revista Fronteiras – estudos midiáticos
23(2):118-139 maio/agosto 2021
Unisinos – doi: 10.4013/fem.2021.232.09

# Deplatforming, demotion and folk theories of Big Tech persecution

## Desplataformização, rebaixamento e teorias folk em relação à perseguição das Big Tech

Emillie de Keulenaar[*]
edekeulenaar@gmail.com

Anthony Glyn Burton[**]
ab@anthbrtn.com

Ivan Kisjes[***]
i.kisjes@uva.nl

**ABSTRACT**
This article examines YouTube's moderation of conspiracy narratives surrounding COVID-19 through an analysis of deplatformed and demoted YouTube videos. Building upon the literature on moderation, it compares the types of content moderated by YouTube from April to October, 2020. In doing so, it seeks to determine the extent to which YouTube's own moderation actions are brought in as part of the conspiratorial narratives surrounding COVID-19, while investigating how it is that moderation becomes entangled with questions of truth and visibility.

**Keywords:** content moderation; conspiracy theories; YouTube.

**RESUMO**
Este artigo examina a moderação de narrativas de conspiração sobre COVID-19 por meio de métodos digitais. Com base na literatura sobre moderação, ele realiza uma comparação dos tipos de conteúdo moderados pelo YouTube de Abril à Outubro de 2020. O artigo busca determinar até que ponto as ações de moderação são trazidas como parte das narrativas conspiratórias em torno da pandemia, enquanto investiga como sua moderação se envolve com questões de verdade e visibilidade.

**Palavras-chave:** moderação de conteúdos, teorias de conspiração, YouTube.

[*] Simon Fraser University (8888 University Dr, Burnaby, BC V5A 1S6, Canadá) e Universidade de Amsterdam (1012 WX Amsterdam, Países Baixos).
[**] Simon Fraser University. 8888 University Dr, Burnaby, BC V5A 1S6, Canadá.
[***] Universidade de Amsterdam. 1012 WX Amsterdam, Países Baixos.

## Introduction

There are many truths stranger than fiction, but none so strange as the emergence of the COVID-19 pandemic at the nadir of a half-decade of expanding fault-lines in the legitimation of public narratives. Compounding this is the fact that even among public health authorities, information about the origin, treatment, and prevention of COVID-19 has not always been certain — from whether the virus has leaked from a lab in Wuhan, to whether asymptomatic people can contaminate others, or if children can be contagious (Iati *et al.*, 2020; O'Leary, 2020). This has driven an imperative for mainstream social media platforms to foster consensus among its user bases, by for example raising up "authoritative sources" on top of search ranking and recommendation results on YouTube and Google Search (Skopeliti and John, 2020), or setting up centralised reference points to the latest local guidelines and information found on the virus on Twitter homepages (Roth and Pickels, 2020).

The problem, it seems, is that these actively moderated platforms are often subject to the misinformation they seek to shut down. In US YouTube infospheres in particular (see, e.g., Flam 2021), they have been characterised by users as "big tech" infrastructures that control the flow of information about sensitive truths, motivated to subvert individuals' self-determination in favour of authoritarian entities (Connolly, 2020). To intervene directly in the "meanings and meaningfulness" of user-generated contents (Langlois, 2014) renders moderation an "essentially contested" governance measure (de Laat, 2012): moderation policies are perceived as drawing external, and thus arbitrary, lines between what users can and cannot claim to be true. From here, a core tension between human-based interventions and platform infrastructures emerges: moderation is a case-by-case answer to the degree to which platforms, within their own infrastructure, simultaneously incentivize and adjudicate the production of misinformation (Burton and Koehorst, 2020).

How, then, is the relationship between platforms and moderated users affected by the former's adjudications as to what is authoritative information on COVID-19? Answering this question requires that we first explore how COVID-19 misinformation is moderated over time, namely how content moderation policies qualify contents related to COVID-19 as authoritative, misleading or false, and how content moderation measures execute these policies in the form of sorting, ranking or deletion techniques. We chose to examine the moderation of eight currently unconfirmed allegations and conspiracy theories on YouTube, in English, between April and November of 2020. We first close-read YouTube's anti-misinformation content moderation policies, namely *Spam, deceptive practices and scams* (YouTube, 2021) and *COVID-19 Medical misinformation* (YouTube, 2020). Repurposing the YouTube video downloader youtube-dl (sic), we built a dataset of 108,537 videos and 31,531,963 comments by collecting the first 60 search results of 98 queries that reflect COVID-19 conspiracy vernaculars (Qanon hashtag "wwg1wga" [when we go one, we go all], "id2020" [COVID-19 vaccine microchip], and "covid depopulation", among others). We then captured moderation information, such as video statuses and daily search rankings. We chose to study YouTube because metadata tied to its content moderation, such as the rankings of search results or information on the availability of videos, remain more accessible than in other platforms thanks to scraping tools like youtube-dl (Garcia Gonzalez et al. 2021). This also applies to access to user-generated contents, such as video metadata, transcripts and user comments.

Secondarily, we assessed how users qualify YouTube as a source of information in relation to the platform's moderation efforts. We used a combination of natural language processing techniques, namely word trees (Wattenberg and Viegas, 2008) of comment sentences containing words relating to demotion, deplatforming and other speech restrictions ("shadowbanned", banned, censored...) and subject-verb-object networks (Milajevs, Sadrzadeh and Roelleke, 2015) of comment sentences claiming what YouTube is and does in relation to moderation.

Our study contributes to two existing areas of research: content moderation and user studies in relation to moderation. We find that, in burying down conspiratorial contents and prioritising "authoritative" sources like mainstream media and public health experts, YouTube places a degree of attention on allegations of censorship contained in such theories. When user discussions and substantiations of conspiracy theories are pushed off the platform, this provides conspiratorial thinkers (1) proof that YouTube is a partisan platform whose moderation policies apply the repressive politics of COVID-wary politicians and (2) evidence they use to prove said theories, such as footage of politicians, pundits and celebrities discussing the implications of the virus. In this sense, we conclude that moderation separates the production and the archiving of misinformation: while YouTube's moderation policies make it an inhospitable place for deliberating conspiracies, it remains a platform *by* and *for* the conspirators.

## Deplatforming, demotion and factual contingencies

Moderation has never been an oddity to social media platforms: Tarleton Gillespie goes so far as to define platforms as existing *for* their moderation (Gillespie, 2018). Still, recent commentaries express a certain surprise for an apparent shift of ethos in platform discourse (Gillespie, 2013): active moderation marks a clear rupture from early Facebook, Twitter, and Google's self-promotion as open, free or participatory alternatives to mass media. This perceived change has been the target of legal scholars concerned with platforms's monopoly over the regularisation of public speech, which some suspect parallels censorship or bias (Jiang, Robertson and Wilson, 2019; Lee, 2020). In response to this, many studies today argue in favour of a "democratisation" of platform moderation with collaborative or participative techniques (De Gregorio, 2020), delegating them to civil society (Elkin-Koren and Perel, 2020), or providing sufficient context for decisions to sanction, quarantine, or delete user-generated contents (Myers West, 2018; Wilson and Land, 2020).

In this context, empirical studies have examined the contingent ways in which platforms intervene within public deliberations of what constitutes right, wrong, true and false information by sorting, ranking and deleting user-generated contents that step outside their jurisdictional boundaries (Rieder, 2017). One of the most prominent techniques so far has been the outright suspension, deletion or "deplatforming" of users or contents (c.f. Rogers 2020) and other research on the "replatforming" of deleted contents in alternative platform ecologies (OILab, 2019). Rogers gives reason to believe that Twitter suspensions have constituted an effective strategy for reducing hate speech and related contents (Rogers, 2020, pp. 13–15), based on the fact that early provocateurs such as Milo Yiannopolous and Alex Jones have mellowed their language while their audiences have thinned on the platforms they migrated to. Similar claims have been made about Reddit's "crowdsourced" moderation (Chandrasekharan *et al.*, 2021) and YouTube's deletion of high-profile conspiracy contents, such as *Plandemic*, a documentary that claims COVID-19 was a planned hoax (Frenkel, Decker and Alba, 2020).

Contrary to hate speech, however, COVID-19 has brought specific challenges to deplatforming (Douek, 2021; de Keulenaar et al., forthcoming). While information on COVID-19 bears critical health risks, the spectrum of true and false information on the virus changes as its epidemiology and public health policies has evolved. Simple speculations about the virus have emerged alongside more elaborate conspiracy theories (Knight, 2021), which, as with claims that the virus leaked from a lab in Wuhan (Maxmen & Mallapaty, 2021), have at times shifted their status from fringe statements to reasonable doubt. This makes it difficult for platforms to opt for outright deletion of "false" or "misleading" contents without being subjected to a loss of legitimacy, be it caused by accusations of epistemic bias or infringements upon public rights to consult information for individual decision-making.

Like its counterparts, then, YouTube has had to find a balance between allowing users to openly speculate about the virus and public health policies, and preventing misleading or false information from possibly causing harm. Juggling one and the other extreme implies a combination of flexible and strict moderation techniques — "hard" and "soft" moderation — including deplatforming, demotion (down-ranking) and promotion (up-ranking) of "authoritative sources" in search and recommendation results (Faddoul, 2020).

Aside from deplatforming, demotion has been studied specifically in relation to search ranking and recommendation mechanisms. While deplatforming may strictly delimit the boundaries of acceptable user behaviour, demotion works to modulate the prominence of problematic contents in the overall assemblage of the platform. Since at least 2015, YouTube has focused on tweaking its ranking algorithms to control the visibility of "authoritative" and "borderline contents", namely by down- or up-ranking each of these types of contents dynamically (YouTube, 2019a). This technique is designed to prevent potentially problematic contents from gathering too much engagement before they infringe YouTube content moderation rules. Describing Facebook's own demotion techniques, Constine (2018) notes that this measure gives a certain flexibility to content moderation by supervising contents that approach the "policy line" separating allowed from prohibited contents.

To understand the motive of demotion techniques, it is useful to look closely at what platforms mean by "borderline contents". On YouTube, "borderline contents" is a term that appeared in June of 2019, a moment when the platform was under heavy criticism for allowing the circulation of historical revisionist, scientific racist and conspiratorial contents (Lewis, 2018; Ekman 2014). At the time, such contents did not immediately infringe upon the platform's guidelines, but could arguably inform and at times incite violent behaviour, as evidenced in the Charlottesville "Unite the Right" rally of August 2017

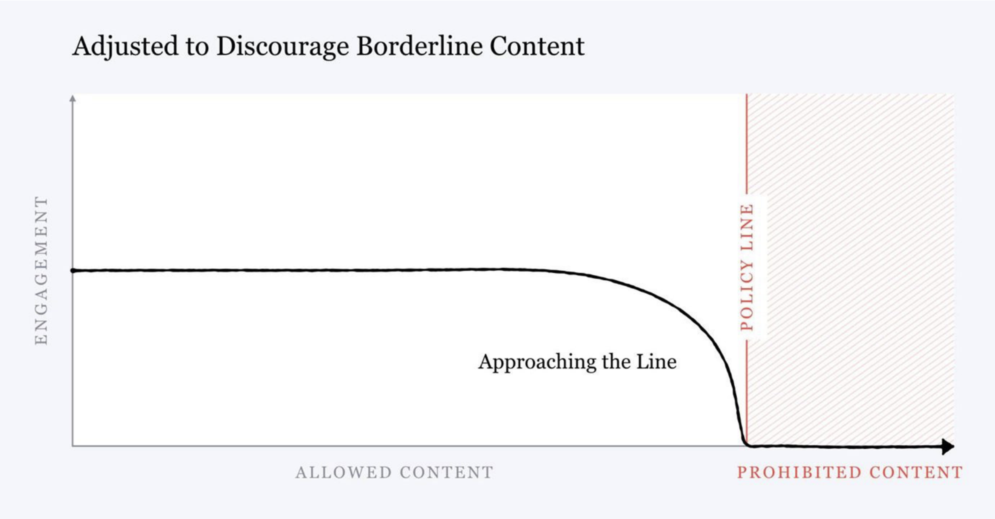


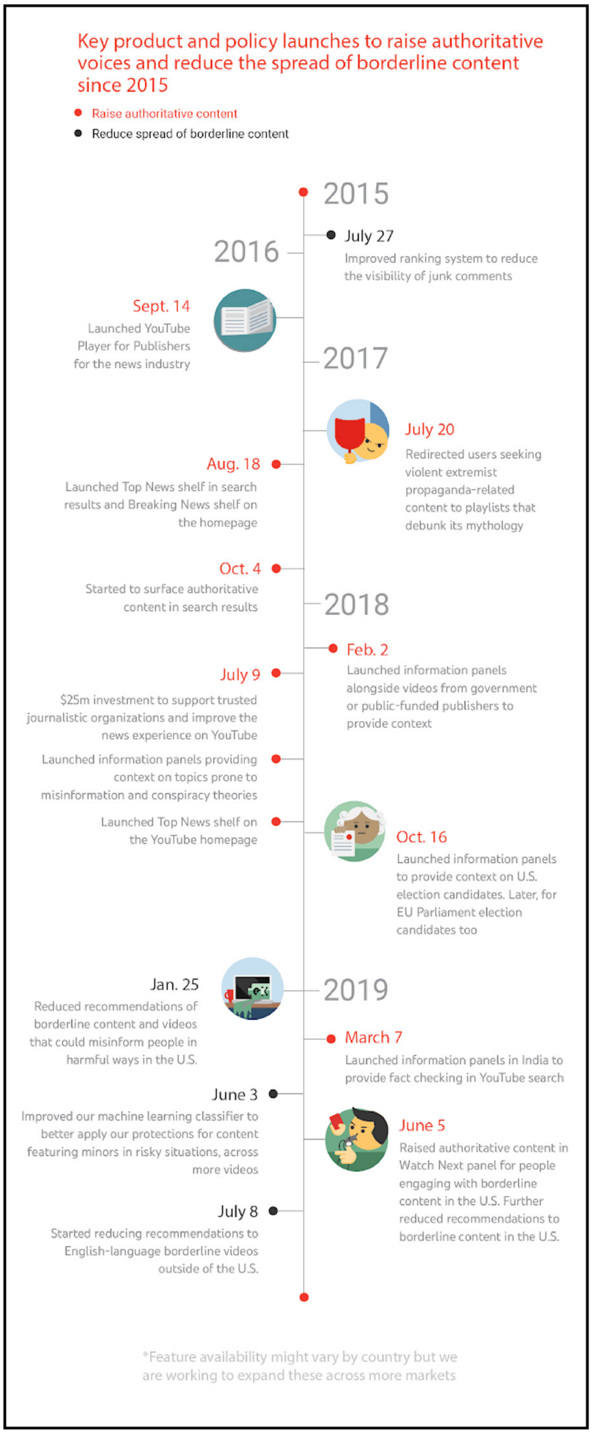
**Figure 1.** Constine's representation of the demotion technique in Facebook's Newsfeed algorithm (in Constine, J. (2018) 'Facebook will change algorithm to demote "borderline content" that almost violates policies', TechCrunch, 15 November. Available at: https://social.techcrunch.com/2018/11/15/facebook-borderline-content/ (Accessed: 20 February 2021).

(Lind 2017). YouTube's solution to closing this gap was to find a concept that could frame inchoate dangers: "borderline contents" indeed refers to what "comes close to — but doesn't quite cross the line of — violating our Community Guidelines" (YouTube, 2019a; YouTube, 2021), with examples as varied as "videos promoting a phony miracle cure for a serious illness, claiming the earth is flat, making blatantly false claims about historic events like 9/11" (The YouTube Team, 2019).

By the beginning of the pandemic, demotion techniques on YouTube had also gained an educational function. While earlier efforts to raise authoritative contents may have been designed to counter-balance misinformation, the gravity and scale of COVID-19 has made it necessary to foster public consensus for local health policies and effectively funnel users down to common sources of information. Besides burying down potentially problematic contents on search and recommendation results, then,  YouTube actively up-ranked what it called "authoritative" or "trusted sources", which are described as mainstream journalistic outlets like "CNN, Fox News, Jovem Pan, India Today and the Guardian", experts in given fields, such as "public health institutions" and local authorities (YouTube, 2021b).

This represents something of a shift for a platform that has long been perceived as running on user-generated content. By observing the top twenty unpersonalized results of four queries over a period of time, Rieder *et al.* find that YouTube's role in determining the ranking of search results is typically only partial: it combines user strategies to up-rank competitors, user engagement (views, up- and down votes, comments, subscriptions) and what the platform classifies as worthy of consumption based on relevance, recency, and user's affinity with recommended contents (Davidson *et al.*, 2010; Covington, Adams and Sargin, 2016). Any platform intervention implies a careful "mediation or curation of [user-generated] content and, consequently, of perspectives or viewpoints", including around different conceptions of "importance" and "authoritativeness" (Rieder, Matamoros-Fernández and Coromina, 2018, p. 52).

YouTube's efforts to introduce external criteria for ranking algorithms imply a problematic relationship with user feedback. Contrary to notions of "importance" or "popularity" informed by user behavior, "authoritative sources" represent an external criteria independent from user preferences. Content moderation and the political and public health implications that render it necessary may thus signal a rupture from previous platform designs as seemingly ethereal spaces by and for users (Gillespie, 2013). How, then, do more proactive content moderation policies affect COVID-19 misinformation as user-generated content?



Key product and policy launches to raise authoritative voices and reduce the spread of borderline content since 2015

● Raise authoritative content

● Reduce spread of borderline content

2015

July 27 — Improved ranking system to reduce the visibility of junk comments

2016

Sept. 14 — Launched YouTube Player for Publishers for the news industry

2017

July 20 — Redirected users seeking violent extremist propaganda-related content to playlists that debunk its mythology

Aug. 18 — Launched Top News shelf in search results and Breaking News shelf on the homepage

Oct. 4 — Started to surface authoritative content in search results

2018

Feb. 2 — Launched information panels alongside videos from government or public-funded publishers to provide context

July 9 — $25m investment to support trusted journalistic organizations and improve the news experience on YouTube

Launched information panels providing context on topics prone to misinformation and conspiracy theories

Launched Top News shelf on the YouTube homepage

Oct. 16 — Launched information panels to provide context on U.S. election candidates. Later, for EU Parliament election candidates too

2019

Jan. 25 — Reduced recommendations of borderline content and videos that could misinform people in harmful ways in the U.S.

March 7 — Launched information panels in India to provide fact checking in YouTube search

June 3 — Improved our machine learning classifier to better apply our protections for content featuring minors in risky situations, across more videos

June 5 — Raised authoritative content in Watch Next panel for people engaging with borderline content in the U.S. Further reduced recommendations to borderline content in the U.S.

July 8 — Started reducing recommendations to English-language borderline videos outside of the U.S.

*Feature availability might vary by country but we are working to expand these across more markets

**Figure 2.** YouTube's timeline of actions taken to raise authoritative contents and reduce the spread of borderline content since 2015. In YouTube (2021) The Four Rs of Responsibility, Part 2: Raising authoritative content and reducing borderline content and harmful misinformation, blog.youtube. Available at:

## Method

The question of how YouTube's moderation practices affect COVID-19 misinformation requires us to first examine how the platform moderates the latter. In order to answer this question, we built a dataset of videos and comments mentioning one or various COVID-19 conspiracy queries between April and November of 2020. We then examine how users respond to YouTube moderation, as well as how they frame YouTube within COVID-19 conspiracies. We use two natural language processing techniques to "distant read" (Moretti, 2013) comments that debate moderation and YouTube, namely word trees (Wattenberg and Viegas, 2008) and subject-verb-object networks (Milajevs, Sadrzadeh and Roelleke, 2015).

In contrast to contemporary YouTube studies (Arthurs, Drakopoulou, and Gandini 2018), we chose to avoid the usage of YouTube's API interface to collect our data, instead using the youtube-dl program alongside scripting mechanisms to build our own database. YouTube's attempt to self-regulate the spread of misinformation on its platform have occurred in lockstep with growing restrictions on mainstream social platforms for researchers to query, access, and analyze data through the API (Bruns 2019). These restrictions tend to centre around data collection that allows users to critique and investigate the operations of the platform's technical structures (Rieder, 2018). The question of this "post-API" evolution in digital methods has been well-documented by scholars over the past three years (see, for example, Freelon 2018 and Perriam et. al., 2019), with the idea that decreasing access illustrates the tenuous nature of access to digital data. The "APICalypse", in the words of Axel Bruns, introduces difficulties in studying "phenomena such as abuse, hate speech, trolling, and disinformation campaigns" alongside the role that the platforms play in the circulation and iteration of these behaviours (Bruns, 2019). The dynamism of our research design–for example, the method we developed to track the deletion of videos–would have faced considerable difficulties if we restricted our data source to the YouTube API.

## Query design

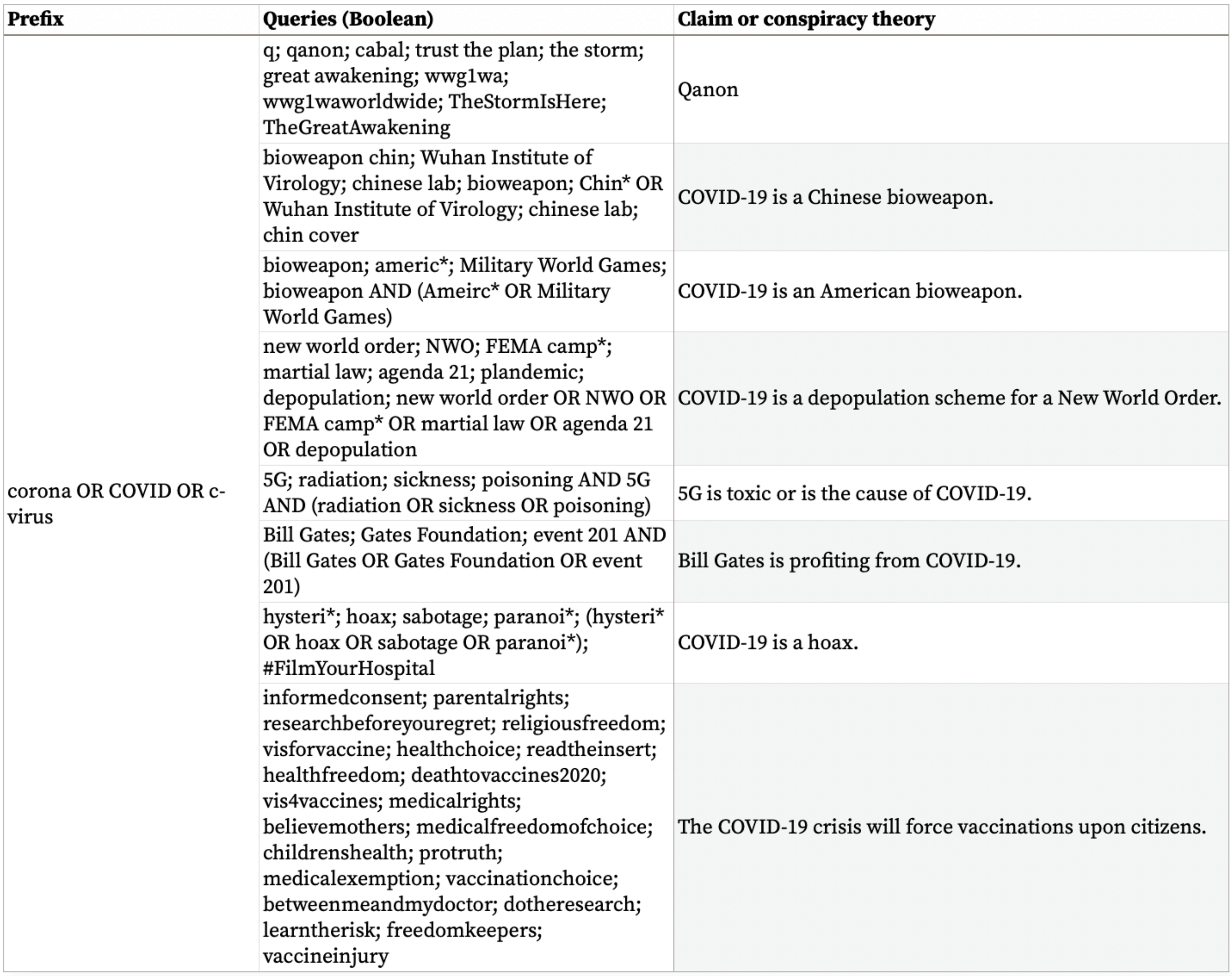
Our queries were designed based on conspiracy theories reported by news media up to late March 2020, and vernaculars proper to messaging boards known to produce such conspiracies (de Zeeuw *et al.*, 2020), particularly 4chan's /pol/ board and 8kun (formerly 8chan).

Conspiracy theories or claims included the idea that COVID-19 is a Chinese or American bioweapon; that 5G is the cause of COVID-19; that Bill Gates has known about the pandemic beforehand and is profiting from it; or that it is simply a hoax (Knight and Birchall, 2020).

These conspiracies translated into 98 queries in total, though we ultimately narrowed our analysis to four queries: "id2020", "wwg1wga", "depopulation" and "5g radiation". This is because we wanted to test search rankings for different types of moderated misinformation: while terms like "depopulation" and "wwg1wga" may be likely to be classified as "borderline content", "id2020" and "5g radiation" are listed in YouTube's *COVID-19 Medical Misinformation Policy* as contradicting the World Health Organisation or public health authorities. Though similar in substance, these two types of misinformation are described and moderated differently according to YouTube policies.

## Data collection

With these queries, we used youtube-dl (Garcia Gonzalez, Amine and M., 2021), an open-source command-line program to download videos and audio from YouTube. Youtube-dl allows one to capture metadata including channel names, channel IDs, video IDs, video comments, video transcripts, engagement (views, likes and dislikes), search rankings, and video status (e.g., "This video has been removed due to copyright."). Due to the propensity of conspiracy videos to be platform-moderated, our youtube-dl script was scheduled



| Prefix | Queries (Boolean) | Claim or conspiracy theory |
|---|---|---|
| corona OR COVID OR c-virus | q; qanon; cabal; trust the plan; the storm; great awakening; wwg1wa; wwg1waworldwide; TheStormIsHere; TheGreatAwakening | Qanon |
| | bioweapon chin; Wuhan Institute of Virology; chinese lab; bioweapon; Chin* OR Wuhan Institute of Virology; chinese lab; chin cover | COVID-19 is a Chinese bioweapon. |
| | bioweapon; americ*; Military World Games; bioweapon AND (Ameirc* OR Military World Games) | COVID-19 is an American bioweapon. |
| | new world order; NWO; FEMA camp*; martial law; agenda 21; plandemic; depopulation; new world order OR NWO OR FEMA camp* OR martial law OR agenda 21 OR depopulation | COVID-19 is a depopulation scheme for a New World Order. |
| | 5G; radiation; sickness; poisoning AND 5G AND (radiation OR sickness OR poisoning) | 5G is toxic or is the cause of COVID-19. |
| | Bill Gates; Gates Foundation; event 201 AND (Bill Gates OR Gates Foundation OR event 201) | Bill Gates is profiting from COVID-19. |
| | hysteri*; hoax; sabotage; paranoi*; (hysteri* OR hoax OR sabotage OR paranoi*); #FilmYourHospital | COVID-19 is a hoax. |
| | informedconsent; parentalrights; researchbeforeyouregret; religiousfreedom; visforvaccine; healthchoice; readtheinsert; healthfreedom; deathtovaccines2020; vis4vaccines; medicalrights; believemothers; medicalfreedomofchoice; childrenshealth; protruth; medicalexemption; vaccinationchoice; betweenmeandmydoctor; dotheresearch; learntherisk; freedomkeepers; vaccineinjury | The COVID-19 crisis will force vaccinations upon citizens. |

**Table 1.** List of queries and their corresponding claims or conspiracy theories.

to obtain the first three pages of search results for all 98 queries, every 20 minutes of every day, between April and October 2020. Our results, summarized in Table 2, indicate that April was a particularly active month for our queries, while June 2020 saw the explicit deletion of many videos that initially landed in our database.

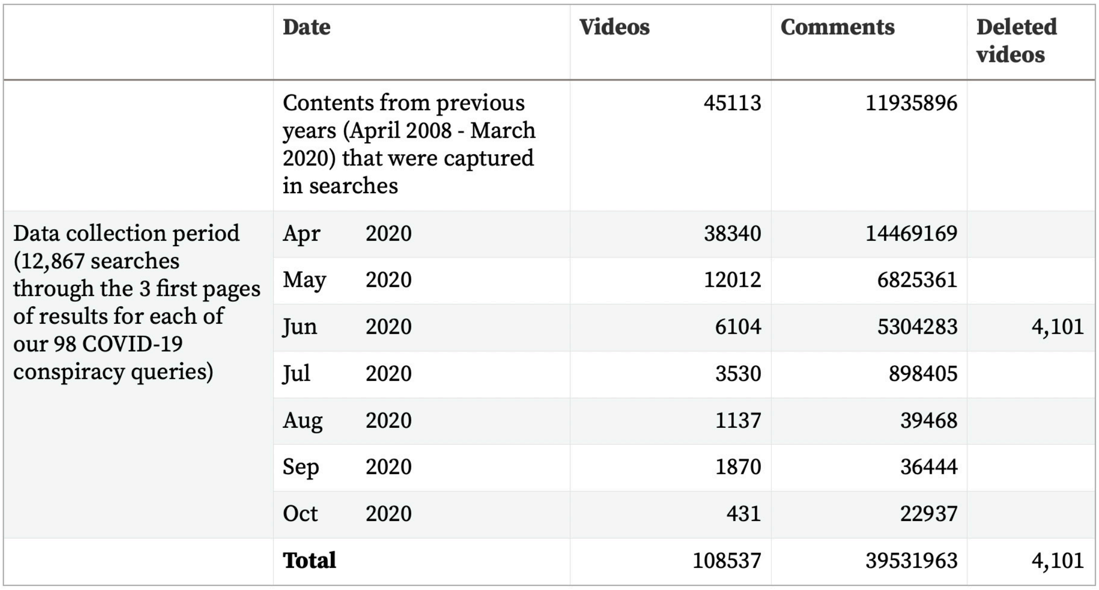
## Understanding YouTube's moderation of COVID-19 misinformation

We used the Wayback Machine's "Changes" tool to perform a close-reading analysis of YouTube's anti-misinformation policies throughout the pandemic (Wayback Machine, 2020). We focused on two policies: (1) *COVID-19 Medical Misinformation* (YouTube, 2020) and (2) *Spam, deceptive practices & scams* (YouTube, 2021a), taking note of (a) the contents they deem problematic; and (b) the moderation techniques they use to sanction them. From these policies, we gathered information on two main content moderation measures: namely, the manipulation of search ranking status and the outright removal of content. The former manifests in demoting "borderline content and harmful misinformation" and up-ranking "authoritative sources"; the latter consists of suspending or deleting "contents that contradict the World Health Organisation and local health authorities" (YouTube, 2020).

As mentioned earlier, borderline contents refer to "videos, comments or channels that do not fall under ban-worthy status for violating the platform's community guidelines, but [come] close to — but [don't] quite cross the line of — violating our Community Guidelines" (YouTube, 2021b). Elsewhere, YouTube describes them as videos "promoting a phony miracle cure for a serious illness, claiming the earth is flat, or making blatantly false claims about historic events like 9/11" (YouTube, 2019b). Though not explicitly phrased as such, we interpreted these examples as referring to conspiracy theories and historical revisionist contents. "Borderline" is the label we gave to videos arguing in favour of COVID-19 conspiracy claims ("WARNING! Digital IDs Will Be Forced On YOU SOON! Why!?" or "Coronavirus COVID-19 | Unfolding Revelation | Agenda ID2020") — distinct, here, from interrogative or speculative titles ("Are Chip Implants the 'Mark of the Beast?'"). We manually coded all search results for the queries "id2020", "wwg1wga", "depopulation" and "5g radiation". As above, we automatically coded (mainstream) news sources for every search result of our four chosen queries.

## Demotion

YouTube describes demotion as up-ranking or "raising" authoritative contents in search results, including news media and other trusted institutional sources



| | Date | Videos | Comments | Deleted videos |
|---|---|---|---|---|
| | Contents from previous years (April 2008 - March 2020) that were captured in searches | 45113 | 11935896 | |
| Data collection period (12,867 searches through the 3 first pages of results for each of our 98 COVID-19 conspiracy queries) | Apr 2020 | 38340 | 14469169 | |
| | May 2020 | 12012 | 6825361 | |
| | Jun 2020 | 6104 | 5304283 | 4,101 |
| | Jul 2020 | 3530 | 898405 | |
| | Aug 2020 | 1137 | 39468 | |
| | Sep 2020 | 1870 | 36444 | |
| | Oct 2020 | 431 | 22937 | |
| | **Total** | 108537 | 39531963 | 4,101 |

**Table 2.** Total number of videos and comments per month, including videos deleted by YouTube or users.

(YouTube, 2021b). It also claims that "borderline contents" are buried in search and recommendation results intentionally (The YouTube Team, 2019). To trace the demotion of videos by borderline channels, we used YouTube search rankings metadata. In three scatterplots, we visualised the ranking position of result per (1) authoritative and borderline quality and (2) query (see **Table 1**), keeping in mind the number of deplatformed videos per month.

## Deplatforming

To complement the restrictions on metadata and API limits, we used youtube-dl to track the status of videos that fell under our query. Between April and June 2020, we were able to track the approximate day that videos previously appearing in query results were no longer available, alongside the message given by YouTube upon visiting a video that was previously captured. We found a total of 4,101 deplatformed videos in June of 2020. To determine why these videos were sanctioned, we first examined the status labels of deleted videos, per query. A majority were not assigned a query, as queries were registered only by the time we began collecting search ranking results in April of 2020. Before then, videos were assigned all queries indiscriminately. In order to determine their contents, we extracted the most prominent words in their audio transcripts using tf-idf (Ramos, 2003).

## Mining for user comments on moderation

In order to discover how users relate to and discuss the impact of moderation on their platform activity, we processed the collected comments of users on the above set of YouTube videos and processed them with word trees — a graphical

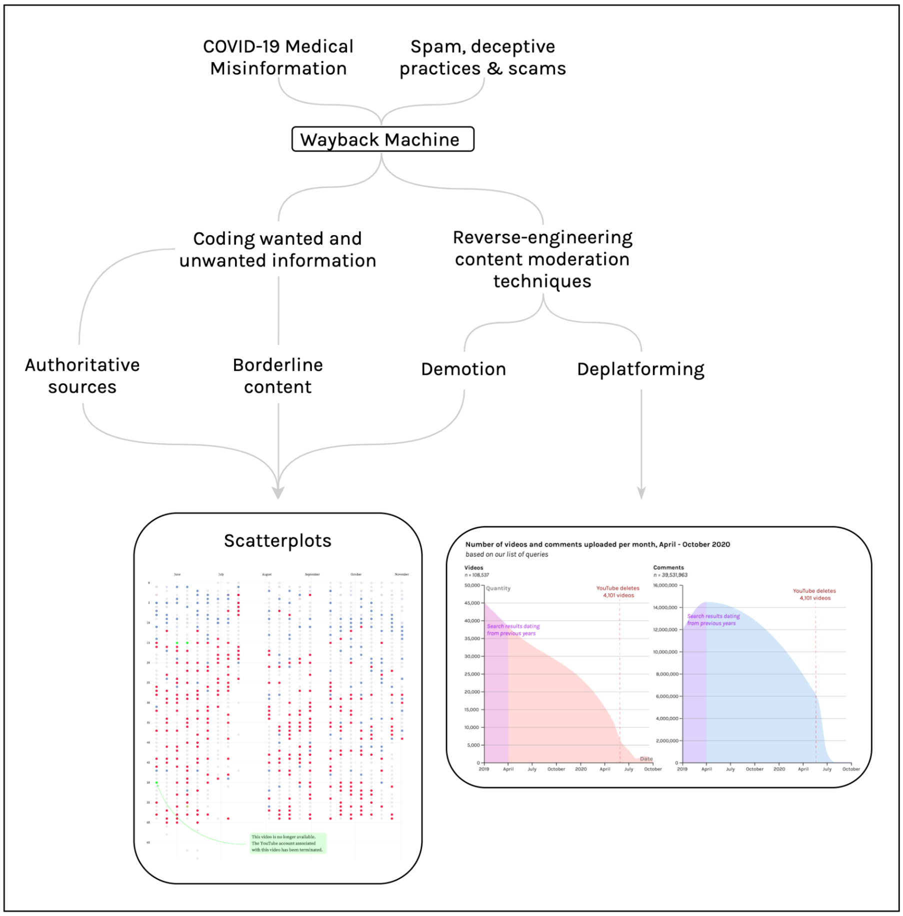

COVID-19 Medical Misinformation
Spam, deceptive practices & scams
Wayback Machine
Coding wanted and unwanted information
Reverse-engineering content moderation techniques
Authoritative sources
Borderline content
Demotion
Deplatforming
Scatterplots
June
July
August
September
October
November
This video is no longer available. The YouTube account associated with this video has been terminated.
Number of videos and comments uploaded per month, April - October 2020
based on our list of queries
Videos
n = 108,537
Quantity
YouTube deletes 4,101 videos
Search results dating from previous years
Date
50,000
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
2019
April
July
October
2020
April
July
October
Comments
n = 39,531,963
YouTube deletes 4,101 videos
Search results dating from previous years
16,000,000
14,000,000
12,000,000
10,000,000
8,000,000
6,000,000
4,000,000
2,000,000
0
2019
April
July
October
2020
April
July
October

**Figure 3.** Analysing YouTube's moderation of COVID-19 conspiracies.

version of keyword-in-contexts (Wattenberg and Viegas, 2008, p. 1221) — and subject-verb-object sentence analysis (Milajevs, Sadrzadeh and Roelleke, 2015). The sentences we extracted were filtered with a list of terms about moderation: “youtube”, “big tech” or “google” and “deplatform*”, “shadowbann*”, “demot*”, “demonetis*” or “demonetiz*”, “suspend*”, “remov*”, “cancel*”, “modera*”, banned.

Our method was informed by Tangherlini et. al.’s *Automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks* (Tangherlini *et al.*, 2020), an ensemble of combined natural language processing techniques that consist in extracting syntactic and semantic elements of conspiracy theories from “noisy” social media posts and visualizing them as storytelling networks. We reconstructed subject-verb-object relations by extracting the relations between named entities; filtering relations based on their reference to conspiratorial narratives; and then filtering resulting networks based on their relationship to particular COVID-19 conspiracy themes. We did this by implementing the Python word2vec library (Rodriguez, 2020), which maps the n-grammatic relations between words through neural networking algorithms. Our basic subjects included “youtube”, “google”, “big tech”, “facebook”, and “twitter”; our verbs were those associated with moderation, such as “cancel”, “remove”, “delete”, “suspend”, “demonetize”; and conspiracy terms included in the query list of **Table 2**. We then normalized the verbs by lemmatizing them and normalized platform shorthands like “yt” and “fb” into their full referents. The results of this can be found in **Figure 13**.

## Methodological obstacles

To investigate demotion and deplatforming, we had to extract data that was not available from YouTube’s API. To investigate deplatforming, for example,

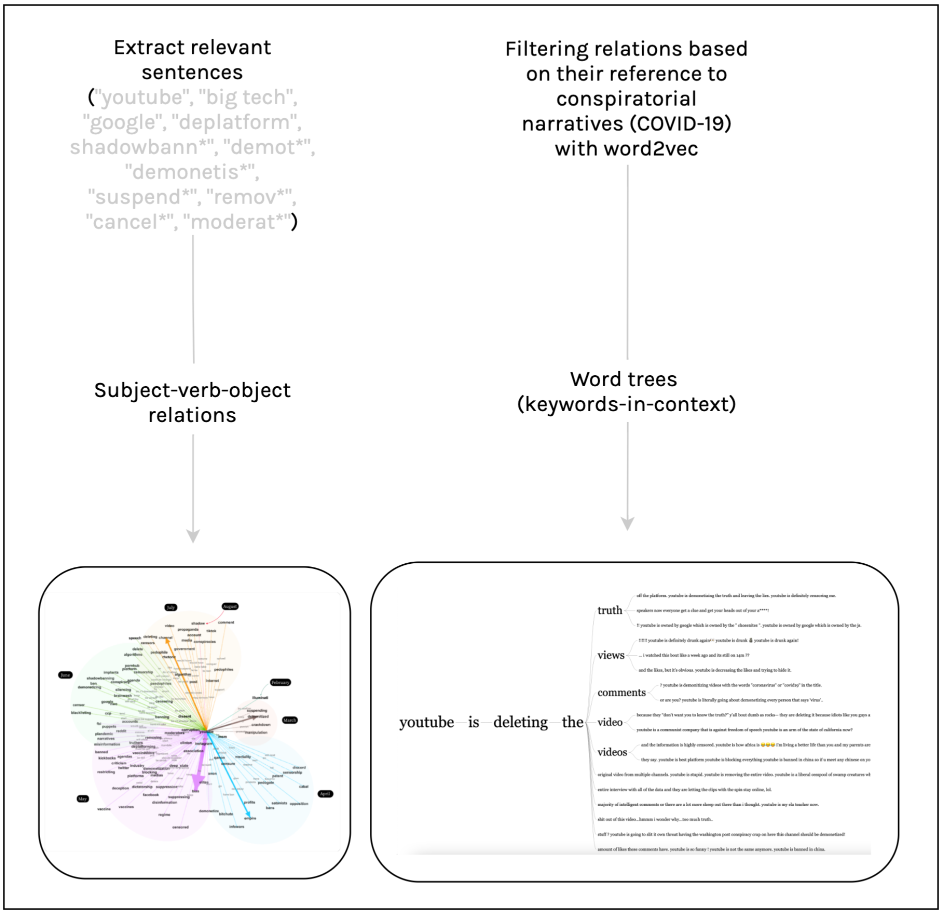

Extract relevant sentences
("youtube", "big tech", "google", "deplatform", shadowbann*", "demot*", "demonetis*", "suspend*", "remov*", "cancel*", "moderat*")
Filtering relations based on their reference to conspiratorial narratives (COVID-19) with word2vec
Subject-verb-object relations
Word trees (keywords-in-context)
youtube is deleting the
truth
views
comments
video
videos

**Figure 4.** Analyzing user reactions to moderation.

we had to track the reason for their removal. Likewise, our method of measuring comments relied on unfiltered access to all comments posted on a given video. These were extracted from the raw data pulled by youtube-dl because querying for a removed video using YouTube's API returns an error, as opposed to the video's potential reason for removal; meanwhile, the comment endpoint of the API does not return all comments (c.f. m. davide, 2019 and Armstead, 2020). This shows us that research that relies on officially-sanctioned platform APIs is not only temporally limited, but actually dealing with a different technical object than those of users (a user clicking a "dead" YouTube link will, indeed, find the reason for the video's removal; a historical trace of problematic content that is hidden from the API endpoint).

Thus, our methods of our study simultaneously act as an argument for methodologies that elide mainstream platform APIs as generative for research. The information given by APIs is, to relative extents, a representational facsimile of the platform itself: APIs provide a particular image of  a programmatic access point to particular data, not the credentials to the platform's databases itself. The API hides important elements relating to the historical presence of problematic content YouTube, and in turn hides the ability to measure how YouTube has related to the problematic content that it hosts over time. These lacks open up the possibility for researchers to offer not only analysis, but critical engagement and questioning that is attuned to the particular content and instantiation of misinformation and other content problematic to platform infrastructures themselves.

# Findings

## Hard and soft moderation on YouTube

Though YouTube's conception of "misinformation" changed significantly since 2013, it has always opted for a mixture of deplatforming and demotion as a solution to any such kind of content. Until 2019, the closest mention of user-generated falsehoods ties mostly to artificial usages of the platform, or "Spam, deceptive practices and scam" (YouTube, 2021a). YouTube's anti-scam policies deal specifically with what the platform classifies as "misleading content", such as false video metadata, blackmail and extortion among users. Sanctions are placed proactively, dispensing user labels and reports for content detection and deletion.

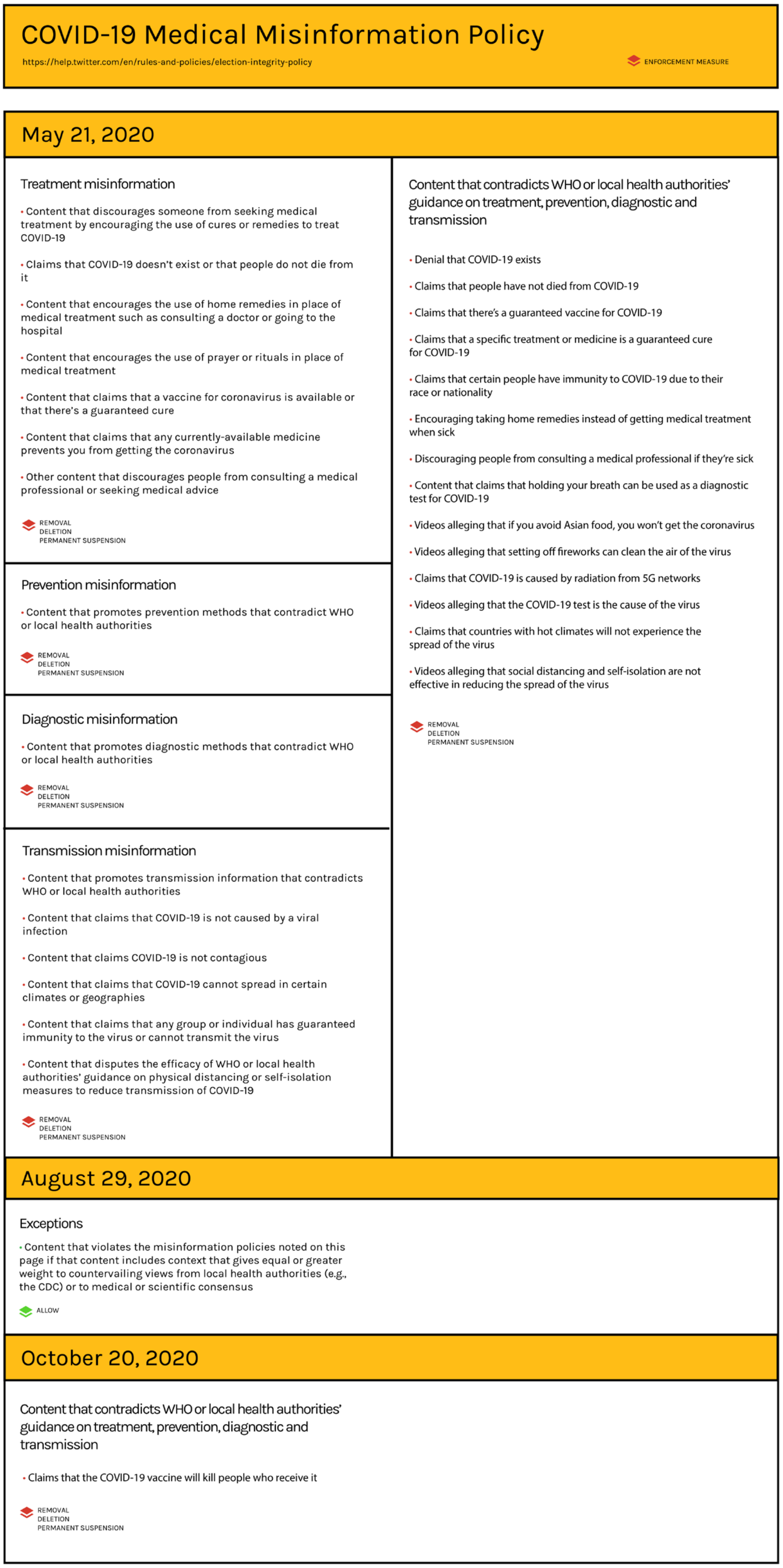
It was not until later that year that "misinformation" is mentioned specifically in the context of ongoing efforts to contain extreme and conspiratorial contents off the platform (YouTube, 2019b). In the context of COVID-19, "misinformation" or "misleading videos" are counterweighted by "authoritative sources", and is described as a series of specific statements that contradict such authorities: as of May 21, 2020, it established a zero-tolerance policy for "content that contradicts the World Health Organisation or local health authorities' guidance on treatment, prevention, diagnostic and transmission." (**Figure 5**). Such contents range from assertions "that COVID-19 doesn't exist or that people do not die from it"; "that COVID-19 is caused by radiation from 5G networks"; or that "the COVID-19 vaccine will kill people who received it."[1] (YouTube, 2020) This hardline approach is justified by the presence of "content where accuracy and authoritativeness are key", and that become in this case crucial to users' health (YouTube, 2021).

On the other hand, YouTube also maintains its demotion policy for borderline contents while increasing its efforts to centralise access to authoritative sources on search, recommendation results and homepages (YouTube, 2021b). Contrary to hardline policies, demotion grants some exceptions to problematic contents: "recommendations systems do not proactively recommend [borderline] content", but borderline videos may still "appear in recommendations for channel subscribers and in search results" (YouTube, 2021b).

## Hard moderation: deplatforming

The effects of "hard moderation" are especially palpable while examining the quantity of misinformation that appears through YouTube's discovery mechanisms. We find that, after implementing its COVID-19 Medical Misinformation policy on May 21, 2021, the numbers of videos making conspiratorial claims decreased steadily

[1] YouTube only makes one exception by August 29th, which it later removes: artistic or critical contents that violate this policy, in the condition they also grant equal balance to "countervailing views from local health authorities [...] or to medical or scientific consensus."



## COVID-19 Medical Misinformation Policy

https://help.twitter.com/en/rules-and-policies/election-integrity-policy

ENFORCEMENT MEASURE

### May 21, 2020

Treatment misinformation

- Content that discourages someone from seeking medical treatment by encouraging the use of cures or remedies to treat COVID-19
- Claims that COVID-19 doesn't exist or that people do not die from it
- Content that encourages the use of home remedies in place of medical treatment such as consulting a doctor or going to the hospital
- Content that encourages the use of prayer or rituals in place of medical treatment
- Content that claims that a vaccine for coronavirus is available or that there's a guaranteed cure
- Content that claims that any currently-available medicine prevents you from getting the coronavirus
- Other content that discourages people from consulting a medical professional or seeking medical advice

REMOVAL
DELETION
PERMANENT SUSPENSION

Prevention misinformation

- Content that promotes prevention methods that contradict WHO or local health authorities

REMOVAL
DELETION
PERMANENT SUSPENSION

Diagnostic misinformation

- Content that promotes diagnostic methods that contradict WHO or local health authorities

REMOVAL
DELETION
PERMANENT SUSPENSION

Transmission misinformation

- Content that promotes transmission information that contradicts WHO or local health authorities
- Content that claims that COVID-19 is not caused by a viral infection
- Content that claims COVID-19 is not contagious
- Content that claims that COVID-19 cannot spread in certain climates or geographies
- Content that claims that any group or individual has guaranteed immunity to the virus or cannot transmit the virus
- Content that disputes the efficacy of WHO or local health authorities' guidance on physical distancing or self-isolation measures to reduce transmission of COVID-19

REMOVAL
DELETION
PERMANENT SUSPENSION

Content that contradicts WHO or local health authorities' guidance on treatment, prevention, diagnostic and transmission

- Denial that COVID-19 exists
- Claims that people have not died from COVID-19
- Claims that there's a guaranteed vaccine for COVID-19
- Claims that a specific treatment or medicine is a guaranteed cure for COVID-19
- Claims that certain people have immunity to COVID-19 due to their race or nationality
- Encouraging taking home remedies instead of getting medical treatment when sick
- Discouraging people from consulting a medical professional if they're sick
- Content that claims that holding your breath can be used as a diagnostic test for COVID-19
- Videos alleging that if you avoid Asian food, you won't get the coronavirus
- Videos alleging that setting off fireworks can clean the air of the virus
- Claims that COVID-19 is caused by radiation from 5G networks
- Videos alleging that the COVID-19 test is the cause of the virus
- Claims that countries with hot climates will not experience the spread of the virus
- Videos alleging that social distancing and self-isolation are not effective in reducing the spread of the virus

REMOVAL
DELETION
PERMANENT SUSPENSION

### August 29, 2020

Exceptions

- Content that violates the misinformation policies noted on this page if that content includes context that gives equal or greater weight to countervailing views from local health authorities (e.g., the CDC) or to medical or scientific consensus

ALLOW

### October 20, 2020

Content that contradicts WHO or local health authorities' guidance on treatment, prevention, diagnostic and transmission

- Claims that the COVID-19 vaccine will kill people who receive it

REMOVAL
DELETION
PERMANENT SUSPENSION

**Figure 5.** Overview of YouTube's Medical Misinformation policy.

on the platform. Of all the 108,537 videos we captured, 4,101 were unavailable by June, 2020 (**Figure 6**). Most of the moderation prompts that occurred stated simply that videos were removed for violating YouTube's community guidelines, or were simply shown as unavailable or as belonging to a terminated user account. Only one small number of videos have been removed for inciting hatred, involving violence or harassment, or because of copyright claims (see **Annex**).

Looking at the types of content YouTube deleted (**Figure 6**), we find that YouTube targets specific claims listed in YouTube's *COVID-19 Medical Misinformation* policy (**Figure 5**), namely allegations that COVID-19 is caused by 5G radiation; that vaccination is an operation to implement microchips; that the virus being a (Chinese) bioweapon or a concocted hysteria; claims that one can use prayers and other spiritual methods for treating the virus. Here, the hardline aspects of deplatforming are especially visible in the absence of tolerance for contents that are usually up to users' discretion, namely religious beliefs.

## Soft moderation: demotion

We find three main tendencies within YouTube's demotion of conspiratorial videos on COVID-19. The first is the effective demotion of videos making claims contrary to the World Health Organisation and local health authorities' guidelines (for example, a video titled "Claims that COVID-19 is caused by radiation from 5G networks") and "borderline contents" (both in red in **Figure 7**), as well as the up-ranking of videos by mainstream news channels or "authoritative sources" (in blue). This is particularly applicable to results for the query "id2020" (**Figure 7**). "Id2020" refers to a microchip that users allege will be sold in combination with COVID-19 vaccines promoted by Bill Gates; YouTube refers to variations of this conspiracy theory as "Claims that the COVID-19 vaccine will contain a microchip or tracking device." (YouTube, 2020)

The second tendency, which casts doubt on long-term effectiveness of the first, is the eventual resurfacing of borderline contents on top of search ranking results. We see that few borderline contents surface in search results for the query "COVID depopulation" up until July of 2020, which YouTube sanctions as videos that "claim that the COVID-19 vaccine will be used as a means of population reduction" (**Figure 8**). In August of 2020, the video "IS THERE AN AGENDA BEHIND THIS VIRUS? - COVID-19 Government Agenda" remains on top for over two months, starting in late August, 2020.

The third tendency is the effect of deplatforming on demoting borderline videos. Queries related to Qanon, such as the Q motto "wwg1wga" (where we go one, we go all), were largely left undetected until YouTube cracked down on the conspiracy around early October (Sandler, 2020). This highlights a symbiotic relationship between the two techniques; as evidenced by **Figures 7, 8 and 9**, in few instances do any of the two work independently of each other.

## Reactions to hard moderation

How are hard and soft moderation techniques interpreted by affected users? At first glance, we see similarities with Myer West's findings (2018) on the "stigmatised" status of moderated knowledge (Barkun, 2017). In the absence of clear (and trusted) justifications for moderation, users effectively draw "connections between related phenomena, developing non-authoritative conceptions of why and how their content was removed" (Eslami et al., 2015 and Kempton, 1986 in Myers West, 2018, 8). Users perceive moderation as an activity of YouTube engaging in the censorship and deletion of various undisclosed truths on the virus (**Figure 10**). Some complement the absence of clear reasons for deletion (see **Annex**) with more elaborate theorisations of YouTube's motives for deleting contents, in that some claim that video testimony of doctors and nurses disappear as part of a general cover-up for the spread of crowdsourced information. In the position of the conspirator, YouTube is perceived as operating for political motives, with some users complaining that YouTube acts as "a liberal cesspool of swamp creatures."

The relations between a wider swath of conspiratorial narratives about the political affiliations and obscure motives of YouTube are made clearer in **Figure 11**, which indicates words that are associated with "YouTube'' in user comments. As we move clockwise throughout the months the graph represents, we see a shift in commenters' interpretations of the motives behind YouTube's content moderation. While in March, commenters remarked that YouTube kept on "suspending", "demonetising" or "manipulating" their contents, in May they accused the platform of "bias" and progressively of suppression, censorship and blacklisting. As critiques become suspicions of persecution, commenters formulate more elaborate explanations of YouTube's political functions as covering governments, supporting pedophiles, or

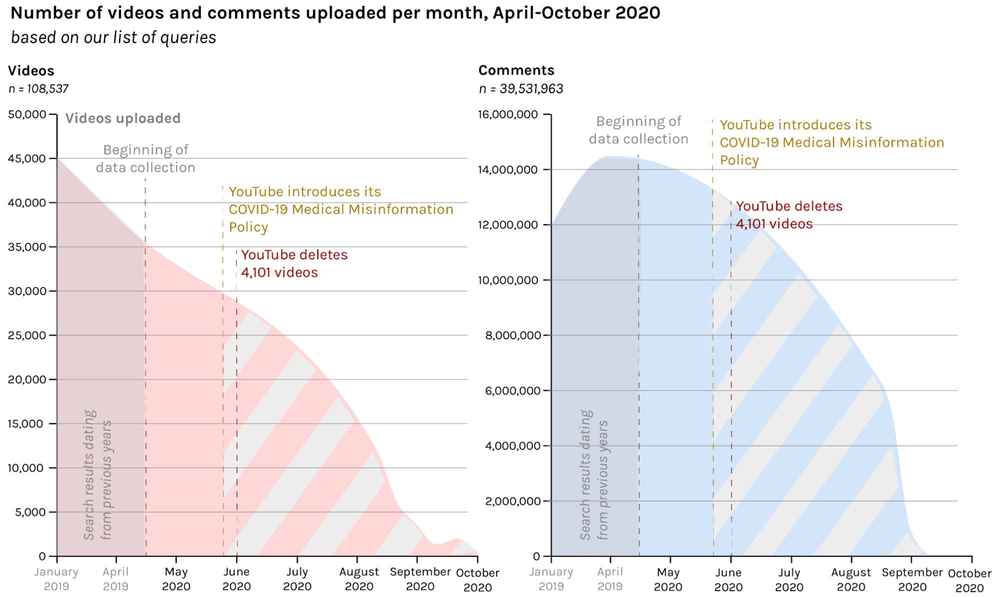


**Figure 6.** Number of deplaformed videos and comments between April and October of 2020. Striped sections represent periods in which COVID-19 misinformation policies come into force.

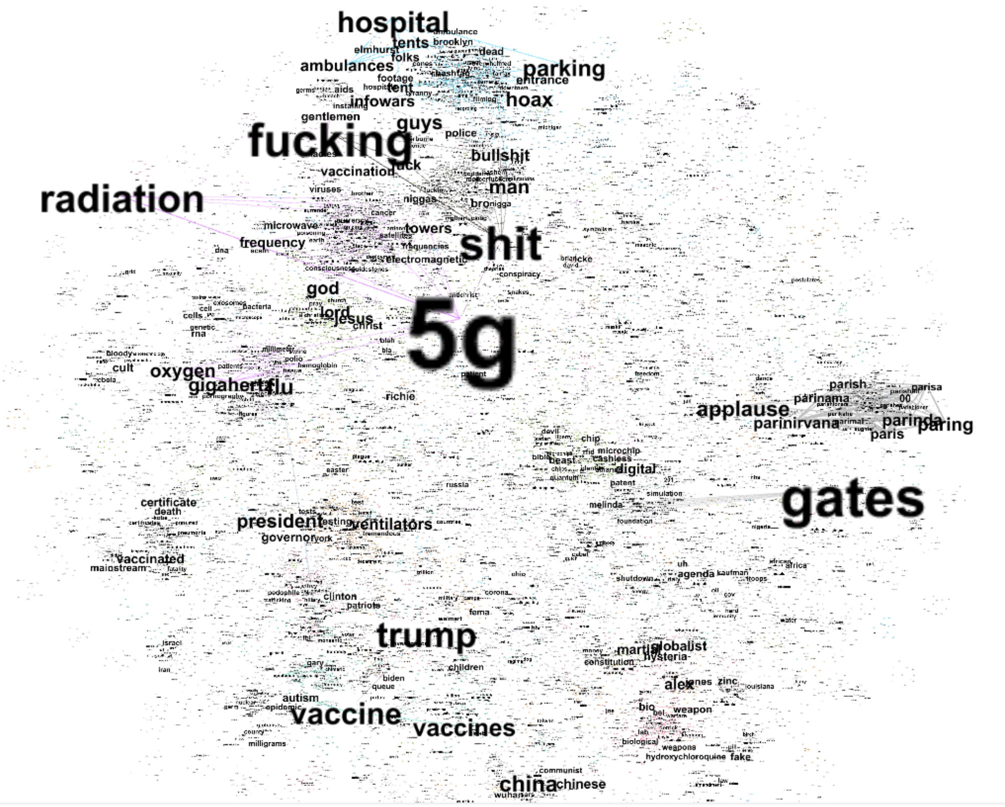
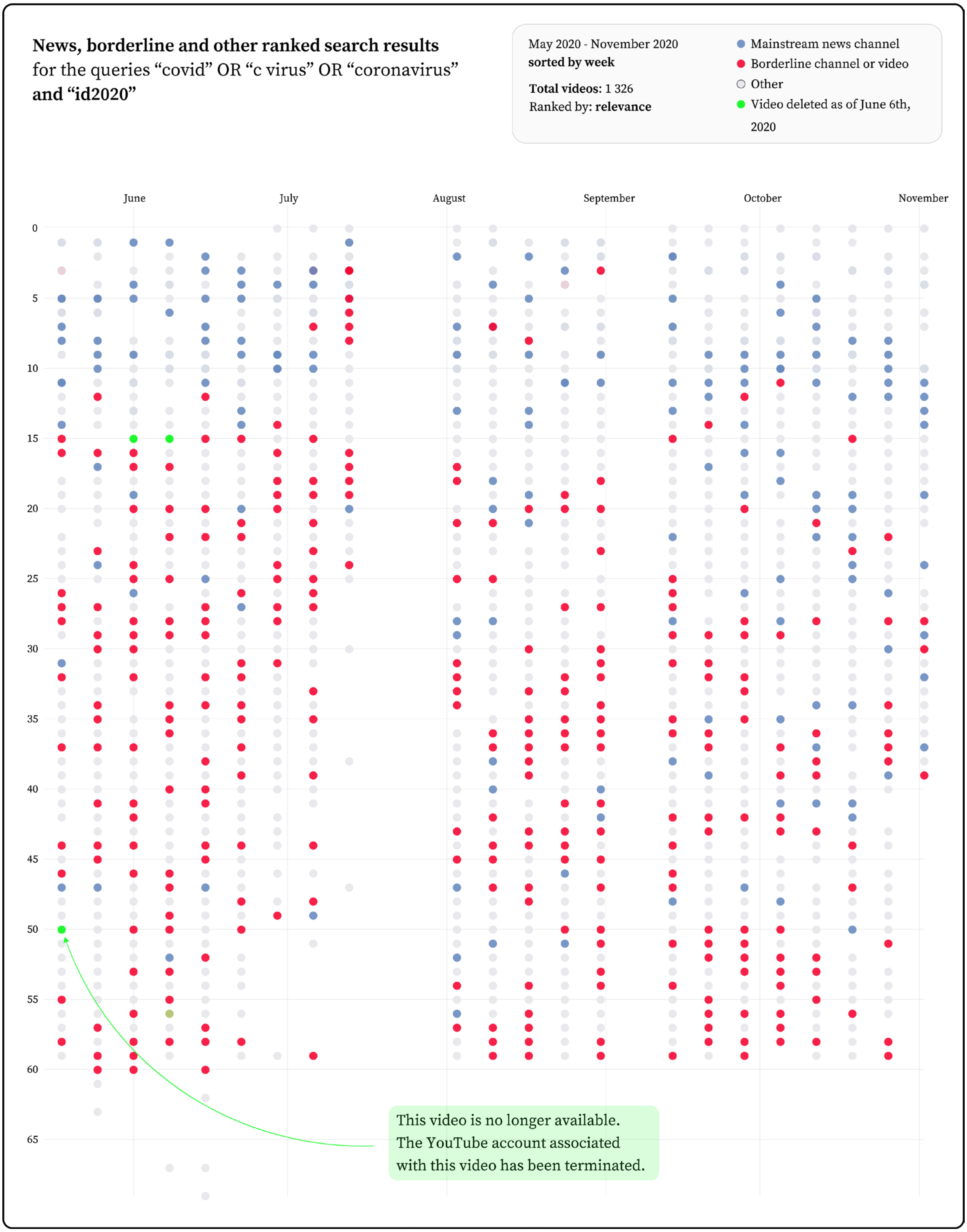


**Figure 6.** Tf-idf network of all 4,101 banned videos, clustered by modularity.



This video is no longer available. The YouTube account associated with this video has been terminated.

**Figure 7.** Demoted and up-ranked search results for the query “id2020”.

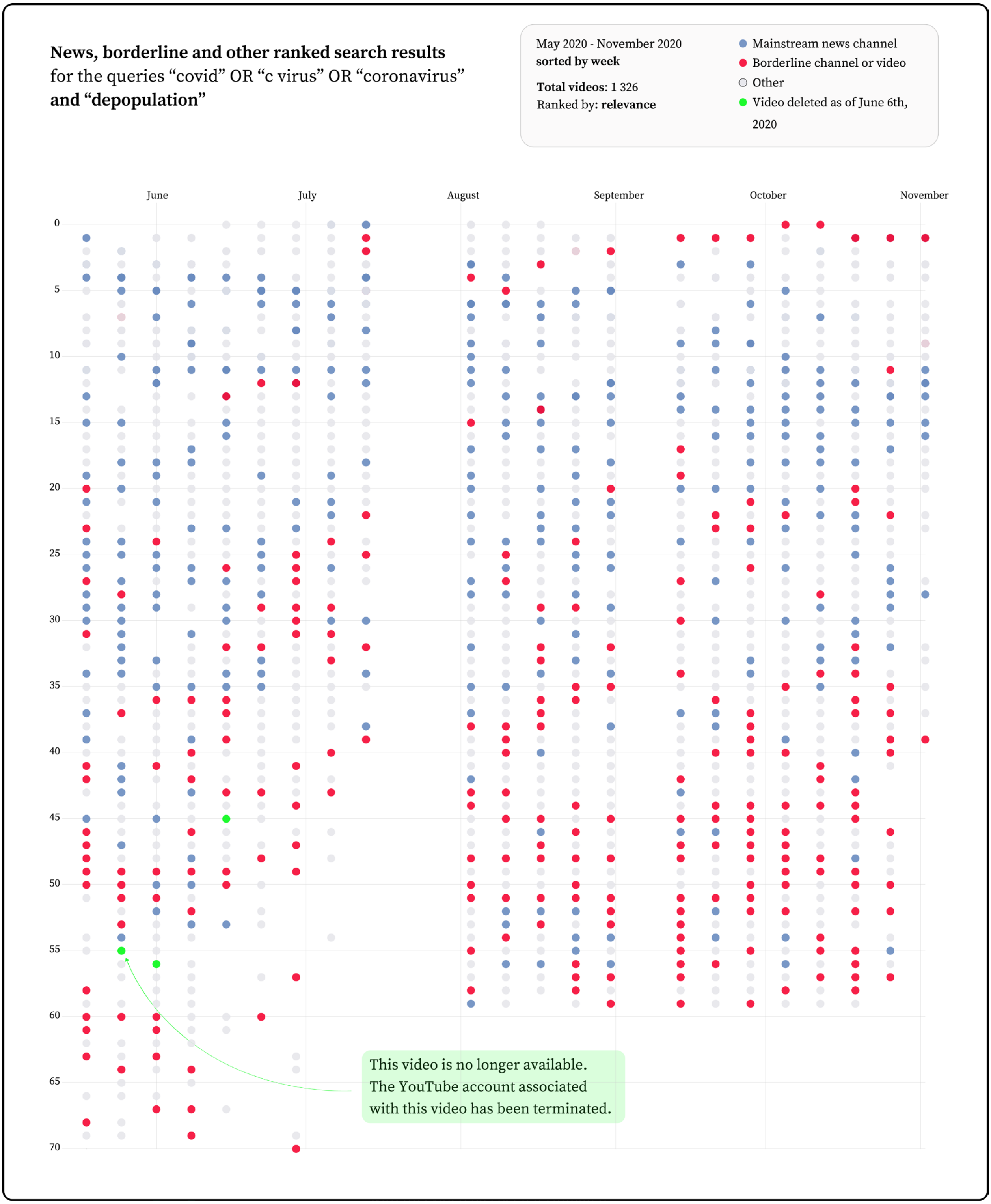


**Figure 8.** Demoted and up-ranked search results for the query "depopulation".

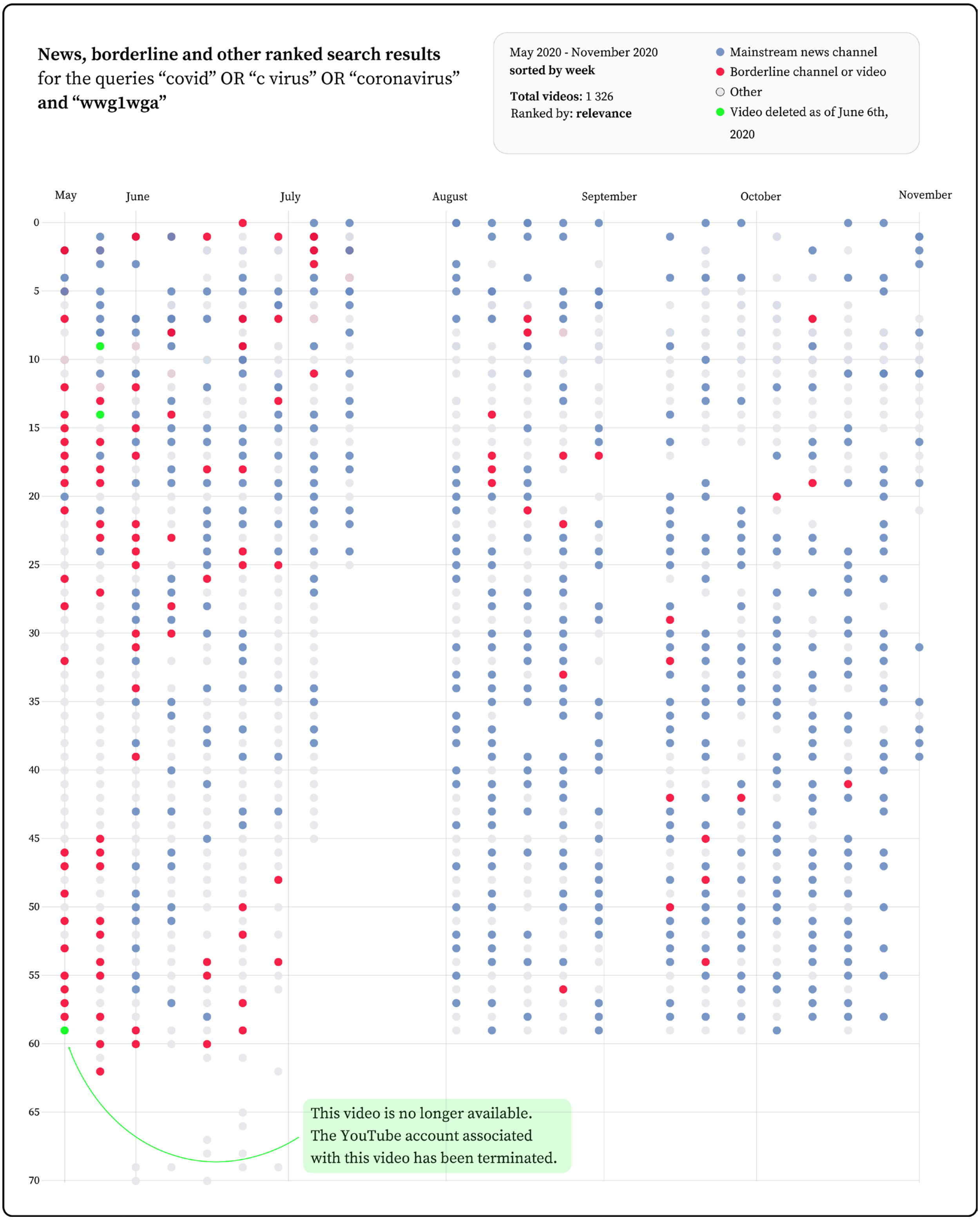
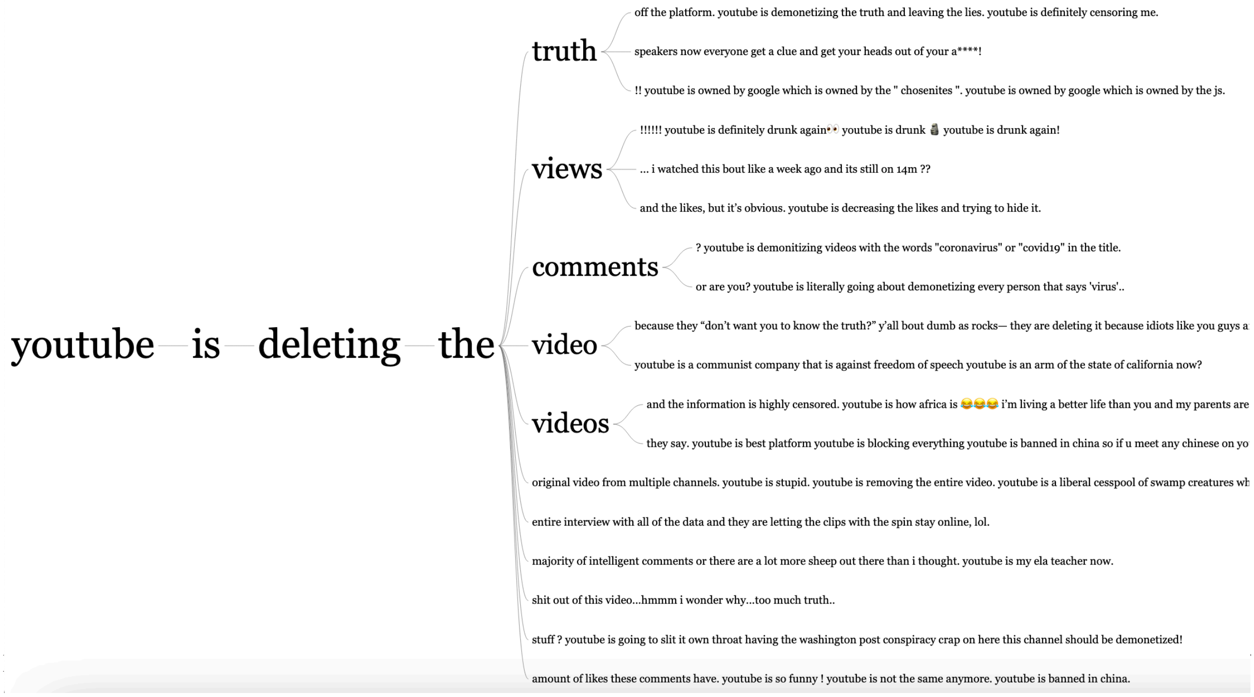


**Figure 9.** Demoted and up-ranked search results for the query "wwg1wga".



off the platform. youtube is demonetizing the truth and leaving the lies. youtube is definitely censoring me.

speakers now everyone get a clue and get your heads out of your a****!

!! youtube is owned by google which is owned by the " chosenites ". youtube is owned by google which is owned by the js.

!!!!!! youtube is definitely drunk again👀 youtube is drunk 🗿 youtube is drunk again!

... i watched this bout like a week ago and its still on 14m ??

and the likes, but it's obvious. youtube is decreasing the likes and trying to hide it.

? youtube is demonitizing videos with the words "coronavirus" or "covid19" in the title.

or are you? youtube is literally going about demonetizing every person that says 'virus'..

because they "don't want you to know the truth?" y'all bout dumb as rocks— they are deleting it because idiots like you guys a

youtube is a communist company that is against freedom of speech youtube is an arm of the state of california now?

and the information is highly censored. youtube is how africa is 😂😂😂 i'm living a better life than you and my parents are

they say. youtube is best platform youtube is blocking everything youtube is banned in china so if u meet any chinese on yo

original video from multiple channels. youtube is stupid. youtube is removing the entire video. youtube is a liberal cesspool of swamp creatures wh

entire interview with all of the data and they are letting the clips with the spin stay online, lol.

majority of intelligent comments or there are a lot more sheep out there than i thought. youtube is my ela teacher now.

shit out of this video...hmmm i wonder why...too much truth..

stuff ? youtube is going to slit it own throat having the washington post conspiracy crap on here this channel should be demonetized!

amount of likes these comments have. youtube is so funny ! youtube is not the same anymore. youtube is banned in china.

**Figure 10.** Word tree for "youtube is deleting the". Every line is a comment (n = 19 comments).

stellation describes the through-lines that build between YouTube's content moderation and a concern regarding the suppression of stigmatised truths (Krug, 2016).

## Reactions to soft moderation

User reactions to demotion are illustrated by comments around "shadowbanning", a vernacular term that refers to algorithmic interventions to reduce the visibility and spread of user-generated contents. Of note is a certain attunement to potential demotion: remarks that a specific video "does not come up at all in youtube search", that certain contents "make it through [a] filter", or that a video that was previously banned is now "coming to the light" (**Figure 12**). This supposed demotion is complemented by suspicions that personal characteristics, beliefs or sanctioned knowledge must be the object of persecution. Emerging from this is the suspicion that moderation is biased, which seems to violate YouTube's claims to provide a *platform* — in the sense of a democratically-accessible venue for speech — for user-generated content ("They pretend to be platforms but they are not"). the production of conspiratorial information altogether. "Authoritative" sources, like videos by scientific experts, local health authorities or mainstream news media, are typically the object of conspiratorial suspicions in the comment sections. One top ranking video for the query "depopulation", "Empty Planet: Preparing for the Global Population Growth" by the Centre for International Governance Innovation, shows political scientist Darrell Bricker and journalist John Ibbitson having an armchair discussion around the thesis that, contrary to popular knowledge, the human population is likely to decline dramatically (Centre for International Governance Innovation, 2019). Commenters see this discussion as evidence that COVID-19 is part of a covert plan to decrease the world's human population, and that the speakers and the institutions they mention are enmeshed in an elite that knew of this policy far before the pandemic struck (**Figure 13**).

## Conclusion

Through the combined historical analysis of content moderation techniques, the moderation techniques related to COVID-19 misinformation and their effects, this study

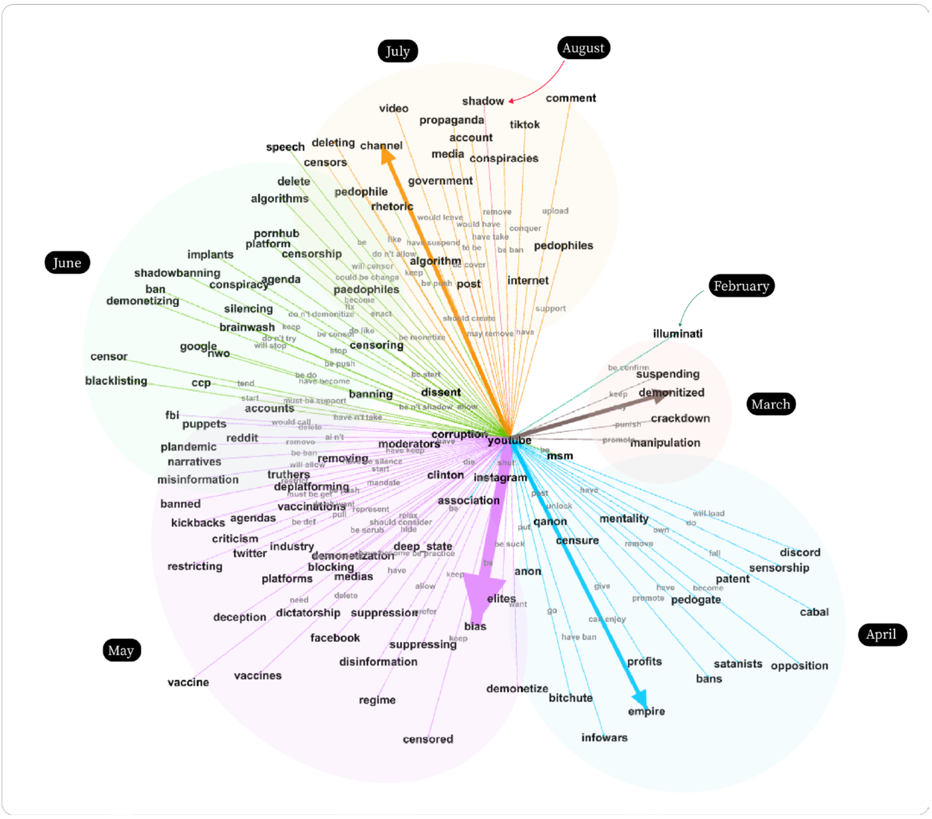
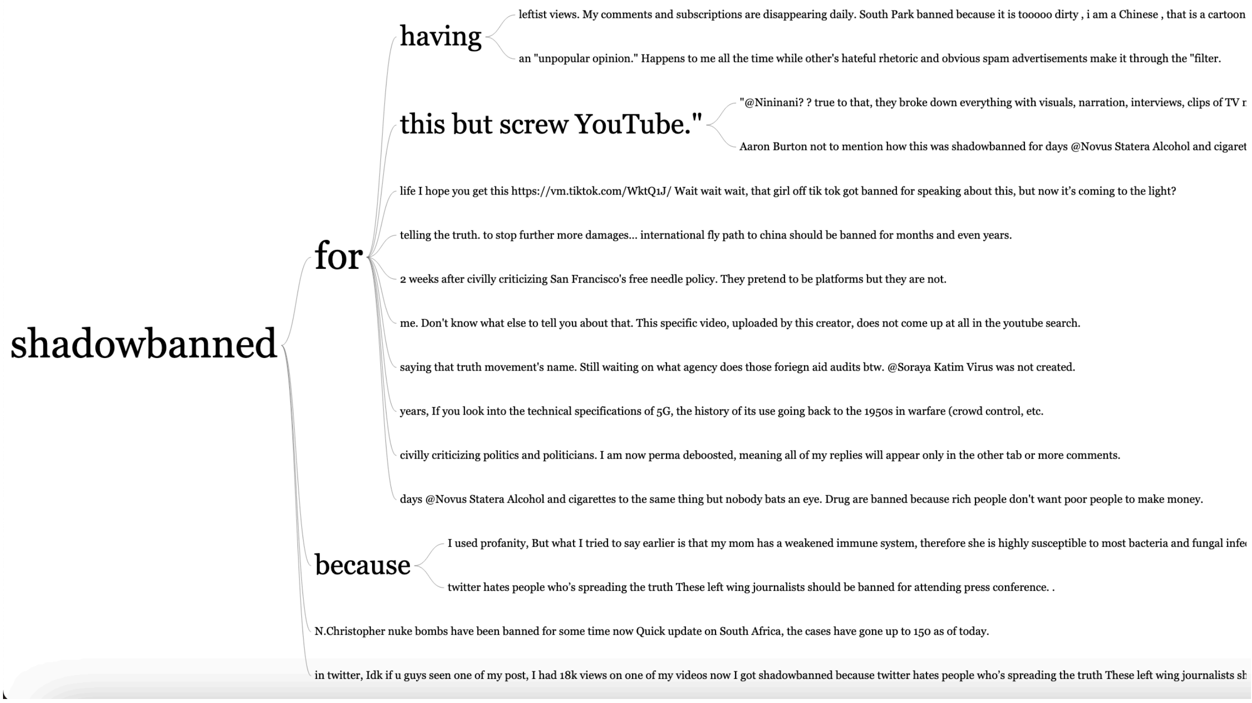


**Figure 11.** Subject-verb-object network of sentences mentioning YouTube, moderation and COVID-19 conspiracies. Colours indicate the months in which comments were uploaded. Line thickness indicates the strength of the n-grammatic association between the root word, "YouTube", and the word at the end of the node.


shadowbanned

for

having

leftist views. My comments and subscriptions are disappearing daily. South Park banned because it is tooooo dirty , i am a Chinese , that is a cartoon

an "unpopular opinion." Happens to me all the time while other's hateful rhetoric and obvious spam advertisements make it through the "filter.

this but screw YouTube."

"@Nininani? ? true to that, they broke down everything with visuals, narration, interviews, clips of TV r

Aaron Burton not to mention how this was shadowbanned for days @Novus Statera Alcohol and cigaret

life I hope you get this https://vm.tiktok.com/WktQ1J/ Wait wait wait, that girl off tik tok got banned for speaking about this, but now it's coming to the light?

telling the truth. to stop further more damages... international fly path to china should be banned for months and even years.

2 weeks after civilly criticizing San Francisco's free needle policy. They pretend to be platforms but they are not.

me. Don't know what else to tell you about that. This specific video, uploaded by this creator, does not come up at all in the youtube search.

saying that truth movement's name. Still waiting on what agency does those foriegn aid audits btw. @Soraya Katim Virus was not created.

years, If you look into the technical specifications of 5G, the history of its use going back to the 1950s in warfare (crowd control, etc.

civilly criticizing politics and politicians. I am now perma deboosted, meaning all of my replies will appear only in the other tab or more comments.

days @Novus Statera Alcohol and cigarettes to the same thing but nobody bats an eye. Drug are banned because rich people don't want poor people to make money.

because

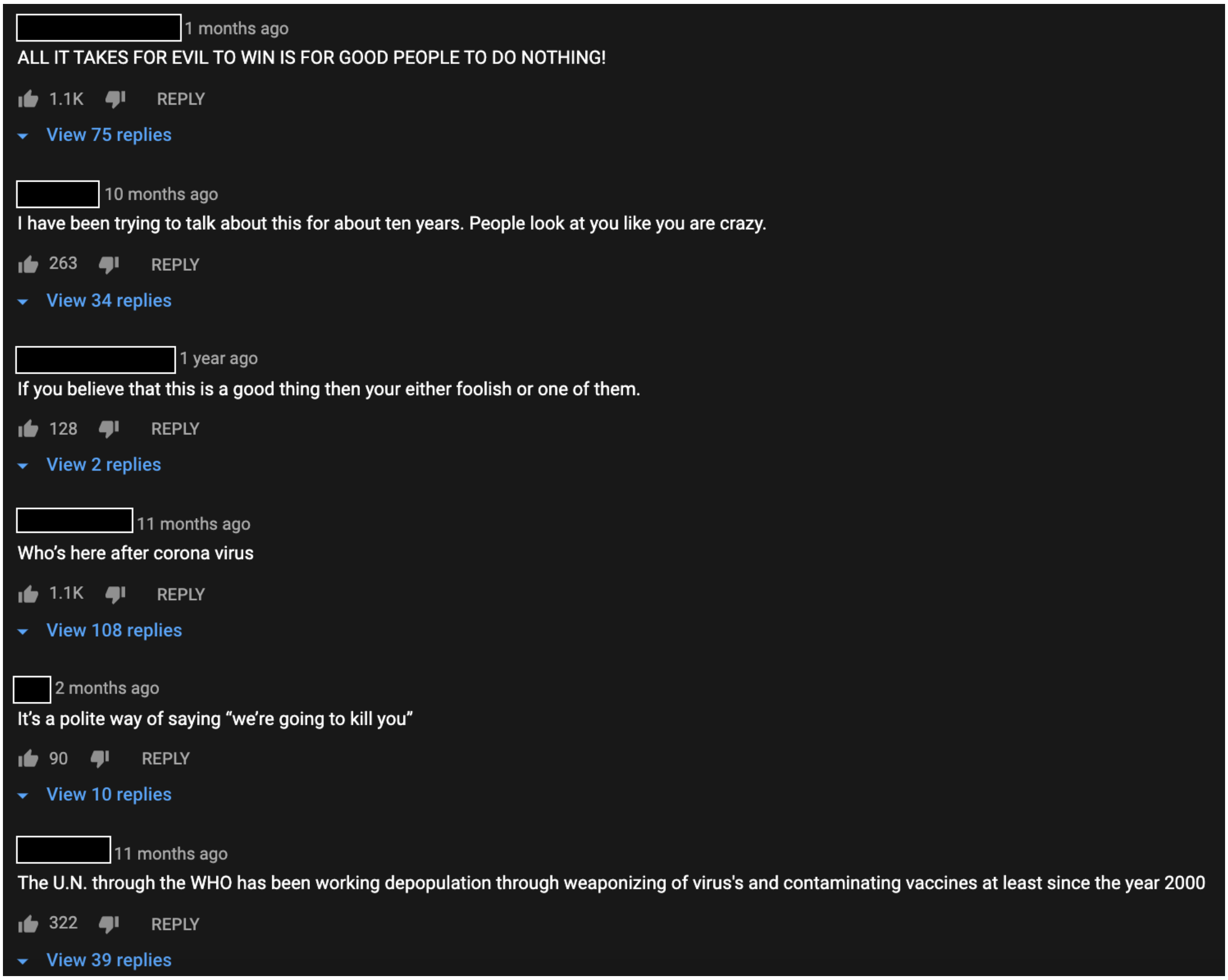
I used profanity, But what I tried to say earlier is that my mom has a weakened immune system, therefore she is highly susceptible to most bacteria and fungal infe

twitter hates people who's spreading the truth These left wing journalists should be banned for attending press conference. .

N.Christopher nuke bombs have been banned for some time now Quick update on South Africa, the cases have gone up to 150 as of today.

in twitter, Idk if u guys seen one of my post, I had 18k views on one of my videos now I got shadowbanned because twitter hates people who's spreading the truth These left wing journalists sh

**Figure 12.** Word tree for "shadowbanned" (n = 16 comments).



[redacted] 1 months ago

ALL IT TAKES FOR EVIL TO WIN IS FOR GOOD PEOPLE TO DO NOTHING!

[redacted] 10 months ago

I have been trying to talk about this for about ten years. People look at you like you are crazy.

[redacted] 1 year ago

If you believe that this is a good thing then your either foolish or one of them.

[redacted] 11 months ago

Who's here after corona virus

[redacted] 2 months ago

It's a polite way of saying "we're going to kill you"

[redacted] 11 months ago

The U.N. through the WHO has been working depopulation through weaponizing of virus's and contaminating vaccines at least since the year 2000

**Figure 13.** Screenshot of top-ranking comments for the video "Empty Planet: Preparing for the Global Population Growth", taken on January 20th, 2021. User names and profile pictures have been anonymised.

found that since 2013, YouTube has applied a combination of "hard" and "soft" moderation techniques to contain the spread of misinformation. The term "misinformation" is however quite new — and more controversial — to its repertoire of content moderation policies. The closest translation of this concept brings us back to YouTube's *Spam, deceptive practices and scam policies*, which to this day aim to suppress false and actively harmful usages or appropriations of the platform: placing false video metadata; gaming the recommender system; posting the same contents repeatedly; or involving users in scams, blackmail and extortion. The circulation of contents that are not immediately condemnable as doctored, coordinated or actively misinformative contents — what "comes close to [...] but doesn't quite cross the line of — violating [...] Community Guidelines" — pushes the platform to both formulate its own conceptions of misinformation as "borderline content", as well as the techniques it uses to contain it.

The advent of COVID-19 as a public health crisis has pushed the platform to take a far more proactive approach to moderating misinformation. At first sight, this approach is visible in YouTube's choice to prioritise "authoritative sources" over contents produced by users at large, regardless of their popularity or affinity with user preferences. Despite the essentially contingent status of knowledge on COVID-19, the platform has had to shrink its margin of tolerance for user-generated contents to a set of very specific statements about the virus. In the absence of stable markers of truth, these statements must facilitate user consensus or convergence around scientific, political and media authorities, or "authoritative sources". In this sense, the choice for authoritative sources indicates an attempt to moderate possible excesses of factual contingency amongst users and the vacuums of (epistemic) leadership they may cause.

This is also evident in a technical sense, through the use of deplatforming and demotion techniques. Both techniques act as correctives, where user-generated contents are either deleted or down-ranked in favour of authoritative statements. Though we have found a relatively small number of deplatformed videos (4,101 between April and October 2020), we see that, combined with demotion techniques, moderation has more often than not successfully targeted "borderline contents" and misleading statements, despite occasional failures to capture all of them or newly emerging ones.

Nevertheless, it is arguably the very success of deplatforming and demotion policies that render the platform vulnerable to misinformation. As YouTube actively qualifies and intervenes in user-generated contents, its once invisible back-end measures and policies surface as operationalised biases. This personalisation of YouTube as an agent for partisan norms breaks the illusion of universal platforms that Gillespie once located in the public-facing "discourse of platforms" (Gillespie, 2013). The idea of a seemingly open host for user-generated contents is, in this case, replaced by a critique that platforms are a universal space for instrumentalizing partisanship, be it for users flagging, deleting and otherwise moderating each other, or for a normative body of thought from elites or the platform itself.

This perceived encroachment sheds light on contemporary attitudes towards platforms. In such sociotechnical imaginaries (Jasanoff 2015, Bucher 2017), YouTube should only hold itself to the informational functions of a media platform, without intervening in the contents that YouTube functionally prefigures. This illustrates a "cognitive mapping" (Jameson 1990) wherein the indiscriminate, purely functionalist rationality of the platform-as-code, or non-human agent, is imagined and understood as its "real" schema. The *positive* functions of YouTube's mode of circulating videos — providing it a home, making it easy to access, and linking it together with similar video content — are taken not only as the platform's function, but its purpose. Despite their entanglements with the very same code that sets up these positive functions, the acts of deplatforming, censorship, and demotion (in short, actions that *impede* free information circulation) are viewed as a violation of this imagined function.

But the degree to which conspiracy theories are entangled with larger metahistorical precursors illustrates the difficulties in applying the binary logic of free-or-censored to thick webs of relation. For example, "authoritative" content may not contain misinformation, but may nevertheless be used to substantiate misinformation — such as videos showing intelligentsia discussing depopulation scenarios a few months before the pandemic. This indicates the ways that "authoritative sources" piggyback onto existing conspiracy narratives and, more importantly, illustrates the primary tension in the spread of misinformation: when ideas are networked so robustly, a single platform's attempts to draw the line between what is acceptable and unacceptable makes a modest impact on their broader information ecology.

What to make of this? While this question certainly can't be answered with a case study on YouTube alone, our conclusion, for now, is that proactive and more explicit moderation contributes to modifying the relationship between platform and user. We have seen that, post-deplatforming and demotion, YouTube is less a platform for *conspiracy-making* (i.e., sharing, discussing and amalgamating evidence of conspiracies) than a platform *by* and *for* the conspirators. That is to say that YouTube is not altogether abandoned by affected user bases, but is complemented by the usage of platforms that afford conspiracy-making — "alt-tech" sites such as Bitchute, Rumble, Parler or Telegram — as an archive of primary sources.

## Referências

Armstead, C. (2020, November 12). python requests—Why is the YouTube API v3 inconsistent with the amount of comments it lets you download before an error 400? Stack Overflow. https://stackoverflow.com/questions/64809133/why-is-the-youtube-api-v3-inconsistent-with-the-amount-of-comments-it-lets-you-d

Arthurs, J., Drakopoulou, S., & Gandini, A. (2018). Researching YouTube. Convergence, 24(1), 3–15. https://doi.org/10/gfzb8w

Barkun, M. (2016) 'Conspiracy Theories as Stigmatized Knowledge', *Diogenes*, p. 0392192116669288. doi: 10.1177/0392192116669288.

Bruns, A. (2019) 'After the "APIcalypse": social media platforms and their fight against critical scholarly research', *Information, Communication & Society*, 22(11), pp. 1544–1566. doi: 10.1080/1369118X.2019.1637447.

Bucher, T. (2017) 'The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms', *Information, Communication & Society*, 20(1), pp. 30–44. doi: 10.1080/1369118X.2016.1154086.

Centre for International Governance Innovation (2019) *Empty Planet: Preparing for the Global Population Decline*. Available at: https://www.youtube.com/watch?v=bSAgHvETNSg (Accessed: 24 August 2021).

Chandrasekharan, E. *et al.* (2021) 'Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit', *arXiv:2009.11483 [cs]*. Available at: http://arxiv.org/abs/2009.11483 (Accessed: 11 November 2020).

Constine, J. (2018) 'Facebook will change algorithm to demote "borderline content" that almost violates policies', *TechCrunch*, 15 November. Available at: https://social.techcrunch.com/2018/11/15/facebook-borderline-content/ (Accessed: 20 February 2021).

Covington, P., Adams, J. and Sargin, E. (2016) 'Deep Neural Networks for YouTube Recommendations', in. ACM Press, pp. 191–198. doi: 10.1145/2959100.2959190.

Davidson, J. *et al.* (2010) 'The YouTube video recommendation system', in *Proceedings of the fourth ACM conference on Recommender systems*. ACM, pp. 293–296. Available at: http://dl.acm.org/citation.cfm?id=1864770 (Accessed: 25 September 2016).

Douek, E. (2020) 'The Year That Changed the Internet', *The Atlantic*, 28 December. Available at: https://www.theatlantic.com/ideas/archive/2020/12/how-2020--forced-facebook-and-twitter-step/617493/ (Accessed: 6 July 2021).

Ekman, M. (2014). The dark side of online activism: Swedish right-wing extremist video activism on YouTube. MedieKultur: Journal of Media and Communication Research, 30(56), 21. https://doi.org/10.7146/mediekultur.v30i56.8967

Faddoul, M. (2020) 'COVID-19 is triggering a massive experiment in algorithmic content moderation', *Brookings*, 28 April. Available at: https://www.brookings.edu/techstream/covid-19-is-triggering-a-massive-experiment-in-algorithmic-content-moderation/ (Accessed: 15 February 2021).

Flam, F. (2021, June 7). Facebook, YouTube Erred in Censoring Covid-19 'Misinformation'. Bloomberg.Com. https://www.bloomberg.com/opinion/articles/2021-06-07/facebook-youtube-erred-in-censoring-covid-19--misinformation

Frenkel, S., Decker, B. and Alba, D. (2020) 'How the "Plandemic" Movie and Its Falsehoods Spread Widely Online', *The New York Times*, 20 May. Available at: https://www.nytimes.com/2020/05/20/technology/plandemic-movie--youtube-facebook-coronavirus.html (Accessed: 21 November 2020).

Garcia Gonzalez, R., Amine, R. and M., S. (2021) *youtube--dl*. youtube-dl. Available at: http://ytdl-org.github.io/youtube-dl/about.html (Accessed: 22 February 2021).

Jameson, F. (1990). Cognitive Mapping. In L. Grossberg & C. Nelson (Eds.), *Marxism and the Interpretation of Culture* (pp. 347–360). University of Illinois Press.

de Keulenaar, E. *et al.* (Forthcoming) 'Authority and misinformation in the process of COVID sensemaking', in Rogers, R. and Niederer, S. (eds) *Mainstreaming the Fringe*. Amsterdam: Amsterdam University Press.

Knight, P. (2021) 'A Perfect Storm?', *Infodemic*, 19 January. Available at: http://infodemic.eu/2021/01/19/a-perfect--storm.html.

Knight, P. and Birchall, C. (2020) 'What are COVID-19 Conspiracy Theories?' Available at: http://infodemic.eu/2020/10/21/what-are-covid-19-conspiracy-theories.html.

Krug, G. J. (2016) 'Alternate Authenticities and 9/11: The Cultural Conditions Underlying Conspiracy Theories', in Williams, J. P. (ed.) *Authenticity in Culture, Self, and Society*. New York: Routledge, pp. 258–273.

de Laat, P. B. (2012) 'Coercion or empowerment? Moderation of content in Wikipedia as "essentially contested" bureaucratic rules', *Ethics and Information Technology*, 14(2), pp. 123–135. doi: 10.1007/s10676-012-9289-7.

Langlois, G. (2014) *Meaning in the Age of Social Media*. New York: Palgrave Macmillan.

Lewis, R. (2018) *Alternative Influence: Broadcasting the Reactionary Right on YouTube*. Data & Society Research Institute, p. 61. Available at: https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf.

Lind, D. (2017, August 12). Unite the Right, the violent white supremacist rally in Charlottesville, explained. Vox. https://www.vox.com/2017/8/12/16138246/charlottesville-nazi-rally-right-uva

m., davide. (2019, June 9). google api—ProcessingFailure error (400) while retrieving CommentThreads list. Stack Overflow. https://stackoverflow.com/questions/56516894/processingfailure-error-400-while--retrieving-committhreads-list

Maxmen, A., & Mallapaty, S. (2021). The COVID lab-leak hypothesis: What scientists do and don't know. Nature, 594(7863), 313–315. https://doi.org/10.1038/d41586-021-01529-3

Milajevs, D., Sadrzadeh, M. and Roelleke, T. (2015) 'IR meets NLP: On the Semantic Similarity between Subject--Verb-Object Phrases', in *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*. New York, NY, USA: Association for Computing Machinery (ICTIR '15), pp. 231–240. doi: 10.1145/2808194.2809448.

Moretti, F. (2013) *Distant reading*. London ; New York: Verso.

Myers West, S. (2018) 'Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms', *New Media & Society*, 20(11), pp. 4366–4383. doi: 10.1177/1461444818773059.

Ramos, J. (2003) 'Using TF-IDF to Determine Word Relevance in Document Queries', *ICML*, p. 4.

Rieder, B. (2018) 'Facebook's app review and how independent research just got a lot harder', *The Politics of Systems*. Available at: http://thepoliticsofsystems.net/2018/08/facebooks-app-review-and-how-independent-research--just-got-a-lot-harder/ (Accessed: 23 February 2021).

Rieder, B., Matamoros-Fernández, A. and Coromina, Ò. (2018) 'From ranking algorithms to "ranking cultures": Investigating the modulation of visibility in YouTube search results', *Convergence*, 24(1), pp. 50–68. doi: 10.1177/1354856517736982.

Rodriguez, D. (2020) *word2vec: Wrapper for Google word2vec*. Available at: https://github.com/danielfrg/word2vec (Accessed: 23 February 2021).

Roth, Y. and Pickels, N. (2020) *Updating our approach to misleading information*, *Twitter Blog*. Available at: https://blog.twitter.com/en_us/topics/product/2020/updating--our-approach-to-misleading-information (Accessed: 31 July 2021).

Sandler, R. (2020) 'YouTube Cracks Down On QAnon, But Doesn't Fully Ban It', *Forbes*, 15 October. Available at: https://www.forbes.com/sites/rachelsandler/2020/10/15/youtube-cracks-down-on-qanon-but-doesnt-fully-ban--it/ (Accessed: 23 February 2021).

Tangherlini, T. R. *et al.* (2020) 'An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, Pizzagate and storytelling on the web', *PLOS ONE*, 15(6), p. e0233879. doi: 10.1371/journal.pone.0233879.

The YouTube Team (2019) 'Continuing our work to improve recommendations on YouTube', *blog.youtube*, 25 January. Available at: https://blog.youtube/news-and--events/continuing-our-work-to-improve/ (Accessed: 21 November 2020).

Venturini, T. *et al.* (2018) 'A Field Guide to "Fake News" and Other Information Disorders', *SSRN*. Available at: http://fakenews.publicdatalab.org/ (Accessed: 25 June 2018).

Wattenberg, M. and Viegas, F. B. (2008) 'The Word Tree, an Interactive Visual Concordance', *IEEE Transactions on Visualization and Computer Graphics*, 14(6), pp. 1221–1228. doi: 10.1109/TVCG.2008.172.

Wayback Machine (2020) *Changes*. Internet Archive. Available at: https://web.archive.org/web/diff/20170118202526/20170120040337/https://www.ice.gov/speeches (Accessed: 2 September 2020).

YouTube (2019a) *Continuing our work to improve recommendations on YouTube*, *blog.youtube*. Available at: https://blog.youtube/news-and-events/continuing-our-work-to--improve/ (Accessed: 20 February 2021).

YouTube (2019b) *Our ongoing work to tackle hate*, *blog.youtube*. Available at: https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate/ (Accessed: 16 February 2021).

YouTube (2020) *COVID-19 Medical Misinformation Policy - YouTube Help*. Available at: https://support.google.com/youtube/answer/9891785?hl=en (Accessed: 2 September 2020).

YouTube (2021a) *Spam, deceptive practices & scams policies - YouTube Help*. Available at: https://support.google.com/youtube/answer/2801973?hl=en (Accessed: 2 September 2020).

YouTube (2021b) *The Four Rs of Responsibility, Part 2: Raising authoritative content and reducing borderline content and harmful misinformation*, *blog.youtube*. Available at: https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/ (Accessed: 19 February 2021).

de Zeeuw, D. *et al.* (2020) 'Tracing normiefication', *First Monday*.