volunteering data about the regrettable experiences they have on YouTube for Mozilla researchers to carefully analyze. As a result, Mozilla gained insight into a pool of YouTube's tightly-held data in the largest-ever crowdsourced investigation into YouTube's algorithm. Collectively, these volunteers flagged 3,362 regrettable videos, coming from 91 countries, between July 2020 and May 2021.

This report highlights what we learned from our RegretsReporter research. Specifically, we uncovered three main findings:

- **YouTube Regrets are disparate and disturbing.** Our volunteers reported everything from Covid fear-mongering to political misinformation to wildly inappropriate "children's" cartoons. The most frequent Regret categories are misinformation, violent or graphic content, hate speech, and spam/scams.
- **The algorithm is the problem.** 71% of all Regret reports came from videos recommended to our volunteers by YouTube's automatic recommendation system. Further, recommended videos were 40% more likely to be reported by our volunteers than videos that they searched for. And in several cases, YouTube

[1] YouTube states that one billion hours of video are watched on YouTube daily. At CES 2018, Neal Mohan, YouTube's CPO stated that at least 70% of this time is from AI-driven recommendations. To the extent these numbers are currently accurate, AI-driven recommendations on YouTube are responsible for 700 million hours of watchtime per day.

[2] Investigations by Mozilla, the Anti-Defamation League, the New York Times, the Washington Post, the Wall Street Journal, and several other organizations, academic researchers and publications have revealed how YouTube can misinform, polarize, and radicalize users.

recommended videos that actually violate their own Community Guidelines and/or were unrelated to previous videos watched.

- **Non-English speakers are hit the hardest.** The rate of YouTube Regrets is 60% higher in countries that do not have English as a primary language (with Brazil, Germany and France being particularly high), and pandemic-related Regrets were especially prevalent in non-English languages.

Within this report, we unpack each of these findings with data analysis and ample case studies. Indeed, just a few examples from Mozilla's research paint a vivid picture. One volunteer reported the animated video "Woody's Got Wood," which is a sexualised parody of the children's film "Toy Story." Another volunteer was recommended a debunked conspiracy theory video about U.S. Representative Ilhan Omar and election fraud. And a third was recommended a video called "Blacks in Power Don't Empower Blacks," which features racist and offensive language and ideas.

Mozilla doesn't want to just diagnose the problem, however—we want to help solve it. In this report's "Recommendations" section is clear guidance for YouTube, other internet platforms, policymakers, and the public. Some of those recommendations include:

- **Platforms must** enable researchers to audit recommendation systems.
- **Platforms must** publish information about how recommendation systems work, as well as transparency reports that give sufficient insight into problem areas and progress over time.
- **Policymakers must** require YouTube to release information and create tools that enable independent scrutiny of their recommendation algorithms.
- **Policymakers must** protect researchers, journalists, and other watchdogs who use alternative methods than those provided by platforms to investigate them.
- **People should** update data settings on YouTube and Google to make sure they have the right controls in place for themselves and their families.

This report also includes a comprehensive methodology section describing Mozilla's research. It details our research questions, browser extension workings, analysis methods, and more.

Introduction

Sam was a 13-year-old boy who suffered from depression—and in his weakest moment he was recruited to join the alt-right. Sam's radicalization happened in part because he went down the rabbit hole of recommended content on YouTube and other social media sites. In an interview with Mozilla, Sam's mother explained that "Platforms like YouTube and others seem to do everything possible to avoid accountability. The companies hide behind all sorts of arguments related to free speech, but this has always seemed to me to be a convenient way to avoid their responsibility and bring more money in. Children are especially vulnerable, especially children who look to the internet for information and support."

Sam's story is not an isolated one. For years now, researchers and investigative journalists have documented how YouTube recommendations can lead people to the internet's darkest and most extreme corners. While YouTube is not *solely* responsible for complex problems like radicalization, evidence suggests that the company's recommendation algorithm plays an outsized part. The algorithm steers people towards this content and, once people are "in" the rabbit hole, it offers up more and more extreme ideas.

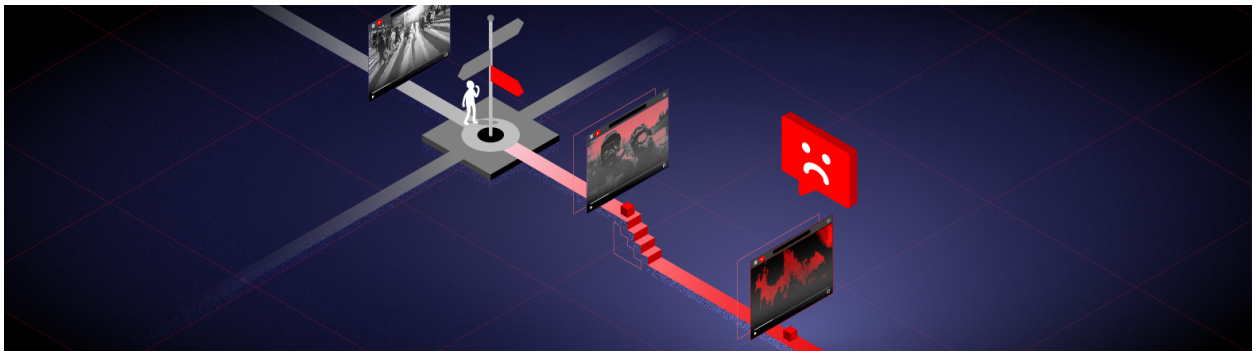
This problem with YouTube's recommendation algorithm is part of a bigger story about the opaque, mysterious influence that commercial algorithms can have on our lives. **YouTube's algorithm drives an estimated 700 million hours of watch time every single day, and yet the public knows very little about how it works.** We have no official ways of studying it. When stories like Sam's emerge, there is no way to understand what happened and what could be changed to prevent the same thing from happening again. At its core, YouTube's algorithm works in the interest of YouTube, not the public.

Since 2019, Mozilla has advocated for transparency into YouTube's recommendation algorithm. Our campaign, called YouTube Regrets, highlighted 28 stories like Sam's—stories of people whose lives were derailed, and in some cases forever altered, by their experiences on YouTube. We published these stories to raise awareness of the human impact of untrustworthy algorithms— and to pressure YouTube to be more responsible and open about how their algorithm works. A few years later, YouTube has yet to do so.

That's why we built RegretsReporter. RegretsReporter is a crowdsourcing tool that scales up the original YouTube Regrets campaign. It empowers Mozilla's volunteer community to contribute data about the experiences that they have on YouTube. RegretsReporter is akin to other browser extensions that use alternative methods to

study some of the most consequential algorithms in the interest of the public. These tools are not substitutes for meaningful transparency to people and public institutions, but they are the best option that we have to hold companies accountable for their impact on people's lives.

Our intention with this research is to share what we have found through RegretsReporter. Our hope is that these findings—which we believe to be the tip of the iceberg—will convince the public, and those who serve the public interest, of the urgent need for full transparency into YouTube's algorithms.



What do we mean by 'regret'?

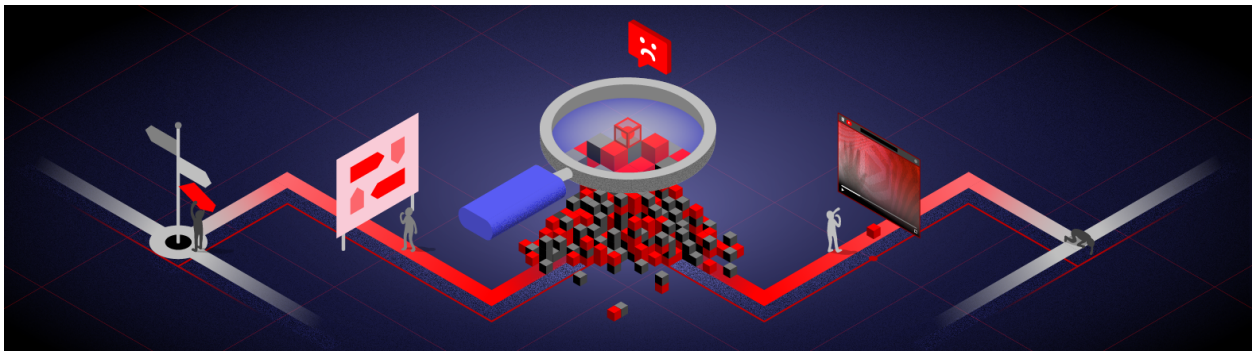
The concept of a "YouTube Regret" was born out of a global, crowdsourced campaign that Mozilla developed in 2019. We collected stories about YouTube leading people down dangerous and bizarre pathways, particularly as a result of their recommendation algorithm. That campaign was our first attempt to find out what kind of experiences people were having on YouTube. We gave no specific guidance on what these stories should be about; submissions were self-identified as regrettable. Since then, Mozilla has intentionally steered away from strictly defining what "counts" as a YouTube Regret, in order to allow people to define the full spectrum of bad experiences that they have on YouTube.

Our people-powered approach centres the lived experiences of people, and particularly the experiences of vulnerable and/or marginalized people and communities, to add to what is often a highly legalistic and theoretical discussion about what constitutes harmful content online. Our qualitative analysis was performed primarily by research assistants from the University of Exeter who used YouTube's Community Guidelines, rules that define what is allowed to be on the platform, to assess the videos and determine whether or not a video should be on the platform and/or recommended by YouTube (described further in the "Analysis Methods" section of this report). When comments were attached to videos reported to us, we used them to inform our

analysis. This methodology does not yield objective distinctions—we are surfacing subjective impressions as a starting point for further conversation and studies.

While our research uncovered numerous examples of hate speech, debunked political and scientific misinformation, and other categories of content that would likely (or have actually been found to) violate YouTube's Community Guidelines, it also uncovered many examples that paint a more complicated picture of online harm. Many of the videos reported to us may fall into the category of what YouTube calls "borderline content"—videos that "skirt the borders" of their Community Guidelines without actually violating them. Since YouTube provides no transparency into how they define and classify borderline content, it is impossible to verify this assumption. Our research also suggests that regrettable, or harmful, experiences online are often the result of how content is targeted at someone over time, which is difficult to understand or remedy by looking at individual videos.

The recommendations that we made in this report—calling for robust transparency, scrutiny, and giving people control of recommendation algorithms—are essential for shifting power into the hands of people who are ultimately impacted by these systems.



Findings

1. YouTube Regrets are disparate and disturbing

The Summary

"Every day, millions of people come to YouTube to be informed, inspired or just plain delighted" —YouTube, 2020, "How YouTube Works"

- **Reports span everything from Covid fear-mongering to political misinformation to wildly inappropriate "children's" cartoons.** The most frequent categories are misinformation, violent or graphic content, hate speech, and spam/scams.

The Story

Public perception of internet platforms has soured in recent years. Facebook and Twitter — which once enjoyed the public's adulation — are now seen as places where hateful content, incivility, and disinformation can flourish.

YouTube, however, has managed to maintain a mostly positive public image — at least compared to its peers. YouTube is still perceived by many as a platform where goofy reaction videos and helpful DIY content thrive, and a community where well-intentioned creators can flourish.

It is true that YouTube can be wonderful. The platform provides education and entertainment to millions of people everyday. But just like its peers, YouTube also has a dark side — even if it's not as apparent in the public's imagination.

In 2019, Mozilla set out to examine this dark side — to better understand the regrettable videos that more and more users were encountering and talking about. We labeled this type of content "YouTube Regrets," and put out a public call for people's experiences. The response was alarming.

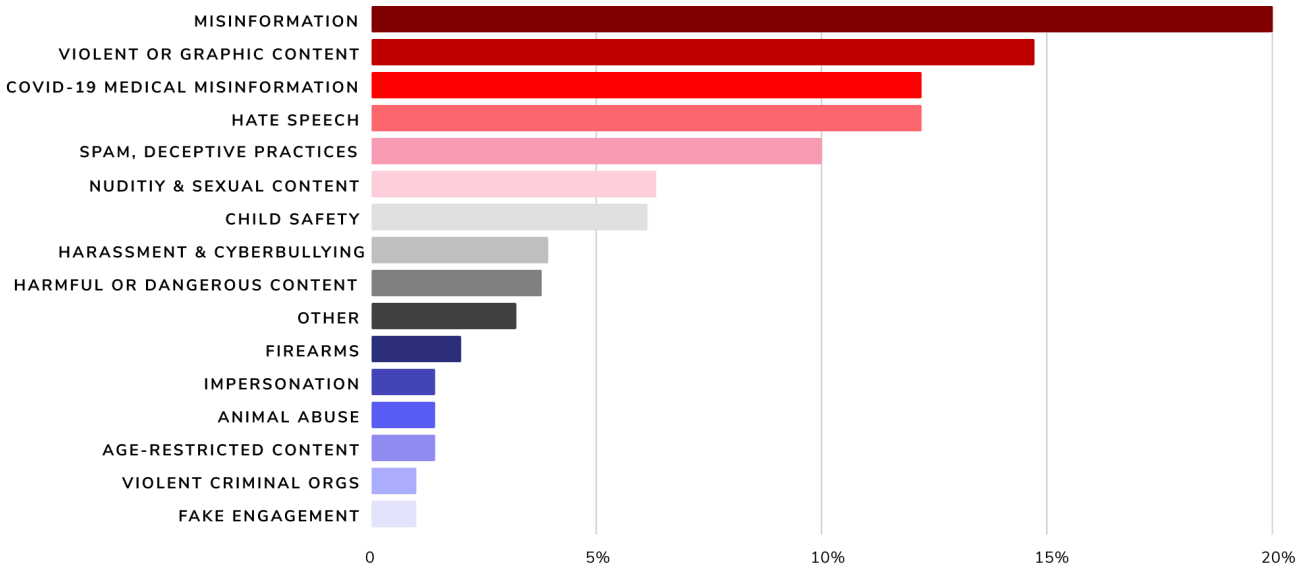
Living alongside those goofy and helpful videos were a mountain of bizarre, and even harmful, videos. Mozilla heard from one person who fell down a rabbit hole of traffic collision content, watching "videos where people clearly didn't survive the accident." A 10-year-old girl, seeking dance videos, ended up encountering videos promoting extreme dieting. "She is now restricting her eating and drinking," her parent told Mozilla. Another person sought out affirming LGBTQ+ content, but instead found hateful, anti-LGBTQ+ videos. "I can only imagine how harmful this would be to people still figuring out their identity," they wrote to Mozilla.

There were many other YouTube Regrets: videos championing pseudo-science, promoting 9/11 conspiracies, showcasing mistreated animals, encouraging white supremacy. It became clear to us at Mozilla that YouTube had a real problem. So we decided to investigate further, to understand the experiences that real people were having on the platform, and the role that YouTube's recommendation algorithm could be playing in amplifying these videos.

The Data

We estimate that 12.2% of reported videos (95% confidence interval of 10.4 to 14.2%) either "should not be on YouTube" or "should not be proactively recommended," based on YouTube's Community Guidelines. Among these videos, the categories identified are shown in the chart below. While the research assistants who classified these videos are not experts in identifying and classifying misinformation and categories of illegal

content, this classification is based on their best judgment and what they personally found to be troubling and worthy of deeper scrutiny. We find that the most frequent categories per this classification are misinformation, violent or graphic content, and Covid-19 misinformation (which we categorize separately as it is of special interest), then hate speech and spam/scams. Other notable categories include child safety, harassment & cyberbullying, and animal abuse. It is clear that YouTube Regrets occur for a multitude of reasons, but that misinformation is a dominant problem. If the Covid-19 misinformation is combined into a single category with the rest of the misinformation, it makes up approximately a third of all categorized Regrets.



We detail some of the most notable examples from across these categories below, and our Annex contains more examples.

A volunteer was recommended a video titled “Omar Connected Harvester SEEN Exchanging $200 for General Election Ballot. ‘We don’t care illegal,’” which spreads an unfounded claim about U.S. Representative Ilhan Omar and voter fraud in the 2020 U.S. election. In a comment, the volunteer wrote that they kept getting recommended extreme right wing channels, despite watching primarily wilderness survival videos, suggesting that this recommendation was part of a larger pattern whereby politically extreme content was continuously targeted at the volunteer.

Omar Connected Harvester SEEN Exchanging $200 for General Election Ballot."We don't care illegal."
542686 views - Sep 28, 2020

A volunteer reported a video titled “BILL GATES HIRED BLM “STUDENTS” TO COUNT BALLOTS IN BATTLEGROUND STATES” which spreads the unfounded and accusatory claim that the founder of Microsoft hired young people from ethnic minorities to commit election fraud.

BILL GATES HIRED BLM “STUDENTS” TO COUNT BALLOTS IN BATTLEGROUND STATES
14 views - Nov 21, 2020

A volunteer reported a video called “The Elites Who Control You” which features a person impersonating a politician or 'elite', stating that they use and lie about Covid-19 as a fear tactic in order to control the masses.

The Elites Who Control You
743641 views - Dec 1, 2020

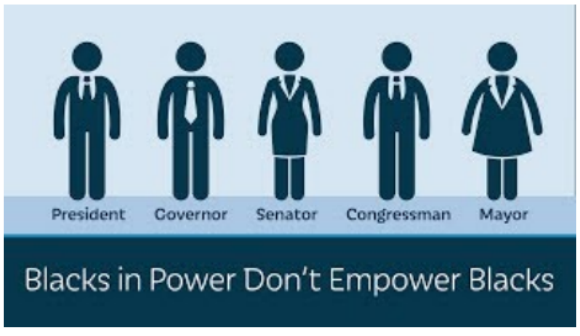
A volunteer reported a video called “7 jokes ending in tragedy”. The stories in the video are very distressing and the thumbnail used for the video is disturbing and shows a lot of blood.

7 SCHERZI finiti in TRAGEDIA
2193169 views - 6 apr 2018

A volunteer reported the animated video “Woody’s Got Wood,” which is a sexualised parody of the children’s film “Toy Story”.

Woody's Got Wood Animated.mp4
22251 views - Sep 17, 2018

A volunteer was recommended a video called “Blacks in Power Don’t Empower Blacks” which features offensive and racist language and ideas.



Blacks in Power Don't Empower Blacks
2365794 views - Mar 26, 2018

A volunteer was recommended a video called “El Arca - It’s Literally Furry Noah’s Arc” which they categorised as being bizarre. The video suggests that the children’s film it is discussing contains sexual and adult content and talks about this in depth.

El Arca - It's Literally Furry Noah's Arc
119145 views - 22 Apr 2021

A video called “‘Biggest fraud’ in US history—up to 300,000 fake people voted in Arizona election: expert | NTD” was recommended to a volunteer.

'Biggest fraud' in US history—up to 300,000 fake people voted in Arizona election: expert | NTD
965541 views - 1 Dec 2020

A volunteer reported a video called “[Documentary 2020] [Adrenochrome — The Darkest Drug of Them All!]” which features a QAnon conspiracy theory about the harvesting of children’s blood.

[Documentary 2020] [Adrenochrome - The Darkest Drug of Them All!]
108279 views - Premiered Mar 23, 2020

2. The algorithm is the problem

The Summary

"*When recommendations are at their best, they help users find a new song to fall in love with, discover their next favorite creator, or learn that great paella recipe. That's why we update our recommendations system all the time—we want to make sure we're suggesting videos that people actually want to watch.*" —YouTube, 2019, "Continuing our work to improve recommendations on YouTube"

- **Around 9% of recommended Regrets have since been taken down from YouTube, including several which were taken down after they were recommended for violating the platform's own Community Guidelines.** Recommended Regrets that were later removed from YouTube had a collective 160 million views at the time that they were reported.
- **Recommendations are disproportionately responsible for YouTube Regrets.** 71% of all Regret reports came from videos recommended to our volunteers, and recommended videos were 40% more likely to be regretted than videos searched for.
- **YouTube Regrets tend to perform well on the platform.** Reported videos acquired 70% more views per day than other videos watched by our volunteers.
- **Regrettable recommendations are often unprovoked.** In 43.3% of cases where we have data about trails a volunteer watched before a Regret, the recommendation was completely unrelated to the previous videos that the volunteer watched.

The Story

In 2020, an investigation by HuffPost told the story of an 11-year-old girl named Allie. Allie was unknowingly filming skits acting out sexual fantasies of predators who found her YouTube channel and would make requests, like asking her to pretend that she fainted on camera. What's worse, the HuffPost investigation concluded that YouTube's recommendation algorithm was actually *helping* predators find channels like Allie's. This disturbing investigation revealed the downside of "algorithmic promotion," and what can happen when seemingly "innocuous" videos that do not violate YouTube's Community Guidelines—like videos of children bathing, doing gymnastics, spreading their legs—are algorithmically grouped and targeted to people who are heavily engaged by that content.

Allie's story is a disturbing real-life example of how YouTube's automated recommendations can make videos "go viral" without considering the context in which those videos are being viewed. Since YouTube is an open content recommendation system, meaning that it recommends user-generated videos without undergoing a vetting process, the decisions made by the algorithm can be more consequential (and more risky) than platforms like Netflix, which only host and recommend content that has been vetted by humans.

Many platforms which automate recommendations of user-generated content rely on algorithms to identify and moderate content. But these algorithms are incredibly limited, and they have to work alongside algorithms designed to recommend content that will "engage" people, based on signals like how long they watch a video. Taken together, this means that algorithms must weigh, among other things, predictions about how likely a video is to 'engage' someone with predictions about how likely a video is to violate YouTube's Community Guidelines. That's a big responsibility, and one that can pit commercial incentives against peoples' welfare.

Our research turned up several cases where YouTube's algorithm seemed to do a poor job of balancing these dual responsibilities. We identified several cases where YouTube's algorithm recommended videos that actually violated their own Community Guidelines and other content policies and were later taken down from the platform, only after racking up millions of views. Our research also found that the recommendations surfaced by YouTube's algorithm were disproportionately responsible for the bad experiences that our volunteers had on YouTube, compared to search or other ways that volunteers could come across videos on YouTube. We also found that YouTube Regrets that our volunteers reported had quickly gained significant viewership compared to other videos on the platform. These examples raise serious questions

about how YouTube prioritizes the various decisions that their algorithm must make, and what trade-offs arise along the way.

YouTube's algorithm is more puppeteer than puppet. What it recommends, why, and how those decisions are made really matters—especially in situations like Allie's. During a US Senate Judiciary Committee hearing in April 2021, Senator Chris Coons (D-DE) asked YouTube to commit to releasing information about how many times they *recommend* a given video, not just how many times that video is *viewed*. This information is critical to understanding the role that YouTube's algorithm plays in the virality of content, and could've helped to understand how Allie's videos, for example, were recommended by YouTube. The YouTube representative would not commit to sharing this critical information.

The Data

Several of the recommended Regrets reported to us have since been taken off of YouTube. YouTube uses a combination of people and machine learning to find content that violates their policies and take it down from the platform. Yet these systems often make mistakes, and when combined with automatic recommendations, in some cases YouTube ends up amplifying content that violates their own rules. Among videos recommended to our volunteers and then reported to us, a total of 189 have been taken down as of June 1, 2021.

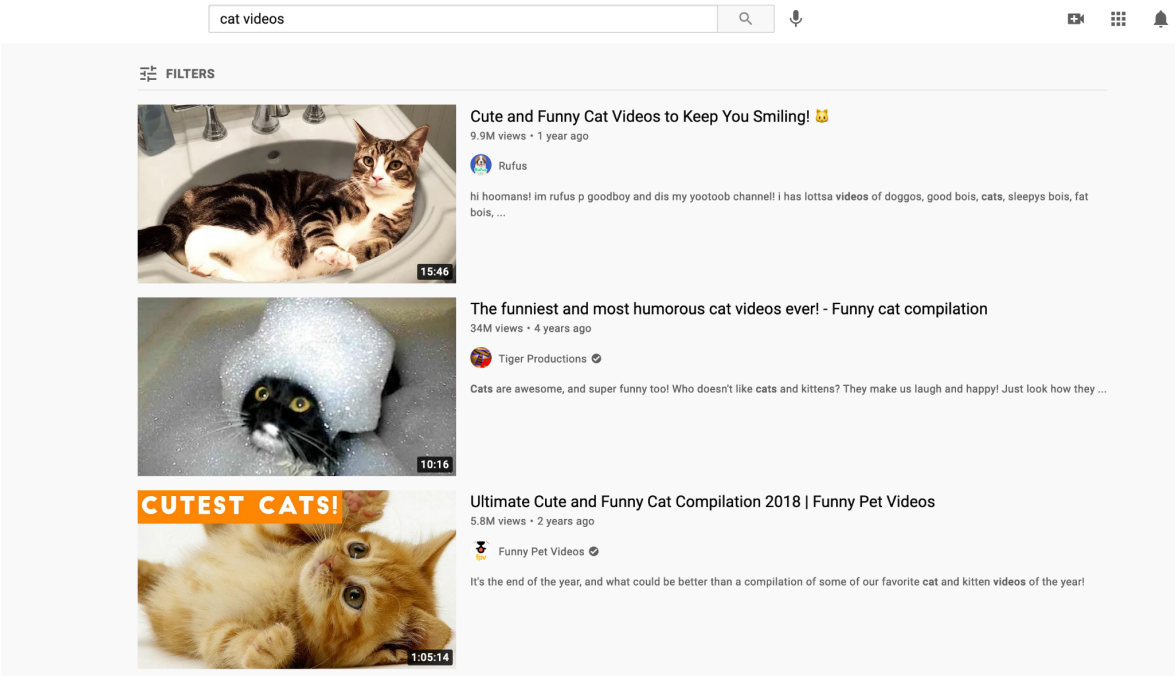
In some cases, YouTube provides a specific reason for why the video was taken down (e.g. Community Guidelines violations, copyright infringement, hate speech laws, or that the video was deleted or made private by the uploader). But in nearly 40% of the cases we analyzed, videos were simply labeled "video unavailable," without identifying the reason why they were taken down.

Collectively, these videos had racked up a collective 160 million views, an average of 760 thousand views per video, accrued over an average of 5 months that the videos had been up at the time they were reported. It is impossible to tell how many of those views were a result of YouTube's recommendation algorithm, since YouTube does not make this data public, but we know that they were recommended at least once (to the volunteer who reported them).

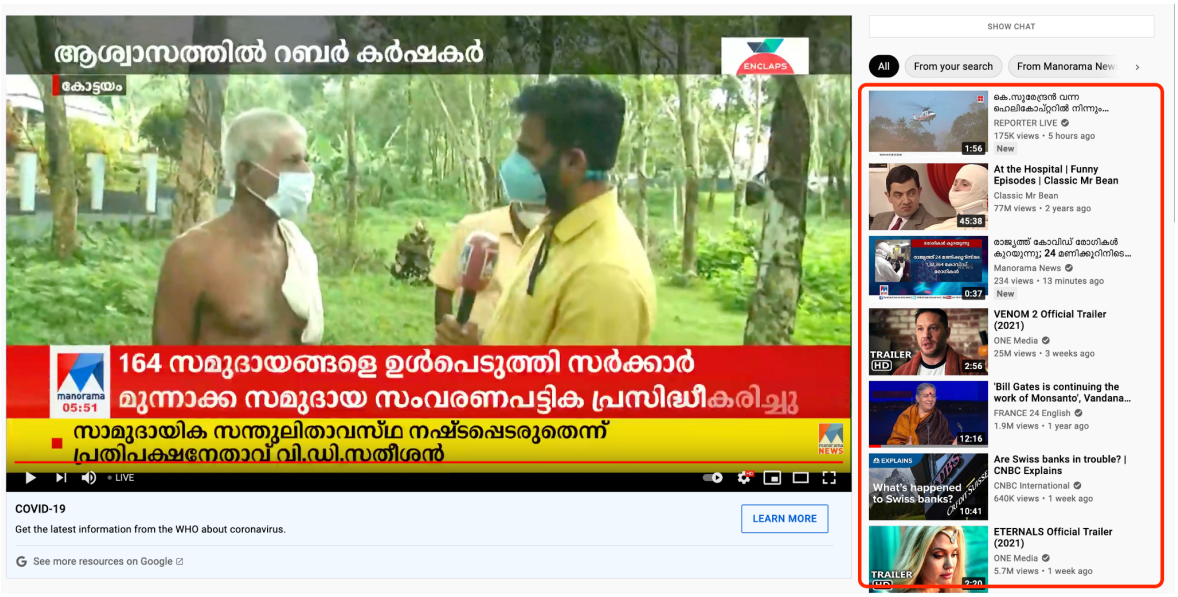
Our data allows us to calculate Regret rates, or the proportion of regretted videos out of the videos that a volunteer watches. This rate can be broken down by various factors—for example, the entry point that took the volunteer to the video (e.g. search, recommendations, direct link). We analyzed two main entry points of interest:

- Searches represent cases in which the volunteer enters a search query and then watches one of the resulting videos that YouTube serves in response to the search.
- Recommendations represent cases in which YouTube is proactively suggesting content to the volunteer.

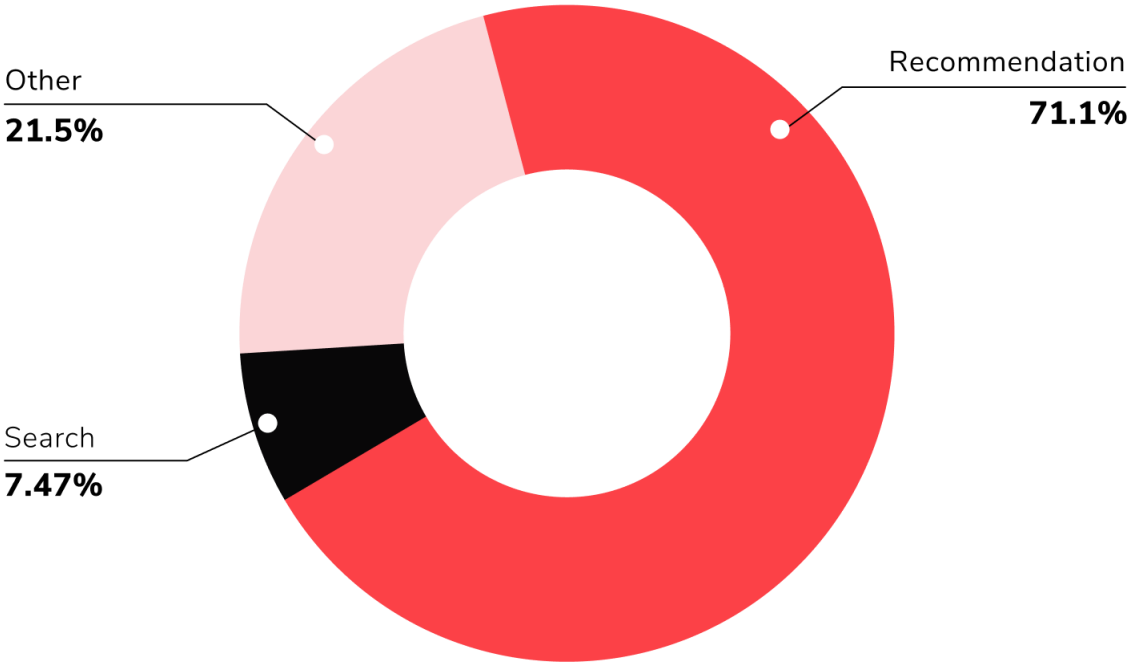
Search



Recommendation



By comparing the Regret rate for recommendations to the rate for searches, it is clear that recommendations are disproportionately responsible for YouTube Regrets. The rate for searches is 9.6 Regrets out of every 10,000 videos watched, while the rate for recommendations is 13.9 Regrets out of every 10,000, or 40% higher. Recommendations also represent 71% of the regretted videos reported to us.



For reference, YouTube's published Violative View Rate, which represents the view rate for videos that violate YouTube's policies, is around 17 videos out of every 10,000. The fact that our rates (13.9 out of 10,000 for recommendations) are not in excess of YouTube's own published findings adds some support to the validity of our concept of regretted videos.

Reported videos also accumulated views faster than other videos. At the time they were reported, YouTube Regrets had a median 5,794 views per day they were on the platform, which is 70% higher than that of other videos our volunteers watched (in the trail prior to reported videos), which had a median of only 3,312 views per day.

In analyzing these "trail" videos, we noticed that some of these trails were totally unrelated to what the volunteer was watching previously. To interrogate this trend, we had research assistants classify the trail videos that accompanied recommended Regrets to determine whether the recommendations seemed related or not. Among recommendations for which we have data on the trail of videos the volunteer followed,

in 43.3% of cases, the regretted recommendation was completely unrelated to the previous videos that the volunteer watched.

For example, one volunteer was watching an Art Garfunkel music video and was recommended a video titled “Trump Debate Moderator EXPOSED as having Deep Democrat Ties, Media Bias Reaches BREAKING Point.” Comments that volunteers submitted along with these videos suggested that volunteers were exhausted and annoyed by being recommended videos that were unrelated to what they watched previously, particularly when those videos went against their beliefs or presented a viewpoint that they disagreed with.

The video being reported

Trump Debate Moderator EXPOSED as having Deep Democrat Ties, Media Bias Reaches BREAKING Point

166310 views - Oct 18, 2020

Video history for this session

Art Garfunkel and his son cover The Everly Brothers live in Napa, May 12, 2019 (4K)

After watching a Mozilla video called “Memes, Misinfo, and the Election - October 12, 2020,” a volunteer was recommended a video called “Global Warming: Fact or Fiction? Featuring Physicists Willie Soon and Elliott D. Bloom.” In a comment, the volunteer wrote that they were surprised to see a video denying climate change suggested after they were watching a Mozilla video.

The video being reported

Global Warming: Fact or Fiction? Featuring Physicists Willie Soon and Elliott D. Bloom

592557 views - Aug 16, 2019

Video history for this session

Memes, Misinfo, and the Election - October 12, 2020

After watching videos about the U.S. military, a volunteer was recommended a video called “Man humilitates feminist in viral video,” which is misogynistic, sexist and discriminatory against women.

The video being reported

Man hum1liates f3m1n1st in v1ral video
368535 views - Mar 6, 2021

Video history for this session


Top 5 Things US Government Denies (Marine Reacts) [Via Recommendation]


Marine Reacts - Can the US Defend an Invasion from Abroad? [Via Recommendation]


North Korean Soldier meet U.S. Soldier For The First Time

3. Non-English speakers are hit the hardest

The Summary

“YouTube’s Community Guidelines are enforced consistently across the globe, regardless of where the content is uploaded.” —YouTube, 2021, “YouTube Community Guidelines Enforcement Transparency Report”

- **Non-English speakers are hit the hardest.** The rate of YouTube Regrets is 60% higher in countries that do not have English as a primary language.
- **Pandemic-related reports are particularly prolific in non-English languages.** Among English-language YouTube Regrets that our research assistants determined should not be on or recommended by YouTube, 14% are pandemic-related. For non-English Regrets, the rate is 36%.

The Story

In 2017, approximately 700,000 people from the Rohingya community, an ethnic Muslim minority group in Myanmar, were forced to flee into neighbouring Bangladesh to escape killings, mass rapes and burning of their villages. Months later, the UN Independent International Fact-Finding Mission on Myanmar released a report highlighting how Facebook had created an “enabling environment” for atrocities committed by the Myanmar military in Rakhine State, home to Rohingya and other ethnic minorities. A Reuters investigation found “poisonous posts calling the Rohingya or other Muslims dogs, maggots and rapists, suggesting they be fed to pigs, and urge

they be shot or exterminated." The hate speech that thrived and amplified on Facebook was generally cited as a significant precursor to the genocide that resulted. As Reuters noted, almost all of these posts were in the main local language, Burmese. And almost all of them violated Facebook's policies.

The point is: The policies of platforms and the enforcement of those policies are very different matters. And unfortunately, platform policies are enforced very differently in different parts of the world. One reason for this is because algorithms used to detect policy violations and to recommend videos rely on language-specific machine learning models. That means that companies have to train their algorithms using data from different languages and country contexts. However, many platforms prioritize training on English-language data, which is why they perform better in these contexts.

While YouTube has not released data about this, they acknowledge that their recommendation systems contribute to the spread of "borderline content," content which skirts the borders of their Community Guidelines without actually violating them. When the company announced policy changes intended to address this problem, they focused first on the United States and other English-speaking countries. More than two years later they announced that they rolled out these changes in every market where they operate, though to our knowledge the company has never released any metrics about the success of these efforts anywhere outside of the United States.

Our research found that YouTube Regrets were 60% higher in non-English speaking countries. Moreover, we found that for the videos that our research assistants determined should not be on or recommended by YouTube, videos related to the pandemic were more prevalent in non-English languages.

When people all over the world use YouTube to access critical information, such as health information during a global pandemic, the consequences of this variability could be disastrous. A recent report from EU Disinfo Lab revealed how a French-language Covid-19 conspiracy video called "Hold Up" (akin to the English-language video "Plandemic"), which violates YouTube's policies on medical misinformation, was still available on the platform more than six months after it was released and had generated millions of views. This charts with research produced by the Election Integrity Partnership which found that during the 2020 U.S. elections, non-English election-related misinformation on Twitter and YouTube lacked labels or other enforcement action, and it was not clear what languages YouTube's policies were being enforced in.

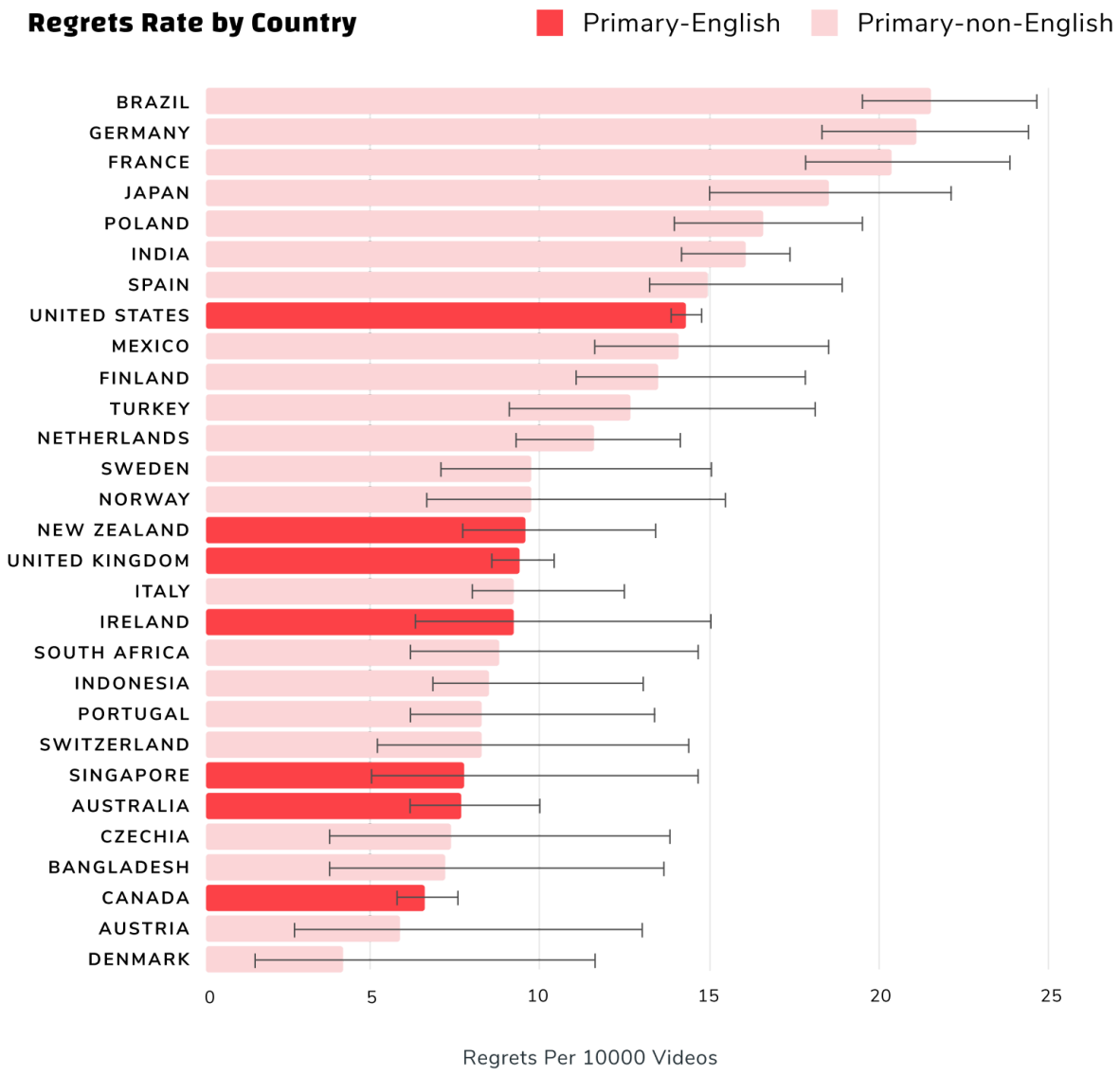
Someone's language should not determine whether they are protected online or not. Platforms like YouTube have a responsibility to everyone who uses their platform, not only people who live in English-speaking countries and speak English. The systemic

lack of transparency and attention to this issue reinforces the importance of transparency tools like RegretsReporter in providing community oversight into how platforms are enforcing (or not enforcing) their own policies consistently.

The Data

We calculated Regret rates by country using GeoIP lookup, which tells us which country our volunteers are accessing YouTube from. We found that the highest Regret rates, over 20 Regrets per 10,000 videos watched, were in Brazil, Germany, and France. In fact, the top seven countries all have a non-English primary language and the U.S. is the only primary-English country in the top 14 highest Regret rates. Among the bottom 14 of the countries shown, with the lowest Regret rates, there are five primary-English countries.

It is clear that Regret rates are lower in primary-English countries. In fact, among countries classified as having English as a primary language, the rate is 11.0 Regrets per 10,000 videos watched (95% confidence interval is 10.4 to 11.7). In countries with a non-English primary language, the rate is 17.5 Regrets per 10,000 videos watched (95% confidence interval is 16.8 to 18.3). Thus we see a clear statistically significant difference, with countries with non-English primary languages having 60% higher Regret rates.



The error bars show a 95% confidence interval for the Regret rate. Due to limited data in many countries, these confidence intervals are wide in some cases, but the findings are all statistically significant. Note that we only show the countries for which we have enough data to make more accurate estimates, as measured by confidence interval width.

We classified the primary language of a country using the CIA World Factbook; we choose the primary language as that with the highest proportion of people speaking it as a first language when possible, or with the highest prevalence, when first-language data is not available. Our classification covers the countries providing the majority of our data (responsible for 94% of Regret reports) as either English or non-English primary language.

Pandemic-related content was a notable segment of our Regret reports and was particularly rampant in non-English videos. We analyzed the videos that our research assistants determined should either not be on YT or not be recommended and found that among those videos in English, only 14% are pandemic-related. But among regretted videos that are not in English, the rate is 36%—more than a third of these Regrets in languages other than English are related to the pandemic.

For example, in this Portuguese livestream video called “Corrupt biden, judicialization of the vaccine and enough of hysteria!,” this person defends conspiracy theories about the coronavirus.

The video being reported

Biden corrupto, judicialização da vacina e chega de histeria!

6 views - Stream iniciado há 57 minutos

This Slovak video was reported to us, which talks about how the pandemic is not true and is a hoax. At the time that this video was reported to us, it had more than 300,000 views.

The video being reported

Celoplošné testovanie: fiasko a konflikt od prvej chvíle

300385 views - 19 Oct 2020

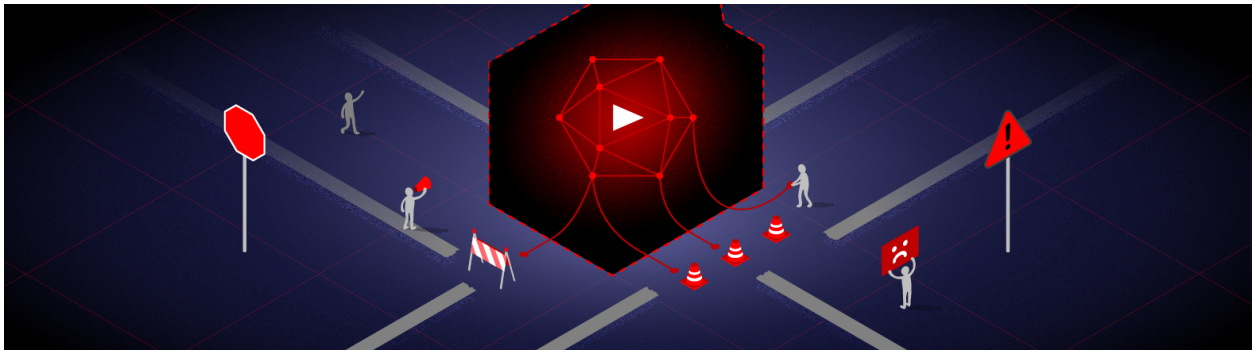
In this Greek audio recording of a podcast/lecture discussing the pandemic, the speaker suggests that the pandemic is a convenient way for the Greek government to further their priorities.

The video being reported

ΧΡΥΣΗ ΑΥΓΗ ΚΑΣΙΔΙΑΡΗΣ ΚΟΡΩΝΟΙΟΣ ΝΤΟΚΟΥΜΕΝΤΑ & ΑΛΗΘΕΙΑ στην ΕΠΙΔΗΜΙΑ ΑΠΑΤΗ φοβου των ΚΑΝΑΛΙΩΝ 2η

1916 views - 5 Απρ 2020

There are many possible explanations for this finding, including cultural differences around reporting or internet usage. However, our data suggests that YouTube has a bigger problem with regrettable content in non-English markets.



Recommendations

Our research suggests that the corporate policies and practices of YouTube, including the design and operation of their recommendation algorithms, is at least partially responsible for the regrettable experiences that our volunteers had on the platform. We believe our research has revealed is only the tip of the iceberg, and that each of these findings deserves and requires further scrutiny. We also recognise that without intervention to enable greater scrutiny of YouTube's algorithms, these problems will continue to go unchecked and the consequences on our communities will build. Despite the progress that YouTube claims to have made on these issues, it is still nearly

impossible for researchers to verify these claims, nor study YouTube's recommendation algorithms.

Our recommendations to YouTube and other platforms

1. **Enable independent audits of recommendation systems.**

   For true accountability it is essential that oversight entities and public interest researchers can 'look under the hood' of platforms like YouTube to better understand and assess how the design and operational practices associated with their recommendation systems may be contributing to online harms.

   To that end, YouTube and other platforms should make a concerted effort to facilitate researchers' efforts to study the risks and harms associated with their recommendation systems. Crucially, these data access regimes must be robust and far beyond what is offered by platforms like YouTube today, because as our report has highlighted, there are simply too many gaps.

   Rather than piecemeal and superficial insight, researchers—on condition of adhering to data protection and security protocols—should be empowered to conduct tests on recommendation systems, gain access to information about the metrics optimized by recommendation systems, access the source code of the recommendation system, access training data used to train machine learning models and review documentation related to trust and safety practices.

   This kind of transparency will help ensure that we can identify the hidden harms in the recommendations ecosystem, and develop the commercial and regulatory solutions to address them.

2. **Publish information about how recommendation systems work as well as transparency reports that give sufficient insight into problem areas and progress over time.**

   Recommendation systems are the key content delivery mechanism for platforms like YouTube, and a crucial site of trust and safety interventions. In recognition of the importance of recommender systems in amplifying harmful content, YouTube's transparency reports should provide meaningful information about the interplay between content moderation measures and recommender systems. This should include granular data that can help researchers understand how recommendation systems may contribute to the spread of harmful content and

independently verify whether the steps that YouTube are taking to reduce recommendations of harmful content are actually working as they say.

YouTube today provides no transparency into how it defines and treats borderline content. YouTube needs to step up and address this transparency gap. YouTube should expand its transparency reporting to include information about how the platform specifically defines 'borderline content', the content moderation methodologies it applies to such content (e.g. downranking; deprioritizing), and aggregate data that can help assess the issues associated with this category of content on the services (e.g. how many times YouTube's recommends borderline content and the overall amount of such content on the platform).

Importantly, this information should be accessible on a designated information site—not simply buried in the Terms of Service—comprehensive, and communicated in an intelligible manner. And to monitor for regional disparities, this information should be broken down by country/geography as well as language.

3. **Give people more control over how their data is used as an input to deliver recommendations, and the output of those recommendations.**

   Platforms should give people more control over which of their data is used to generate recommendations. People should also have full insight into and control over other information that is considered in making recommendations. For instance, people should have the option to exclude data collected from other related products/services (e.g., Google data used to inform YouTube recommendations), from previous engagement with certain content/pages/users, or data about other people that is used to generate recommendations.

   Platforms should allow people to customize the recommendations or content displayed to them in order to better protect their safety on the platform by providing additional controls or choices. This should include, for example, the ability to exclude certain keywords, types of content, or channels from recommendations. It should also include the ability to control whether and to what extent 'borderline content' or specific categories of content on the platform (e.g. news, sports) appear in recommendations.

   These controls should be accessible through centralized user settings as well as incorporated into the interface that displays algorithmic recommendations. Where possible, platforms should use plain language that describes the outcome that will result from using a control rather than the signal that the control will

send (e.g., "Block future recommendations from this channel" instead of "I don't like this recommendation").

4. **Implement rigorous, recurring risk management programs that are dedicated to recommender systems.**

   Platforms should systematically identify, evaluate, and manage on a continuous basis the risks to individuals and the public interest that may arise from the designing, functioning, or use made of the recommender system. Such a risk assessment should take account of both of the probability of a harm occurring as well as the potential magnitude of harm. By following this more comprehensive approach to risk management, platforms would be able to better take stock not only of business and reputational risks but also of the intended and unintended externalities caused by their services.

5. **Allow people to opt out of personalization.**

   Platforms, including YouTube, should provide people with an option to opt-out of personalized recommendations in favor of receiving chronological, contextual, or purely search term-based recommendations. Access to the service should not be conditional on seeing recommendations (or any other personalization choices made by people).

Our recommendations to policymakers

1. **Require YouTube and other platforms to release information and create tools that enable researchers to scrutinize their recommendation algorithms through audits and data access provisions.**

   Policymakers should introduce regulations requiring platforms to provide transparency into their recommendation algorithms, as well as robust data access frameworks that enable independent research into social media platforms. This is already being proposed in the European Commission's proposal for a Digital Services Act (DSA) and policymakers in a number of other jurisdictions have signalled a desire to advance similar transparency and oversight regimes. Policymakers must recognize that YouTube and other platforms are not stepping up to provide this much-needed transparency voluntarily, and that regulatory interventions are necessary.

2. **Ensure policy frameworks are attuned to the issues and risks unique to content recommender systems.**

If and when crafting frameworks for online content responsibility, policymakers should ensure that those regulations incentivize trust and safety responsibility for content recommender systems. There are numerous different approaches for how this might be achieved (e.g. through obligations to undertake risk assessments or offer greater end-user controls), but in principle, regulations that concern online content responsibility should ensure that platforms take due account of the risks when designing and operating automated systems that amplify content at scale.

3. **Protect researchers, journalists, and other watchdogs who use alternative methods to investigate platforms.**

   Policymakers should create safe harbor provisions or other protections that shield researchers who are conducting public-interest research independent of platform-provided data access channels from legal threats. This could prevent platforms from becoming bottlenecks in granting access to data. To this end, existing legislation should be amended or clarified. For instance, significant uncertainty remains as to the scope of research exceptions provided for in the EU's General Data Protection Regulation (GDPR); and despite the U.S. Supreme Court's encouraging recent van Buren ruling, some open questions remain on how the U.S. Computer Fraud and Abuse Act (CFAA) applies to public interest research online. To resolve such uncertainty, safe harbor provisions could bar platforms from blocking tools used for research or imposing rate limits. Provisions could provide better protections for researchers scraping data from platforms, or conducting sock puppet audits, when this would currently be in violation of platforms' Terms of Service.

Our recommendations to people who use YouTube

1. **Get informed about how YouTube recommendations work.**

   Check out Mozilla's video series "Mozilla Explains: Recommendation Engines" and "Mozilla Explains: Recommendation Engines part 2" for more in-depth information about how YouTube and other platforms that deliver algorithmic recommendations work.

2. **Check your data settings on YouTube and Google to make sure you have the right controls in place for yourself and your family.**

   Many people aren't aware that these settings exist, since they can be hard to find from the YouTube homepage. Our pro tip is to make sure you check your "watch"

and “search” history to edit out any videos that you don’t want influencing your recommendations, or consider turning this off altogether. This can be particularly helpful if you share a computer with other people. You can also visit this page to learn about how to turn off Autoplay on YouTube from your mobile, TV, or computer.

3. **Download RegretsReporter to contribute your data to our crowdsourced research!**

   Mozilla will continue to operate RegretsReporter as an independent tool for scrutinizing YouTube’s recommendation algorithm. We plan to update the extension to give people an easier way to access YouTube’s user controls to block unwanted recommendations, and will continue to publish our analysis and future recommendations.

Conclusion

YouTube’s algorithm plays a huge role in shaping what we believe both as individuals and as a society. Research by Mozilla and countless other experts has confirmed that there are significant harms associated with YouTube’s algorithms. If these experiences are as common as our research suggests and replicated across other AI systems (which is realistic given commercial offerings like Google’s Recommendations AI), the consequences will be significant.

There are many experts who argue that these problems are not actually errors with the algorithm—rather, they are the output of YouTube’s algorithm working exactly how it should and that there is a fundamental misalignment between algorithms optimized to further business incentives and those optimized for the well-being of people. That may well be true. What is definitely true is that algorithms that are this consequential should not be deployed without proper oversight. And transparency is an essential first step.

“Borderline” content is inherently subjective and difficult to regulate, whether in code or in public policy. But it is because it is difficult that it is so important to study, debate, and attempt to understand. We were only able to do this because we had the skills and resources to create a tool that would enable us to do so—and we shouldn’t have to do that. YouTube could embrace openness and enable research like ours, but they have chosen not to. They have instead built a house of cards by trying to manage these consequential decisions entirely on their own. Communities around the world are

impacted by the decisions that YouTube makes in their Silicon Valley boardrooms, and those communities deserve insight and input into these decisions.

YouTube has made it clear that they are taking the wrong approach to managing this responsibility. We will only get closer to the right approach with greater openness, transparency, accountability, and humility.

Methodology

Research Questions

Our research began with the following questions:

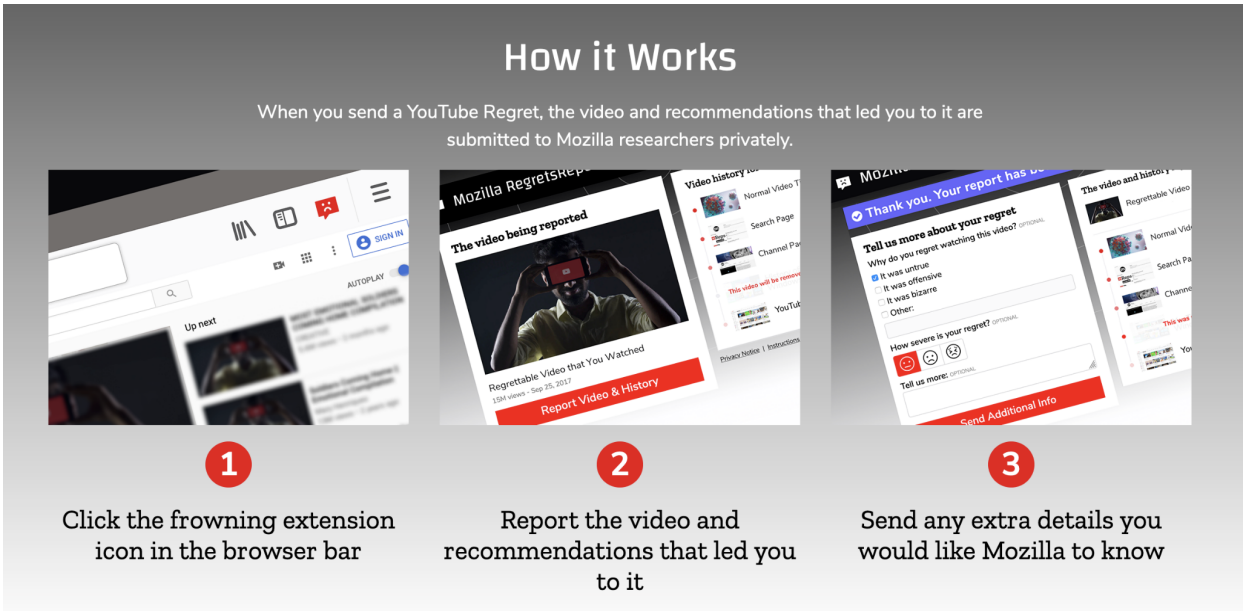
1. **What videos are being reported as regrettable?**
2. **What categories does reported regrettable content fall under?**
3. Are there identifiable patterns in terms of frequency or severity of reported regrettable content?
4. Are there specific YouTube usage patterns that lead to encountering and reporting regrettable content?
5. **Are there geographical variations in Regret rates or types?**
6. **What is the frequency of reported regrettable experiences and how does this vary by entry point?**
7. Does the report frequency change for people after they send their first report and how does this vary by category or severity?
8. **What are the qualitative characteristics of the YouTube Regrets and the recent watch history preceding the reports? Are there identifiable patterns?**

Ultimately, our dataset allowed us to answer questions 1, 2, 5, 6 and 8 (marked in bold).

RegretsReporter Extension

RegretsReporter is a browser extension, available for Firefox and Chrome, that enables volunteers to send data to Mozilla about regrettable videos that they watch on YouTube. With RegretsReporter installed, volunteers can click the RegretsReporter icon

in their browser bar on a YouTube video page to generate a report form. The report form prompts them to answer a series of questions about their experience and gives them a summary of the data that will be sent to Mozilla.



When a report is sent, Mozilla receives information about the reported video, including the title, description, view count, entry point (recommendation, search, etc.) and a link to the video watch page, as well as the optional volunteer-provided data such as category, comment, and severity rating. This allows us to understand the experience being reported and see what types of videos are regretted.

Mozilla may also receive a “trail” of how the volunteer arrived at the reported video, if the volunteer chooses to send this information. The trail includes, at maximum, the last five pages loaded on the YouTube site, within at maximum the last five hours before making the report. This information allows us to analyze the chain of recommendations or other actions that led the volunteer to the reported video. In order to calculate the rates or frequency of Regrets, RegretsReporter also receives data about each volunteer’s YouTube use, recording no detail of which videos are watched or exactly when YouTube is used, but instead an overall measure of how many videos are watched. This usage data is essential to calculate Regret rates broken down by various factors. Finally, any data that is sent to Mozilla includes the country that each volunteer is visiting YouTube from based on GeoIP lookup, allowing us to analyze how Regret rates vary between countries.

RegretsReporter is open source, which means that anyone can access the code. RegretsReporter was built using Mozilla’s lean data principles, which means that it

collects **only** data that is necessary to achieve our research aims. RegretsReporter does not collect or store any personal data about the volunteers who use it.

People-powered dataset

Since YouTube does not make data available to researchers, we developed a citizen science approach to collecting data needed to answer our research questions. Building on and inspired by past investigations, RegretsReporter is the largest-ever crowdsourced investigation into YouTube's recommendation system. Our dataset is powered by 37,380 volunteers across 190 countries who installed the RegretsReporter browser extensions for Firefox and Chrome. Of our total volunteer contributors, 1,662 submitted at least one report, for a total of 3,362 reports coming from 91 countries, submitted between July 2020 and May 2021. Volunteers who downloaded the extension but did not file a report were still an important part of our study. Their data—for example, how often they use YouTube—was essential to our understanding of how frequent regrettable experiences are on YouTube[3] and how this varies between countries.

This people-powered approach captures the real lived experience of people who use YouTube and, critically, allows us some insight into the algorithm despite YouTube's unwillingness to provide data to researchers. However, there are methodological limitations to our approach, including:

- Selection bias: Our volunteers are a particular group of people and our findings may not generalize across all YouTube users.
- Reporting bias: There may be many factors that affect whether a volunteer reports a particular video.
- Regret concept: The concept of a YouTube Regret is (intentionally) non-specific, and different volunteers may make reports based on different notions of "regret" than others.
- The observational nature of the study means that, while we can confidently state "what" is happening, we are not able to confidently infer the "why". For example, we do not know why YouTube chose to recommend any particular video to any particular volunteer.

[3] Calculating frequency of Regrets depends both on the number of Regrets experienced and on the total number of videos watched - 1 Regret out of 5 videos watched is very different from 1 Regret out of 100 videos watched. Thus, the number of videos watched by our volunteers is critical, even for volunteers that made no reports.

Despite these limitations, our results offer insight into problems on YouTube from the lived perspective of real people from all around the world. We believe that the limitations of our research underscore the need for YouTube to provide data to researchers.

Analysis methods

The RegretsReporter extension transmits collected data through Mozilla's telemetry system and then stores it in Mozilla's data warehouse. Analysis was performed using BigQuery and Python, primarily in the Google Colab environment. The analysis code used is publicly available here.

We used several different methods in our analysis, employing desk research, statistical methods, and qualitative analysis to address our research questions.

Statistical methods entailed calculation of counts, proportions, and rates, while slicing by relevant variables. Inferential techniques were applied including calculation of confidence intervals and evaluation of hypothesis tests. All differences described in this report were statistically significant at the $p < 0.05$ level, and all statistical significance was evaluated at this level.

We began the qualitative analysis process by convening a working group of experts (named individually in the Acknowledgements section of this report) with backgrounds in online harms, freedom of expression, and tech policy. The working group was tasked with watching reported videos from the dataset, and then identifying themes. Over the course of three months, the working group developed a conceptual framework for classifying some of the videos, based on YouTube's Community Guidelines. The working group decided to use YouTube's Community Guidelines to guide the qualitative analysis because it provides a useful taxonomy of problematic video content and also represents a commitment from YouTube as to what sort of content should be included on their platform.

A team of 41 research assistants (named individually in the Acknowledgements section of this report) employed by the University of Exeter then used this conceptual framework to assess each reported video. The framework asked the research assistants to answer questions such as:

- Do you think this video should be on YouTube?
- Do you think this video should be recommended on Watch Next and on the YouTube Homepage?

- If you've answered NO to any of the questions above: Which Community Guideline category do you believe the video may violate?

The research assistants also analyzed the qualitative characteristics of the recommendation trails that led to the regretted video (when available) and answered additional questions such as the primary language of the video and whether it was about the Covid-19 pandemic, which helped other elements of our analysis.

From 7-15 June, the research assistants analyzed 1141 or 33.9% of the total 3362 reports.

Report data was collected starting on July 22 of 2020, first with a select beta test group, followed by general availability in September of 2020. We cut off the data for analysis on May 31, 2021. There were two data cleaning procedures applied:

Inactive criteria. Some volunteers continued to leave the extension installed, but made no use of the reporting functionality. Given that the intent for installing the extensions was to make reports, we consider these volunteers as inactive if they have made no reports for 56 days (two 28-day months) after their last report, or since installation in the case they have made no reports. The volunteer's usage data is not counted after this 56-day period ends. If the volunteer goes inactive and then makes a report, they are reactivated and their full usage data is counted.

We feel this cutoff period is appropriate as 91% of volunteers that made a report did so within the first 56 days after installing RegretsReporter

Outlier criteria. Two volunteers appear to have used the extension in a manner that is not consistent with our general population of volunteers. These two volunteers filed 231 and 109 reports respectively, while the next most prolific volunteers filed 65, 53, and 37. We felt that inclusion of the data from these two volunteers would detract from the generalizability of our results, so excluded them from analysis.

Disclosures

Mozilla is firmly committed to being carbon-neutral and will significantly reduce our greenhouse gas (GHG) footprint year over year aligning to, and aiming to exceed, the net zero emissions commitment of the Paris Climate Agreement. We use Google Cloud Platform (GCP) to host RegretsReporter and process queries. GCP is 100% carbon-neutral, and RegretsReporter primarily uses GCP infrastructure based in Oregon, which uses 89% carbon-free energy. The full scope of carbon emissions

associated with RegretsReporter will be calculated and disclosed in Mozilla's next Greenhouse Gas (GHG) Inventory, due to be released in 2022.

All data collected and processed for this project was handled by either Mozilla employees or by researchers contractually bound by Mozilla's privacy, security, data processing, and confidentiality policies.

We acknowledge the breadth of research into the psychological impacts of content moderation and related work. The team who worked on this research, including research assistants from the University of Exeter who watched the bulk of the videos reported to support our analysis, were provided with psychological support during the course of this research.

Google is the default search engine in Firefox in many regions of the world. Despite this relationship, Mozilla has a long history of advocating for greater transparency and accountability from YouTube and other online platforms.

References

Adamczyk, Roman. "What's the Hold-up? How YouTube's inaction allowed the spread of a major French COVID-19 conspiracy documentary." *EU DisinfoLab,* 12 May 2021, https://www.disinfo.eu/publications/whats-the-hold-up%3F-how-youtubes-inaction-allowed-the-spread-of-a-major-french-covid-19-conspiracy-documentary/

Alexander, Julia. "YouTube claims its crackdown on borderline content is actually working." *The Verge*, 3 Dec. 2019, https://www.theverge.com/2019/12/3/20992018/youtube-borderline-content-recommendation-algorithm-news-authoritative-sources

Bergen, Mark. "YouTube Executives Ignored Warnings, Letting Toxic Videos Run Rampant." *Bloomberg*, 2 Apr. 2019, https://www.bloomberg.com/news/features/2019-04-02/youtube-executives-ignored-warnings-letting-toxic-videos-run-rampant

Boyd, Ashley. "Senate Hearing Confirms YouTube Won't Fully Release Recommendations Data Without More Pressure from Public and Congress." *Mozilla*, 28 Apr. 2021, https://foundation.mozilla.org/en/blog/senate-hearing-confirms-youtube-wont-fully-release-recommendations-data-without-more-pressure-from-public-and-congress/

Bridle, James. “Something is wrong on the internet.” *Medium*, 6 Nov. 2017, https://medium.com/@jamesbridle/something-is-wrong-on-the-internet-c39c471271d2

Campbell, Eliza and Spandana Singh. “The flaws in the content moderation system: The Middle East case study.” *Middle East Institute*, 17 Nov. 2020, https://www.mei.edu/publications/flaws-content-moderation-system-middle-east-case-study

Chen, Annie Y., Brendan Nyhan, Jason Reifler, Ronald E. Robertson, and Christo Wilson. “Exposure to Alternative & Extremist Content on YouTube” *ADL*, Feb. 2021, https://www.adl.org/resources/reports/exposure-to-alternative-extremist-content-on-youtube

Cobbe, Jennifer and Jatinder Singh. “Regulating Recommending: Motivations, Considerations, and Principles.” *European Journal of Law and Technology (EJLT) 10(3)*, 30 Dec. 2019, https://ejlt.org/index.php/ejlt/article/view/686

Cook, Jesselyn and Sebastian Murdock. “YouTube Is A Pedophile’s Paradise.” *HuffPost*, 20 Mar. 2020, https://www.huffpost.com/entry/youtube-pedophile-paradise_n_5e5d79d1c5b6732f50e6b4db

Córdova, Yasodara, Adrian Rauchfleisch and Jonas Kaiser. “The implications of venturing down the rabbit hole.” *Internet Policy Review*, 27 Jun. 2019, https://policyreview.info/articles/news/implications-venturing-down-rabbit-hole/1406

Fisher, Max and Amanda Taub. “How YouTube Radicalized Brazil.” *The New York Times*, 11 Aug. 2019, https://www.nytimes.com/2019/08/11/world/americas/youtube-brazil.html

Fisher, Max and Amanda Taub. “On YouTube’s Digital Playground, an Open Gate for Pedophiles.” *The New York Times*, 3 Jun. 2019, https://www.nytimes.com/2019/06/03/world/americas/youtube-pedophiles.html

Geurkink, Brandi. “Congratulations, YouTube… Now Show Your Work.” *Mozilla*, 5 Dec. 2019, https://foundation.mozilla.org/en/blog/congratulations-youtube-now-show-your-work/

Geurkink, Brandi. “Our recommendation to YouTube.” *Mozilla*, 14 Oct. 2019, https://foundation.mozilla.org/en/blog/our-recommendation-youtube/

Leerssen, Paddy. “The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems.” *European Journal of Law and Technology (EJLT) 11(2)*, 31 Oct. 2020, http://www.ejlt.org/index.php/ejlt/article/view/786

Lewis, Rebecca. “Alternative Influence: Broadcasting the Reactionary Right on YouTube.” *Data & Society*, Sep. 2018, https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf.

Maréchal, Nathalie and Ellery Roberts Biddle. “It's Not Just the Content, It's the Business Model: Democracy’s Online Speech Challenge.” *Ranking Digital Rights*, 17 Mar. 2021, https://rankingdigitalrights.org/its-the-business-model/

Molter, Vanessa. “Platforms of Babel: Inconsistent misinformation support in non-English languages.” *Election Integrity Partnership*, 21 Oct. 2020, https://www.eipartnership.net/policy-analysis/inconsistent-efforts-against-us-election-misinformation-in-non-english

Nicas, Jack. “How YouTube Drives People to the Internet’s Darkest Corners.” *The Wall Street Journal*, 7 Feb. 2018, https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478

Ohlheiser, Abby. “They turn to Facebook and YouTube to find a cure for cancer — and get sucked into a world of bogus medicine.” *The Washington Post*, 25 Jun. 2019, https://www.washingtonpost.com/lifestyle/style/they-turn-to-facebook-and-youtube-to-find-a-cure-for-cancer--and-get-sucked-into-a-world-of-bogus-medicine

Ribeiro, Manoel Horta, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira. “Auditing radicalization pathways on YouTube.” *In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)* (pp. 131–141), 27 Jan. 2020., https://doi.org/10.1145/3351095.3372879

Roberts, Sarah T. “Behind the Screen: Content Moderation in the Shadows of Social Media.” *Yale University Press*, 2019, https://doi.org/10.2307/j.ctvhrcz0v

Roth, Camille, Antoine Mazières and Telmo Menezes. “Tubes and bubbles topological confinement of YouTube recommendations.” *PLoS ONE 15(4)*, 21 Apr. 2020, https://doi.org/10.1371/journal.pone.0231703

Sanna, Leonardo, Salvatore Romano, Giulia Corona and Claudio Agosti. “YTTREX: Crowdsourced Analysis of YouTube’s Recommender System During COVID-19 Pandemic.” *Information Management and Big Data (pp.107-121)*, May 2021, https://doi.org/10.1007/978-3-030-76228-5_8

Singh, Spandana. “Why Am I Seeing This? How Video and E-Commerce Platforms Use Recommendation Systems to Shape User Experiences.” *Open Technology Institute*, 25 Mar. 2020, https://www.newamerica.org/oti/reports/why-am-i-seeing-this/

Singh, Spandana. “Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content.” *Open Technology Institute*, Jul. 2019, https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/

Tufekci, Zeynep. “YouTube, the Great Radicalizer.” *The New York Times*, 10 Mar. 2018, https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

Vermeulen, Mathias. “The keys to the kingdom. Overcoming GDPR-concerns to unlock access to platform data for independent researchers”. *Knight First Amendment Institute Draft paper*, 27 Nov. 2020, https://doi.org/10.31219/osf.io/vnswz

“Answers to Your Questions About the Dark Side of the Internet.” *Mozilla*, 3 Sep. 2019, https://foundation.mozilla.org/en/blog/answers-your-questions-about-dark-side-internet/

“Continuing our work to improve recommendations on YouTube.” *YouTube Official Blog*, 25 Jan. 2019, https://blog.youtube/news-and-events/continuing-our-work-to-improve

“How YouTube Works.” *YouTube*, 2021. https://www.youtube.com/intl/howyoutubeworks/

“2020 Ranking Digital Rights Corporate Accountability Index.” *Ranking Digital Rights*, 2020, https://rankingdigitalrights.org/index2020/explore-indicators

“The Four Rs of Responsibility, Part 2: Raising authoritative content and reducing borderline content and harmful misinformation.” *YouTube Official Blog*, 3 Dec. 2019, https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce

“What Happened After My 13-Year-Old Son Joined the Alt-Right.” *Washingtonian*, 5 May 2019, https://www.washingtonian.com/2019/05/05/what-happened-after-my-13-year-old-son-joined-the-alt-right

Acknowledgements

This report was written by Jesse McCrosky and Brandi Geurkink.

Contributing authors included: Kevin Zawacki, Anna Jay, Carys Afoko, Maximilian Gahntz and Owen Bennett.

We want to thank the members of our working group for sharing their expertise with us to improve our analysis framework: Gabrielle Guillemin, Dia Kayyali, Chico Camargo, Udbhav Tiwari, Jason Chuang, and Amber Sinha. The views in this paper do not necessarily reflect theirs, nor their employers. We also want to thank the research assistants at the University of Exeter, supervised by Dr. Chico Camargo, for their diligence and hard work analyzing the reported videos: Julia Dominika Burzyk, Maria Campbell, Giulia Catelani, Hoi Ying Chang, Sharon Choi, Hannah Cox, Isabel Dally, Ben Entwisle, Alice Gallagher Boyden, Laura Garratt, Adriano Giunta, Lisa Greghi, Rosemary Griggs, Matthew Gurney, Connie Hitchin, Olliver Hopkins, Oana Ionescu, Elliot Jones, Ritvika Kedia, Ruslan Kudryashov, Smriti Lakhotia, Michael Lewis, William Lewis, Mitran Malarvannan, Lois Mander, Zachary Marre, Alina McGregor, Inês Mendes de Souza, Ayodele Ogunyemi, Henry Payne, Sadaf Sahel, Matej Svoboda, Jia Tang Zhi, Martina Toneva, Olivia Warnes, Katherine Williams, Ching Yin Yan, Suchitra Bansode, Mingting Hong, Ayush Shrivastav, Feng Xu.

We are grateful to our colleagues across civil society who provided input, guidance and feedback on this report including Nathalie Maréchal, Spandana Singh and Mathias Vermeulen. The views in this paper do not necessarily reflect theirs, nor their employers.

Thank you to Fred Wollsén who built RegretsReporter, and to the team at Reset Tech for funding this work.

Finally, we want to thank the 37,380 volunteers who contributed data through RegretsReporter. Our lean data policy means that we don't know who they are (and so we cannot thank them individually), but without their contributions this research would not have been possible.

Annex: Examples of YouTube Regrets by category

[Download Addendum Report PDF]