# Preliminary Draft

July 23, 2021

Title proposition 1: "Big Tech is censoring me": using social media data to verify the platforms' regulation policies regarding misinformation.

Title proposition 2: Illustrating Facebook, Twitter and YouTube regulation policies against misinformation with a few chosen examples

Title proposition 3: Misinformation Policies of Mainstream Social Media Platforms: third-party monitoring, methods and illustration.

## Contents

# 1 Introduction

Section 230 in the United States Communications Decency Act provides immunity for website platforms against the content created by users. Similar regulations exist in the European Union via the E-commerce Directive (2000) in articles 12 and 15.[1] Nevertheless, there is growing pressure for mainstream social media platforms, such as Facebook, Twitter or YouTube, to moderate the available content. In particular, platforms take explicit actions when content is in violation of local laws in different jurisdictions, such as laws regarding defamation of a racial nature, dissemination of symbols from unconstitutional organizations, privacy protection, digital security, electoral laws. For example, Facebook reports having implemented a total of 64.7 thousand content restrictions based on local law across all countries in 2020.[2]

Furthermore, mainstream platforms are increasingly engaging in editorial tasks by implementing targeted policies to insure that each platform's rules are not violated. Community guidelines of Facebook, Twitter and YouTube can be summarized in a handful of categories, regarding safety, privacy and authenticity; which include sub-categories such as violence, terrorism, child sexual exploitation, abuse, harassment, hateful conduct, suicide or self-harm, illegal or regulated goods and services, platform manipulation and spam (see Appendix 6.1 for references). While specific to each platform, the previously cited categories correspond in most cases to well defined concepts that fall into legal frameworks in many countries, unlike misinformation. The intricacies of constructing a legal framework for misinformation arises from the difficulty of identifying and qualifying a piece of online content as false or misleading, among an overwhelming quantity of daily produced content, without infringing existing laws.[3] In particular, a number of recent studies point towards the idea that "Fake News" or disinformation is a small subset of the total supply of information on online social networking platforms (e.g. Grinberg et al. (2019) [5] and Broniatowski et al. (2020) [2]). Yet, this seemingly small subset is generating great concern in traditional media and in society in a broader sense. [4]

Hence, in the present article, we focus on mainstream platforms' policies and interventions regarding content with low credibility or false information, commonly referred to as *Fake News* (see Lazer et al. (2018) [7]). The *Fake News* phenomenon is still ill-defined by the academic community, as it encompasses several combined features such as spreading inaccurate, false or misleading information, with or without the intention of influencing or manipulating a target pool of audience. The growth of social networking platforms over the last decade in terms of number of users worldwide and volume of content, has modified the information ecosystem in terms of production of information and its mediation. Many users can now produce and share content which includes news related information, without having to abide by strict editorial processes that ensure accuracy of information and reliability of sources. In particular, false or inaccurate content produced and shared on social networking platforms concerning the political life or public health may have a potentially harmful impact on the society, in the rare event that it goes viral. This gave rise to

[1] See section 4 in Bayer (2019) [1] for a comprehensive overview of the mentioned articles.

[2] See Facebook Transparency Center, Content restrictions based on Local Law: transparency.fb.com/data/contentrestrictions. We summed the count of content restrictions over all countries reported in the table, for $H1$ and $H2$ of the year 2020.

[3] See A guide to anti-misinformation actions around the world on the website of Poynter Institute.

[4] For example see the February 2020 speech of the Director General of the WHO at the Munich Security Conference, where he says "But we're not just fighting an epidemic; we're fighting an infodemic." Furthermore, the European commission recognizes the spread of online disinformation as a problem and has put together in 2018 a Code of practice on Disinformation, which is a set of self-regulatory standards to fight disinformation.

a set of heterogenous fact-checking policies across mainstream platforms. For example, Facebook has a substantial partnership program with Fact-checking partners certified by the non-partisan International Fact-Checking Network. Facebook uses a number of signals and machine learning models to predict misinformation and surface it to fact-checkers.[5] Twitter seems to have a different approach where they focus on providing context rather than fact-checking[6] and the platform is testing a new system based on the wisdom of the crowds to tackle misinformation (see Twitter Birdwatch). As for YouTube, this platform utilizes the schema.org ClaimReview markup, where fact-checking articles created by eligible publishers can appear on information panels (see Appendix 6.1 for references).

During the COVID-19 global health pandemic platforms have upgraded their guidelines to include a set of rules to tackle the propagation of potentially harmful content (see Appendix 6.1 for references). Those policies are enforced via existing actions used by the platforms to tackle other rules' violations, such as: labelling content to provide more context or indicate falsehood, publishing a list of terms or topics that will be flagged, suspending accounts, implementing strike systems, reducing the visibility of content, etc. As each platform is a private company, those *new* policies are not coordinated and are implemented in different ways across platforms. Such targeted policies show the willingness of mainstream platforms to enhance the quality of the online conversation, but also sheds light on the lack of specific policies to tackle misinformation in general. In particular, policies regarding misinformation are not part of the set of platform rules or community guidelines (as of July 2021). The 2019 report of the Facebook Data Transparency Advisory Group (DTAG) states that "*DTAG was not tasked with evaluating any of the following: (...) Facebook's policies with respect to "fake news" or misinformation, as neither of these categories were counted as violations within the first two versions of the Community Standards Enforcement Report*".

Misinformation specific interventions by mainstream platforms are hard to monitor, study or verify by third parties (e.g. academic community, data journalists, NGOs), as in many cases misleading or false content does not qualify as a violation of a given platform's rules and it does not explicitly appear as a separate category in available transparency reports. This makes the study of online misinformation, the assessment of the impact of platforms' actions to tackle misinformation and their relevancy a burdensome task for the academic community. Hence, in the present article we explain how to verify with data mining mainstream platforms' current actions regarding content with low credibility or false information. We do so by providing a series of examples for different interventions and platforms. We chose to focus on three platforms: Facebook, Twitter and YouTube. Both Facebook and YouTube are in the top three most popular social media platforms in terms of number of users.[7] We further choose Twitter because it is a social networking platform with the most news-focused users, according to the Pew Research center (2019) [6]. To collect data from these three platforms, we either used the APIs (Application Programming Interfaces), or web scraping, i.e. retro-engineering the HMTL code of a web page to extract meaningful data, see Table 1 for a summary. Minet [10], a webmining tool developed by the Médialab SciencesPo, was often

[5]See the section Frequently asked questions: 'How does Facebook use technology to detect potential misinformation?"

[6]To the best of our knowledge, Twitter does not have a page which summarizes its fact-checking strategy. The Twitter Safety Team tweeted on June 3, 2020 the following: "We heard: 1. Twitter shouldn't determine the truthfulness of Tweets 2. Twitter should provide context to help people make up their own minds in cases where the substance of a Tweet is disputed. Hence, our focus is on providing context, not fact-checking." Tweet ID 1267986503721988096.

[7]See for example the ranking of the most popular social networks as of April 2021 on Statista: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

used, and we scripted our own data mining code when it was necessary.

| | Application Programming Interface (API) | Web Scraping |
|---|---|---|
| | CrowdTangle API and Buzzsumo API | Code created for this article |
| | Twitter API v2 | Minet tw scrape |
| | YouTube API v3 | Code created for this article |

**Table 1:** Summary of ressources used for data collection.

More specifically, we survey in this article a number of common policies against misinformation used by Facebook, Twitter and YouTube. We classified those common policies into three broad categories: (*i*) temporary or permanent suspension of users, (*ii*) introducing flags and notices, and (*iii*) reducing the visibility of some content. We compile in Appendix 6.1 in table 5, a list of links that redirect to the policies, regulations and transparency centers of Facebook, Twitter and YouTube that we discuss throughout the article. Furthermore, the examples provided to illustrate how to monitor a given policy were picked out of a list of domain names with several failed fact-checks. To be more precise, some domain names with failed fact-checks were picked because a given platform communicated about an intervention or because an intervention (e.g. suspension) was announced on the social media accounts linked to a given domain name. Finally, we discuss how an increased effort of transparency regarding specific content can help the community of researchers study and assess the impact of platforms' policies regarding misinformation.

# 2 Temporary & permanent suspension

Mainstream social media platforms may suspend the account of a specific user when they deem that the platforms' rules have been violated. Account suspension can be temporary or permanent. When the suspension is temporary the user is prohibited for a limited period of time from posting content on their account, but created content prior to suspension remains available to the user and their followers. However, when the suspension is permanent, in most cases, followers or subscribers have no longer access to the content prior to the suspension and the user can no longer use the account to create new content. In what follows, we focus on the implementation of this policy by several platforms and provide simple examples to illustrate.

## 2.1 Facebook

When an account is permanently suspended by Facebook, it disappears from the platform. That is, the data can no longer be scrapped and it also disappears from the CrowdTangle API.[8] Facebook publishes on monthly basis a *coordinated inauthentic behavior* report, where it informs how many personal accounts, pages or groups were deleted and to which *deceptive network* they may have belonged to.[9] But as long as external persons do not have access to deleted accounts data, these reports cannot be verified by independent researchers or journalists.

[8]CrowdTangle is a public insights tool owned and operated by Facebook, that exclusively tracks public content from Facebook public groups and pages.

[9]See the April 2021 report as an example.

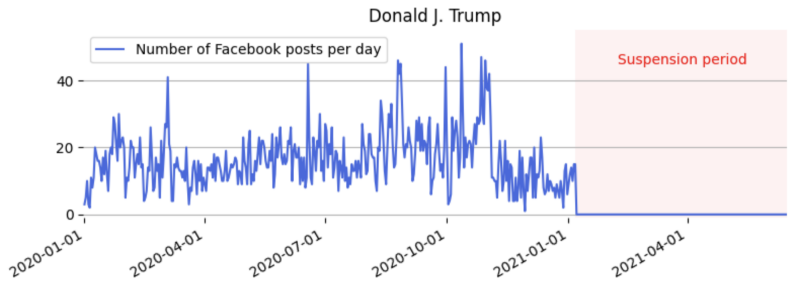


**Figure 1:** Number of Facebook posts published each day by the Facebook page *Donald J. Trump* between January 1, 2020 and June 15, 2021. The data corresponds to 6 083 posts retrieved from the CrowdTangle API using the *posts* endpoint.

Facebook can also apply a temporary suspension, and in this case the data can often be collected and analyzed. For example, Donald Trump's official Facebook page has been suspended following the Capitol attack on January 6, 2021.[10] Nevertheless the page's data is still present in the CrowdTangle API. Thus, after manually adding this page to the CrowdTangle dashboard, we collected the 6 083 posts it had published between January 1, 2020 and June 15, 2021 using the *posts* endpoint.[11] We used Minet command line tool [10] to collect the data. We can verify on figure 1 that the *Donald J. Trump* page has not published any content since January 6, 2021, and that this behavior is not consistent with the page's previous activity: an average of 16 posts were published each day on Facebook before the suspension.

## 2.2 Twitter

Twitter has implemented a strike system as part of their Civic Integrity Policy and their COVID-19 misleading information policy. Violations of both policies can entail strikes, where two strikes lead to a 12-hour account lock and five or more strikes lead to permanent suspension from the platform. A list of notable Twitter temporary and permanent suspensions can be found on Wikipedia. The 12-hour account lock is hard to observe in the data, especially for users who do not have an over the clock tweeting activity. In this section, we provide one example of a temporary suspension of a Twitter account, that seems to be the result of a manual decision concerning a Tweet which violated the rules.

The Twitter account $@LifeSite$ of the website lifesitenews.com has been suspended for at least two periods of time: from end of 2019 until fall 2020 for 308 days, then again since January 2021 for having violated Twitter Rules[12]. In particular, this website has several failed fact-checks concerning the published articles, according to Iffy.news.[13]. We collected the activity (tweets, replies, quotes, retweets) on their Twitter account via the Twitter API, using the historical search endpoint. We then plotted the number of Tweets, Retweets, Quotes and Replies per day, as shown in panel *a* of

[10] See https://www.facebook.com/zuck/posts/10112681480907401

[11] See the endpoint documentation for more details: https://github.com/CrowdTangle/API/wiki/Posts.

[12] See Lifesitenews's article discussing the reason for the suspension: https://www.lifesitenews.com/news/lifesite-is-dumping-twitter-and-so-should-you. Twitter rules can be found at: https://help.twitter.com/en/rules-and-policies/twitter-rules.

[13] See https://mediabiasfactcheck.com/life-site-news/

figure 2). The two periods of temporary suspension are clearly observed in the data as the user(s) of the account were not allowed to use the functionalities of the Twitter Platform.

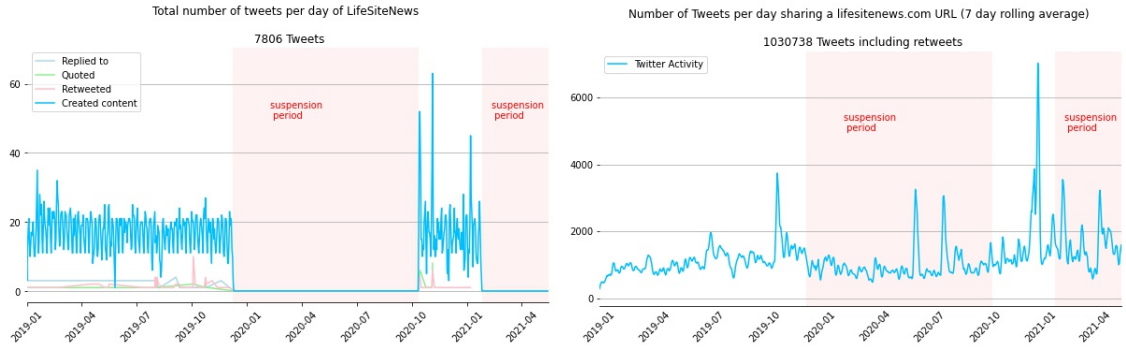


**Figure 2:** Left Panel: number of Tweets per day of the Twitter account @*Lifesite* linked to the website lifesitenews.com from January, 2019 until April 2021. Right Panel: number of Tweets per day that have shared a lifesitenews.com URL link from January, 2019 until April 2021.

To further assess the impact of this double temporary suspension, we collect via Minet Command Line Tool [10], all the tweets that have shared during the same period a url link containing lifesitenews.com. Panel (*b*) of figure 2, shows that during both periods of temporary suspension, other users still shared lifesitenews.com links and that the level was only slightly below the tweeting and retweeting levels prior to the first temporary suspension. More specifically, there was an average of 960 tweets (including retweets) per day over the first temporary suspension period of 308 days from December 9, 2019 until October 12, 2020, against an average of 977 tweets (including retweets) per day during the exact same period one year earlier. Finally, panel (*b*) points towards the limitations of suspending an account to limit the spread of its content.

## 2.3 YouTube

In this section, we turn to the channel's temporary or permanent suspension policy of YouTube. Whenever a channel publishes a video that violates the community guidelines for the first time they will usually receive a warning and the content will be removed. For the second time the channel will start receiving strikes. A first strike results in limiting the access of the YouTube channel for one week, like uploading videos, streaming and other activities. Then a second strike is similar but the suspension will be for two weeks. A third strike results in the termination of the channel. The strike count of a channel lasts 90 days. In the special case, where a video is in extreme violation of the guidelines, the publishing channel may get terminated without a warning.

To illustrate the implementation of this policy we provide two examples for the temporary suspension of the following two YouTube channels: One America news Network and Tony Heller. The website of One America News Network has a "low" factual reporting score according to iffy.news.[14] Tony Heller posts regularly blog posts on the website *realclimatescience.com*, which also has a "low" factual reporting score according to iffy.news.[15] Both are active on YouTube and both have

[14]See mediabiasfactcheck.com/one-america-news-network/.
[15]See mediabiasfactcheck.com/real-climate-science/

communicated via their Twitter account about restrictions applied by YouTube over their content (see figure 3).


**Figure 3:** Left Panel: Tweet announcing moving to rumble by OANN (Twitter), Twitter ID 1372238828425998336. Right Panel: Tony Heller's tweet after getting suspended from YouTube, Twitter ID 1310703852769796097.

First, we investigate the temporary suspension of the YouTube channel of *One America News channel*. This channel received a first strike on November 24, 2020 for the promotion of a false cure for COVID-19 according to the News outlet NBCnews (2020) [3]. We collected the activity of the channel OANN (video counts, view counts) using the YouTube API v3, between November 2020 and January 2021. For the video counts, we used the playlist endpoint to retrieve the videos uploaded with their publishing date and for the view count we used the IDs of the videos we had from the playlists and via the videos endpoint we retrieved the view counts on June 2021.

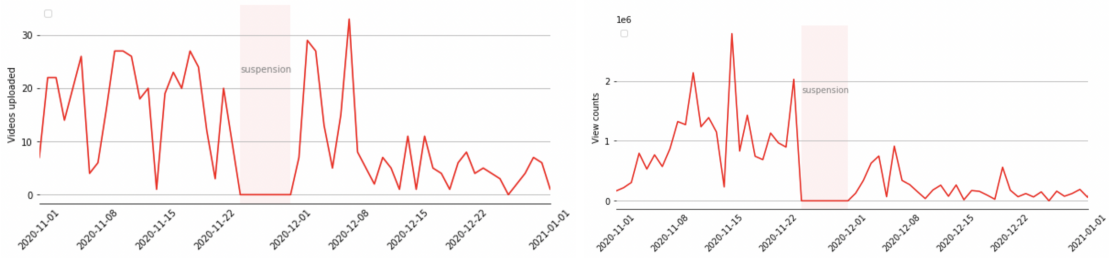


**Figure 4:** Left Panel: Number of YouTube videos uploaded each day by the youtube channel *One America news Network* November 1, 2020 and January 1, 2021. Right Panel: accumulated view counts for videos. The metrics correspond to the videos' publishing date and the data is retrieved from the youtube API with the *playlists* and *videos* endpoints.

In addition, as shown in figure 4 when comparing the month before the suspension from 2020/10/24 to 2020/11/24 and one month after from 2020/12/01 to 2021/01/01 it was found that the view count decreased by -73% and the videos uploaded by -55%. Besides that, OANN decided to move officially to Rumble on March 17, 2021 as announced on their Twitter account (see figure 3) and their upload activity on their YouTube channel is close to zero since that announcement.

We now turn to our second example, the temporary suspension of the YouTube channel Tony Heller. This channel got its first strike after posting a video about an anti-covid-lockdown doctor getting arrested (see screenshot in figure 3). The suspension period was for one week from September 29 until October 5. We applied the same methods as in the previous example for the data collection.

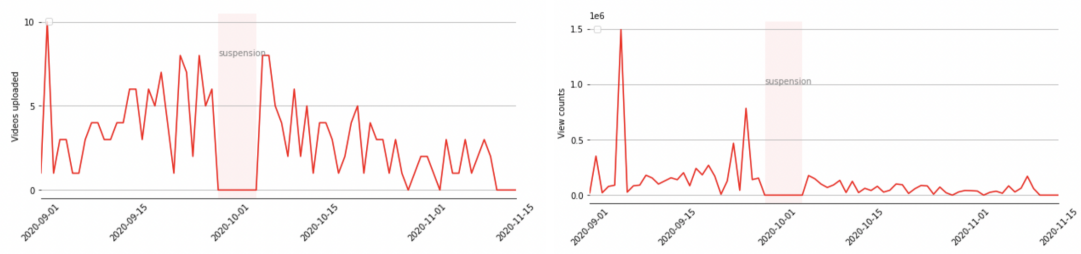


**Figure 5:** Left Panel: Number of YouTube videos uploaded each day by the YouTube channel *Tony Heller* between September 1, 2020 and November 15, 2020. Right Panel: accumulated view counts for videos uploaded by the same YouTube channel. The date corresponds to the videos' publishing date.

Figure 3 shows the daily number of videos uploaded by the channel. The suspension period can be observed clearly in the historical data of the channel. observing the reach of the audience Figure 3: Tony heller tweet after getting suspended from YouTube (Twitter) using view counts one month before the suspension starting from 2020/08/28 to 2020/09/28 and one month after the suspension from 2020/10/05 to 2020/11/05 the channel witnessed a drop of view counts by $-69.5\%$ and the videos published in the channel were less by $-29\%$. This drop in views can show that the suspension period may have a good impact on reducing the audience interest or reach to the channel.

# 3 Flags, Notices and labels

## 3.1 Facebook

To the best of our knowledge, two types of flags can currently be seen on Facebook posts, videos and pictures: $(i)$ information banners that do not refute the message in the post but that give a link to an authoritative source, such as "Visit the COVID-19 Information Centre for vaccine resources" and $(ii)$ fact-check flags that provide a "judgement" on the text of the post or the link, shared, such as "False information Checked by independent fact-checkers" (see Figure 6). The fact-check flags can be of varied nature : 'False', 'Partly false', 'Missing context', 'False headline', 'Altered media', 'Opinion', 'Satire', 'Not eligible' and even 'True'.

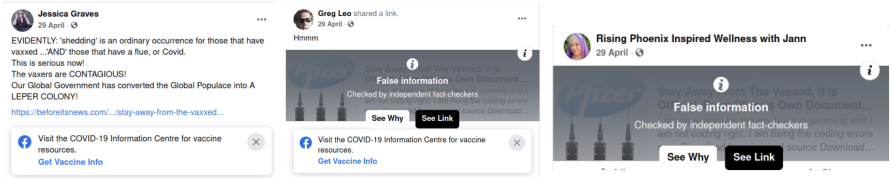


**Figure 6:** Examples of Facebook posts having shared a link fact-checked as False by one of Facebook's partners. Screenshots taken on July 8, 2021: facebook.com/groups/1220117708132394/permalink/2386760061468147, facebook.com/groups/473809623000471/permalink/1333200093728082, facebook.com/691911990845585/posts/3835629366473816).

No information regarding the flags can be found on Buzzsumo or CrowdTangle, the two APIs we use to access Facebook data. The only way to verify Facebook's flag policy is thus to scrap Facebook. For this paper, we added a new feature in minet to scrape Facebook posts and automatically verify

for the presence of the flags.

We first searched for all the Facebook posts having shared a link rated as 'False' by Science Feedback with the *search* endpoint in minet[16]. Twenty Facebook posts were collected this way, and the newly developed scraper was used to verify whether they were flagged with an information banner, a fact-check flag, both or none. Three posts were unavailable, and thus could not be categorized by the scraper. As Science Feedback has sent a False fact-check for this link to Facebook, we would expect all these posts to contain a fact-check flag, but we had no expectations for the information banner.

| Number of posts | | | |
|---|---|---|---|
| without a flag | with an information flag | with a fact-check flag | with a fact-check and an information flags |
| 0 | 7 | 4 | 7 |

**Table 2:** Count for the Facebook posts with the different types of flags having shared a link fact-checked False by one of Facebook's partners

Surprisingly only 11 posts out of the remaining 17 had a 'False information' flag (Table 2, see left panel of Figure 6 for an example). We observed that the flagged posts were also the ones in which the false link was expanded (i.e., a banner was visible with an image of the link and that can be clicked on, see the middle and right panels of Figure 6 for examples). As the 'False information' flag is applied on the link banner, and not on the link itself, a user is thus able to share a False link on Facebook without the 'False information' banner if the link is not expanded.

We also observed that most of the posts sharing this False link (13 out of 17, Table 2) had an information banner saying 'Visit the COVID-19 Information Centre for vaccine resources' (see the first two examples of Figure 6). We could not identify why the banner was applied on some posts and not on others (we saw no clear difference in the post messages for example). It should be noted that some posts displayed both the information and the fact-check banners, as in the middle panel of Figure 6.

## 3.2 Twitter

Alongside other social networking platforms, when the content of a tweet violates the Twitter rules, a notice can be added to provide more context according to Twitter's Help Center. At the tweet level, notices take the form of a label or an interstitial. Labels are context specific (e.g. COVID-19 or presidential elections) and redirect users to a webpage to get more context, for example *Get the facts about COVID-19'* (see right panel of figure 7). Interstitials are presented as a greyed box on top of a tweet, which indicate sensitive content, violations of Twitter rules, withheld tweets for violation of local laws or even tweets from suspended accounts, for example *The following media includes potentially sensitive content* (see left panel of figure 8). At the account level, notices can also indicate whether an account has been temporarily or permanently suspended.

To the best of our knowledge, when using the Twitter API v2, there is no field which indicates whether a tweet is labeled or not; while the interstitial "possibly sensitive" and "withheld" are both Tweet fields that can be recovered[17] from the API. Hence, in order to investigate the presence

[16]Click here to access the link and its fact-check by Health Feedback.

[17]See https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/tweet.

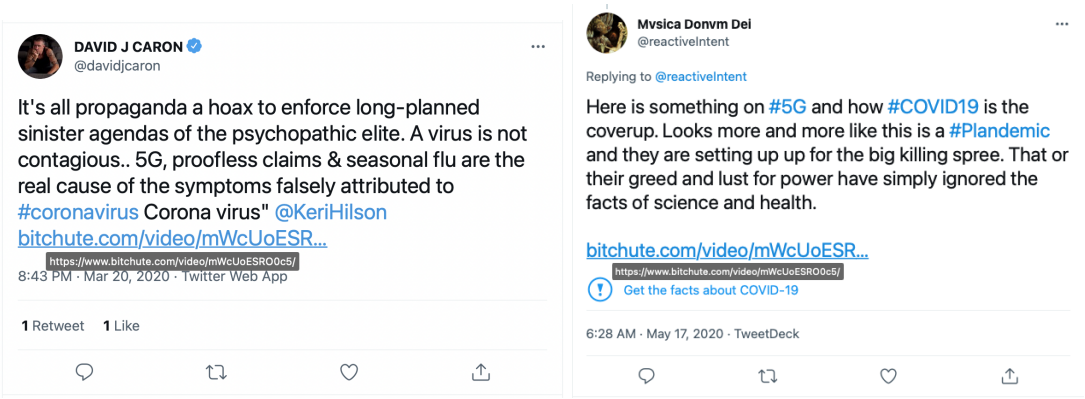


**Figure 7:** Two tweets sharing the same link marked as False by Science Feedback, screenshots taken on June 20, 2021. Left Panel: Tweet ID 1241088065462026242 without a label. Right Panel: Tweet ID 1261876171584745472 containing a label.

of labels of all types, we resorted to scrape Twitter data using Minet [5]. This tool was recently enriched upon our request in order to capture whether a tweet contains a label or not, via the *minet twitter scrape* command.

In this section, we take a deeper look at how labels and notices are introduced by Twitter, to indicate content which is inaccurate or false. To that end, we gathered a set of 3094 links redirecting to articles which were marked as $False$ between April 2019 and February 2021 by Science Feedback, a fact-checking organization verifying the credibility of science-related viral information. As a second step, we collected (on June 30, 2021) via Minet Command line tool [10] all the tweets that have shared a link which belongs to the set of 3094 links marked as $False$. This data collection resulted in 323 938 tweets, excluding retweets. Only 28 tweets contained the label *Get the facts about COVID-19*, 5 tweets contained the label *Learn about US 2020 election security efforts* and only 1 had the following label *This claim about election fraud is disputed.* Furthermore, we noticed that the labeling rule might not be applied uniformly on a given set of tweets sharing the exact same link, among the set of collected tweets. To give an example, 657 tweets had shared the exact same link redirecting to a video on Bitchute, entitled *Important information on coronavirus 5G Kung Flu.* Among those 657 tweets only 3 contained the label *Get the facts about COVID-19* (see figure 7). This points towards the non-automation of the tweet labelling process and that it might be that these 3 tweets were the only ones reported by other users.

We further examine the placement of interstitials that indicate a possibly sensitive content. We find that only 9344 (2.97%) tweets out of 323 938 containing a link marked as False, have an interstitial *potentially sensitive content.* To check whether the interstitial *potentially sensitive content* is automated or not, we pick out one tweet having shared a link marked as False[18], which also contains the interstitial *potentially sensitive content* (see figure 8). Then we look in the set of 323 938 tweets for other tweets who have shared the exact same link. We find a total of 32 tweets. Among those 32 tweets only 5 tweets had the interstitial "potentially sensitive content". Again this points towards the non-automation of interstitials placement. Furthermore, notice that the

[18]The link marked as False redirects to the article *5G Technology and induction of coronavirus in skin cells – US National Library of Medicine (what David Icke has been saying for months)* and can be found here.


**Figure 8:** Left Panel: Tweet containing a URL link hidden behind an interstitial. Right Panel: when clicking on *view*, to view the content hidden behind the interstitial. Tweet ID 1285866521533861888. Screenshots taken on July 1, 2021.

appearance of interstitials may depend on the settings of a user's account and country or language specific regulations. In particular, in the settings of Twitter account, one can deactivate the display of interstitials for *sensitive content*, by ticking the box *Display media that may contain sensitive content* in the section *content you see*.

Finally, the number of tweets containing the interstitial *potentially sensitive content* (9344 tweets) widely outnumbers the tweets which contain a flag (32 tweets), among the set of 323 938 tweets containing a link marked as False (see table 3). Furthermore, we noticed that among the tweets which contain the interstitial *potentially sensitive content*, there are tweets which contain links redirecting to misleading articles about COVID-19 and yet they are not marked by the flag *Get the facts about COVID-19*. We also noticed that the 9344 tweets marked with the interstitial *potentially sensitive content* do not necessarily contain graphic violent[19] content but rather links that redirect to articles marked as *False* by Science Feedback. When navigating through the Twitter Help Center, we found no explanation for this large discrepancy between using the interstitial *potentially sensitive content* and flags which provide more context to users.[20]

## 3.3 YouTube

YouTube may provide an information panel for videos with topics that are prone to misinformation like COVID-19, moon landing, and climate change. Information panels provide ressources concerning a potentially controversial topic from independent third parties or authoritative ressources. YouTube states that these panels exist regardless of the point of view expressed in a given video. In addition, these panels are not yet available in all languages and countries.

[19]The Twitter Help center explains the placement of interstitials of possibly sensitive content as follows: "We may place some forms of sensitive media like adult content or graphic violence behind an interstitial advising viewers to be aware that they will see sensitive media if they click through".

[20]Twitter Safety account announced in a Tweet (see Tweet ID : 1379515615954620418) on April 6, 2021 that their team "will begin deploying automated tools to build on (their) efforts to label tweets that may contain misleading information around COVID-19 vaccinations".

| 323 938 Tweets containing a link marked as False | | | |
|---|---|---|---|
| with the interstitial "potentially sensitive content" | with the flag "Get the facts about COVID-19" | with the flag "Learn about US 2020 election security efforts" | with the flag "This claim about election fraud is disputed" |
| 9344 ( 2.88450 %) | 28 (0.00864%) | 5 (0.00154%) | 1 (0.00030%) |

**Table 3:** Summary of the number of tweets among the set of collected tweets containing a link marked as False by Science Feedback, which contain a flag or an interstitial.

To further study the assignment of information panels to videos, we compiled a list of 171 YouTube videos, fact-checked by Science Feedback between April 2019 and March 2021 and marked as containing false or inaccurate information. We collected the information panels, when present, in each video by connecting to a US server and scraping the content of the web page. Four types of information panels were found in the list of visited videos as shown in figure 9: information about Climate change, COVID-19 Vaccine, COVID-19 and US elections.[21]

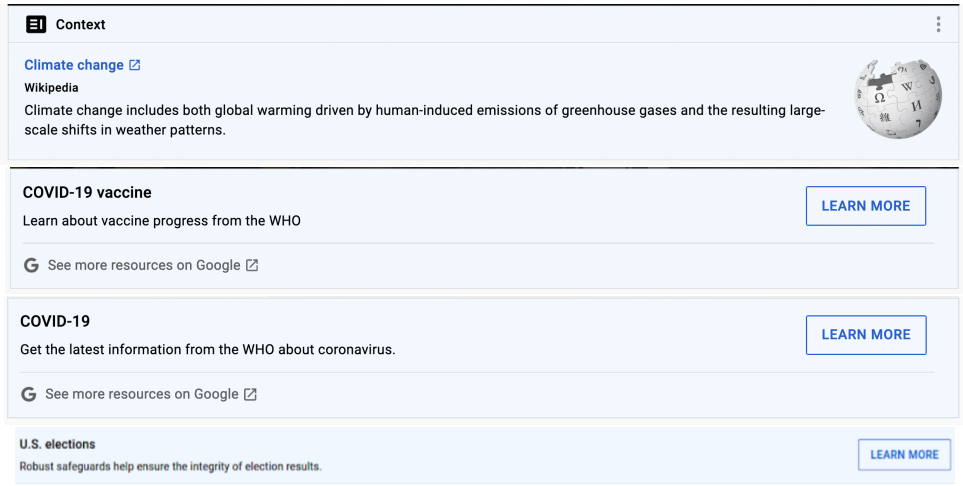


**Figure 9:** Screenshots taken on July 21, 2021. Examples of YouTube information panels displayed under Youtube videos respectively about: Climate change, COVID-19 Vaccine, COVID-19 and US elections.

We classified the list of 171 YouTube videos by topic based on its content as shown in table 4. We recorded the number of videos containing an information panel within each category. For COVID-19 and COVID-19 vaccine related videos, more than half of the videos contained an information panel. Nevertheless we noticed that among a set of duplicates of the exact same YouTube video uploaded with different video titles, some contained an information panel while others did not (see figure 10 for an example). In addition, we noticed for COVID-19 related videos, when the video title did not include keywords like (Testing, Pandemic, COVID-19, coronavirus), the video might

[21] See: youtube.com/watch?v=OwqIy8Ikv-c, youtube.com/watch?v=9sW0OmzcmL0, youtube.com/watch?v=usyQgPU-VrI and youtube.com/watch?v=8PAwc8nlE_Q. Last accessed on 21/07/2021.

not contain a panel associated with it, and in some cases when the video title contains variations of word COVID like (C O V I D 19 or Cv19) it wouldn't include a panel either. Therefore, we suspect that YouTube is automatically adding panels to the videos based on the video title and not the content of the video.

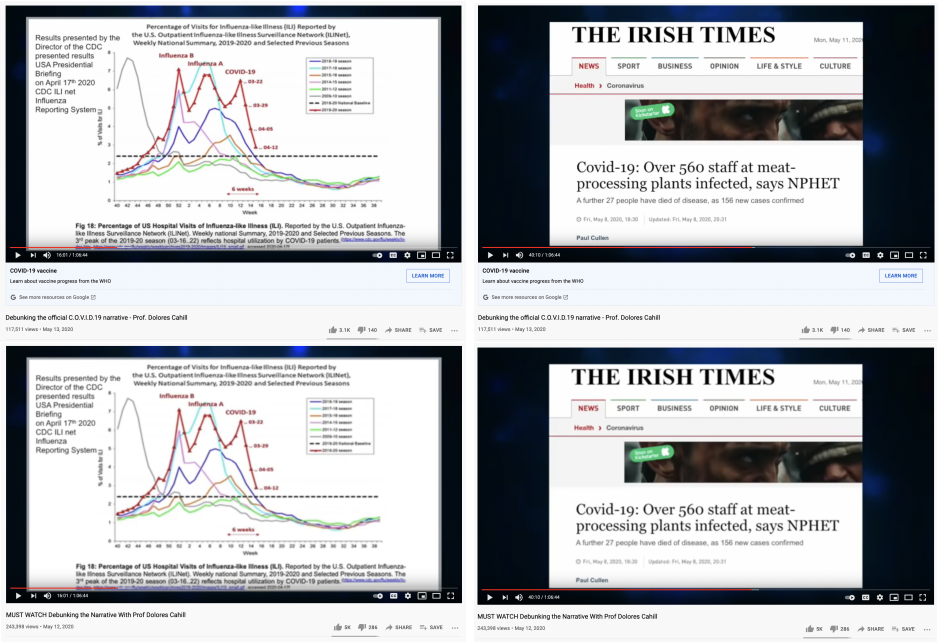


**Figure 10:** Screenshots taken on July 21, 2021. The screenshots show two duplicates of the same YouTube video uploaded under different titles. Top panels: a YouTube video about COVID-19 which contains an information panel, see https://www.youtube.com/watch?v=d9GbVZOcT18. Bottom panels: a duplicate of the YouTube video in the top panels uploaded under a different title and it does not contain an information panel, see https://www.youtube.com/watch?v=A4RvrEKoNxc.

For climate change we found that most of the videos (13 out of 18) did not display an information panel. Lastly, no information panels were found in the general health category set, that included videos that were not related to COVID-19 but contained misinformation concerning cures for cancer, abortion, and viruses. Therefore, it is likely that YouTube add information panels below videos concerning controversial topics that can have misleading information like the climate change, COVID-19, flat earth and vaccine. It appears as if YouTube is focusing its efforts on the current "hot" topics with a wide public impact (elections, COVID), leaving room for misinformation to spread concerning other topics.

# 4 Reducing the visibility

Mainstream social media platforms can reduce the visibility of the content created or shared by specific users, whenever they violate the platforms' rules. The implementation of this policy varies across platforms and is not easy to verify ex-post. In what follows we provide means to verify this policy on Twitter and Facebook.

| Category | Number of YouTube videos **with** an information panel | Number of YouTube videos **without** an information panel |
|---|---|---|
| Climate change | 5 (28%) | 13 (72%) |
| COVID-19 Vaccine | 21 (75%) | 7 (25%) |
| COVID-19 | 73 (63%) | 43 (37%) |
| General Health | 0 (0%) | 11 (100%) |

**Table 4**

## 4.1 Facebook

One of Facebook's measures to regulate misinformation is to reduce the spread of misleading content through their ranking system. Facebook ranks each post and/or ad by assigning to it a relevancy score, where a high score leads to a high likelihood of the post and/or the ad to appear on a user's newsfeed. Doing so, Facebook can make a post or a whole account less visible by decreasing the relevancy score of its content; this is precisely the *reduce* measure (see the article on Facebook Newsroom (2018) [8]). This measure can be verified by looking at the number of views (reach) of a post, but this metric is not available via the APIs used to access Facebook data: CrowdTangle or Buzzsumo. Hence we can indirectly investigate the *reduce* measure by looking at the engagement metrics (likes, comments, shares) related to a given post; which are available on CrowdTangle and Buzzsumo. If a post reaches less users because it has a lower ranking, then it is less likely to receive likes, comments and shares, relative to a post with a higher ranking.

To illustrate, we investigate the case of the website $Infowars$. This website appears in the Misinformation Directory of FactCheck.org, among other websites who have posted deceptive content[22]. Furthermore, the factual reporting of $Infowars$ has been rated *very low* by the Media Bias / Fact Check resource of Iffy.news and it has several failed fact-checks reported by Feedback.org.[23]

On May 2, 2019, Facebook announced they would prohibit users from sharing Infowars content unless, they are explicitly condemning the material.[24] To verify the measure, we used the "/posts/search" endpoint[25] of the CrowdTangle API, to collect 37 242 Facebook public posts that had shared a URL link containing "infowars.com", published between January 1, 2019 and December 31, 2020.[26]

The number of public posts sharing an $Infowars$ link remained globally stable throughout 2019 (see figure 11 top panel). Thus the measure announced by Facebook doesn't seem to have prevented users from sharing an $Infowars$ link. Nevertheless, a clear drop in engagement was observed on

[22]See https://www.factcheck.org/2017/07/websites-post-fake-satirical-stories/.

[23]See the Iffy.news page https://mediabiasfactcheck.com/infowars-alex-jones/ and Feedback.org https://open.feedback.org/media/RM.

[24]See the article in Wired by Martineau (2020) [9] and see the section *So what happened with InfoWars? They were up on Friday and now they are down?* about.fb.com/news/2018/08/enforcing-our-community-standards/.

[25]see the documentation for more details: github.com/CrowdTangle/API/wiki/Search.

[26]We found in the collected data some Facebook posts that did not directly share an Infowars link (but rather a YouTube or Facebook video containing an Infowars link in its description), thus we excluded such posts from our data to keep only the 27 721 posts directly sharing an Infowars link.

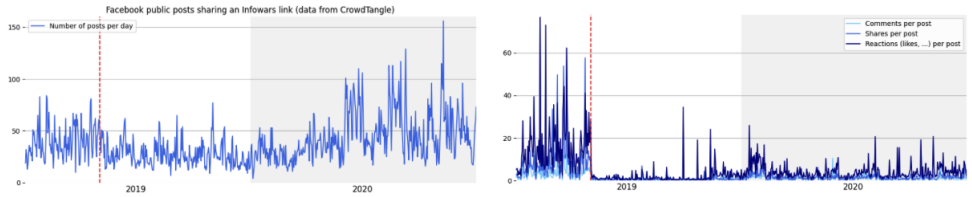


**Figure 11:** Public Facebook posts sharing an Infowars link in 2019 and 2020 collected from the CrowdTangle API. The red line marks the date of May 2, 2019, when Facebook has announced the ban regarding Infowars. (Top panel) Number of daily posts. (Bottom panel) Engagement metrics: average number of reactions, shares and comments per post.

May 2, 2019 (see figure 11 bottom panel). The number of reactions, shares and comments per post have decreased respectively by $-94\%$, $-96\%$ and $-93\%$ when comparing the two months before and after May 2, 2019. This suggests the measure taken by Facebook in May 2019, is not a ban per se, but rather a *reduce* measure. This is because users could still post $Infowars$ links, but these posts generated less engagement. It should be noted that the engagement metrics increased again by the end of 2019 / beginning of 2020, suggesting that the *reduce* measure may have been lifted a few months after its implementation.

Another measure that platforms can apply is to prevent users from sharing specific types of content, in this example URLs coming from a specific domain name. The Beauty of life (thebl.com/) is a US-based media company that shares pro-Trump views and conspiracy theories such as QAnon.[27] Facebook has announced on December 20, 2019 that *The BL is now banned from Facebook* for coordinated inauthentic behavior (see Facebook Newsroom (2019) [4]), which includes using fake accounts that misrepresent one's identity or using methods to artificially boost the popularity of content. Coordinated inauthentic behavior is a distinct phenomenon from disinformation according to Facebook, as "most of the content shared by coordinated manipulation campaigns isn't probably false".[28] Nevertheless, for the case of the Beauty of Life, both misinformation and coordinated inauthentic behavior were attested, according to the fact-checking organization Snopes which had reported about The BL's activity to Facebook[29].

To verify Facebook's ban of The BL domain name, we first tested whether we could post a Facebook message containing a url from thebl.com. This turned out to be impossible. But such manual verification cannot inform us whether the ban applies indeed to all Facebook users and accounts (as we used only our own personal accounts), nor when it has started. To further investigate this policy, we collected data from the Buzzsumo API[30]. We used the "/search/articles" endpoint to collect the engagement metrics of the 13 634 articles crawled from the *thebl.com* website between January 1, 2019 and June 15, 2021.

The number of Facebook reactions, shares and comments dropped to zero for TheBL's articles published after December 1, 2019 (see figure 12 top panel), indicating the start of the ban. We can note that although the ban was communicated in an article [4] published on December 20, 2019, it seems to have actually started on December 1, 2019. The communication around the ban appeared

[27] See wikipedia article.

[28] See: https://about.fb.com/news/2019/10/inauthentic-behavior-policy-update/.

[29] https://www.snopes.com/news/2019/11/12/bl-fake-profiles/, https://www.snopes.com/news/2019/11/12/bl-fake-profiles/, https://www.snopes.com/news/2019/12/13/facebook-bl-cib/

[30] BuzzSumo is a commercial content database that tracks the volume of user interactions with internet content on Facebook, Twitter, and other social media platforms.

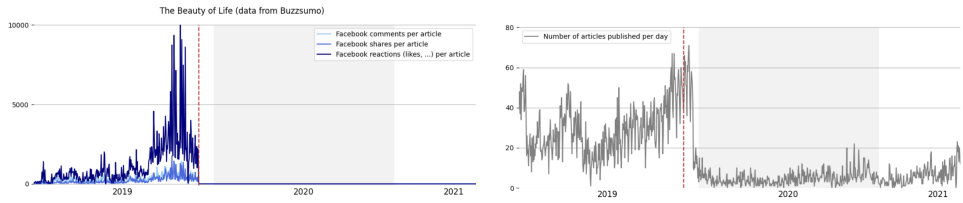


**Figure 12:** Articles from The Beauty of Life website (thebl.com) published between January 1, 2019 and June 15, 2021 and gathered from the Buzzsumo API. Left panel: Facebook engagement metrics (average number of reactions, shares and comments per article). Right panel: Number of articles published per day. The red line marks the date of December 1, 2019.

to have discouraged The Beauty of Life to proceed with their activity. Indeed the number of articles they published daily was around 50 until December 20, 2019, when it decreased drastically to reach around 5 to 10 articles published per day (see figure 12 bottom panel).

Using Buzzsumo data, we ascertained that links from thebl.com were not shared on Facebook anymore. The ban started on December 1, 2019, and appeared to be still enforced in June 2021.

## 4.2 Twitter

Twitter can take action against a tweet which violates the Twitter rules, by limiting its visibility[31] on users' timelines and in search results. To illustrate we provide an example for the website *globalresearch.ca*, which has several failed fact-checks according to Feedback.org and iffy.news - two websites which provides databases of websites with low factual reporting levels.[32]

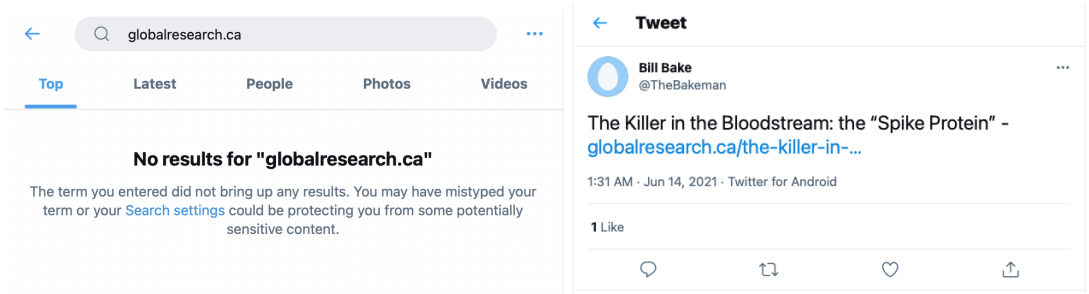


**Figure 13:** Screenshots taken on June 14, 2021. Left Panel: screenshot that shows that no results can be found when searching for *globalresearch.ca* via the Twitter search box. Right Panel: example of a tweet posted on June 14 containing the domain name *globalresearch.ca*.

The website *globalresearch.ca* is linked to the Twitter account *@CRG_CRM*. When a user searches via the twitter search-box for this domain name, no results appear as shown in the screenshot in the left panel of figure 13, taken on June 14, 2021. To further investigate the possible

[31] See the paragraph *Limiting Tweet visibility*: https://help.twitter.com/en/rules-and-policies/enforcement-options.

[32] For *globalresearch.ca* see open.feedback.org/media/AEKA and https://mediabiasfactcheck.com/global-research/.

implementation of a reduced visibility measure, we search via the Twitter API for tweets, excluding retweets, containing the query *globalresearch.ca* from January 1, 2021 until June 10, 2021. As shown in the left panel in figure 14, we find a strictly positive number of tweets containing the domain name *globalresearch.ca* throughout May 2021 and the first week of June 2021. Hence, the visibility of tweets containing this domain name has been reduced because users can no longer access tweets containing the domain name *globalresearch.ca* via the search box. Nevertheless users are not restrained from posting tweets containing this domain name, as shown in the screenshot in panel (*b*) of figure 13, found by taking the tweet ID of one of the collected tweets via the Twitter API. Furthermore, those collected Tweets have strictly positive engagement metrics as shown in panel (*b*) of figure 14. Hence, the users who tweet articles from the *globalresearch.ca* website receive tweet level engagement from their own followers.

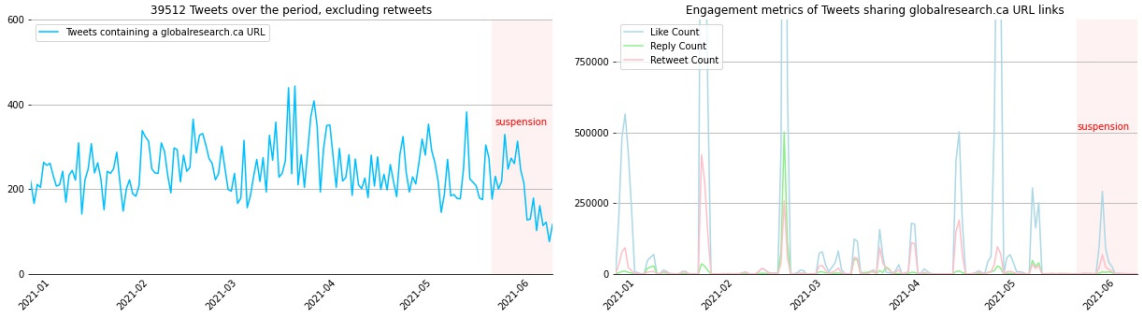


**Figure 14:** Left Panel: daily number of Tweets, excluding retweets, containing the query *globalresearch.ca* from January 1, 2021 until June 10, 2021. Right Panel: engagement metrics of Tweets containing the query *globalresearch.ca* from January 1, 2021 until June 10, 2021. Data collected via the Twitter API v2 on June 16, 2021.

Finally, the above mentioned account was recently suspended from Twitter. We noticed the message about the account suspension (see the right panel in figure 15) on May 25, 2021. To the best of our knowledge, no official communication by Twitter has announced the suspension nor the exact date at which it was implemented. Hence the account may have gotten suspended before that date. Furthermore when a user attempts to click on a link which contains *globalresearch.ca* in a Tweet, a warning message appears and indicates that the link may be unsafe (see the left panel in figure 15). For the case of this Twitter account, it is likely that a mix of interventions was used: reducing the visibility via the search box, suspending the related Twitter account and returning a warning message when users click on a link containing this domain name.

### 4.3 YouTube

What about something about recommendations ? authoritative content

## 5 Discussion

Sporadic points for the discussion, no structure yet.

- We do not provide an exhaustive list of methods on how to investigate the platforms' policies, as other APIs and some databases could be useful for that matter. We rather provide

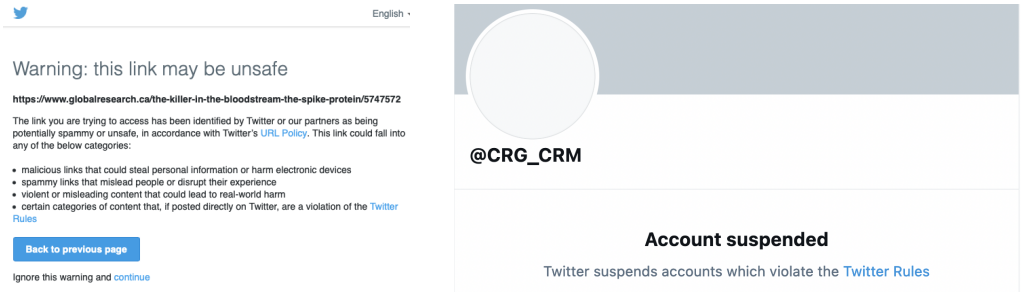


**Figure 15:** Screenshots taken on June 14, 2021. Left Panel: warning message that appears when a user attempts to click on a link which contains the domain name *globalresearch.ca*. Right Panel: account suspension message of @*CRG_CRM* linked to *globalresearch.ca*.

a methodology to investigate key policies, that can be useful to researchers or journalists interested in implementing external monitoring. For a detailed list of interventions of multiple platforms we invite the reader to check the following airtable compiled by Slatz and Leibowicz (2021) [11].

- For a previous research project (add reference to the article) we searched on CrowdTangle for public accounts sharing specific content associated with misinformation in November 2020, and selected 94 Facebook pages corresponding to our criteria. We then tried to collect these pages' posts in January 2021, and discovered that 11 pages could not be found anymore. This highlights an important issue when studying misinformation trends on Facebook: some data disappears from the CrowdTangle API as accounts are deleted or changed to *private*.
- To facilitate the verification in the policy applications, we would generally recommend for the platforms to be more transparent. But too much transparency on how the regulation policies are exactly implemented can actually backfire. For example YouTube is certainly applying an 'information' banner on all videos mentioning COVID-19 and related terms in their title. Misinformation accounts are trying to avoid the official banners by using terms as 'C.O.V.I.D' or 'C O V I D'. If YouTube was totally transparent on that matter and published the list of 'dangerous' words that leads to an information banner, this list would of course help us to understand YouTube's policies but it would also help the misinformation actors to escape the regulation. There is thus a balance between communicating enough so the public can know precisely how the platforms are regulating their content, but without giving too much information that would allow the policies to be bypassed.
- There are other ways to collect data from platforms, and besides Buzzsumo, other API are also aggregating data from multiple social platforms. For example Newsguard, blablabla... In this tweet, we can see the interface of XXX being used in this study from a data journalist: ex
- Recommendation: inform about sources, example inform users that this user shared x failed fact-checks.
- Make point about "recycling" existing policies (e.g. to tackle terrorism) and apply it to misinformation+

- Make point about the communication of platforms for their policy: 4R of YouTube, RRI of Facebook,
- Make point that platforms when they take actions regarding misinformation, they rarely cite misinformation as the reason. E.g. beauty of life, coordinated inauthentic behavior.

  https://misinforeview.hks.harvard.edu/article/tackling-misinformation-what-researchers-could-do-with-social-media-data/
- After a channel gets suspended or a video is deleted the data related to them is removed from the YouTube API. This is a problem that affected the study to investigate the suspension reasons or the removal. For instance, the original list of videos from table 1 had more than 200 videos available in March 2021, however, by June 2021 30 videos got deleted from YouTube because they contain content that is against the YouTube guidelines and their data got removed from the API.
- User side: psychological effects (Pennycook, etc.), multi-platforms, indirect effects, strategies...
- business model
- Pas accès au "reach" ! ranking not available. Fb nous a donné certaines données (les données Condor) et dedans on a le nombre de clicks par exemple
- "All four Pages have been unpublished for repeated violations of Community Standards and accumulating too many strikes. While much of the discussion around Infowars has been related to false news, which is a serious issue that we are working to address by demoting links marked wrong by fact checkers and suggesting additional content, none of the violations that spurred today's removals were related to this." Newsroom.
- Researchers should be able to access the data of deleted accounts on the main platforms.
- Currently the flag or notice banners data is not present in the platforms' API datasource, and this prevents researchers and journalists from easily investigating that matter. They have to use or build a scraper to access the data.
- Lack of transparency in the platforms' communication that makes their policies harder to verify.
- Finally, we found irregularities in the number of Infowars articles collected from Buzzsumo. While Infowars usually publishes 20 to 30 articles per day, only 53 articles were collected in the 31-day period between June 11 to July 11, 2020. A temporary crawling problem coming from Buzzsumo may have caused this lack of data. Because no database is perfect, we would like to highlight the importance of cross-checking information between different sources when possible.
- We hope this paper can illustrate how the platforms policies can be verified. All the code needed to collect the data, and create the figures can be found on a public GitHub repo: https://github.com/medialab/truth-and-trust-online-2021/, so that other researchers and journalists can easily reuse our methodology.

# References


[1] Judith Bayer. Between anarchy and censorship public discourse and the duties of social media. *CEPS Paper in Liberty and Security in Europe*, (2019-03), May 2019.

[2] David A. Broniatowski, Daniel Kerchner, Fouzia Farooq, Xiaolei Huang, Amelia M. Jamison, Mark Dredze, and Sandra Crouse Quinn. Debunking the misinfodemic: Coronavirus social media contains more, not less, credible content. *mimeo*, 2020.

[3] Ahiza García-Hodges. Youtube suspends oann for violating its covid-19 policy. *NBCnews, https://www.nbcnews.com/news/all/youtube-suspends-oann-violating-its-covid-19-policy-n1248845*, November 2020.

[4] Nathaniel Gleicher. Removing coordinated inauthentic behavior from georgia, vietnam and the us. *Facebook Newsroom https://about.fb.com/news/2019/12/removing-coordinated-inauthentic-behavior-from-georgia-vietnam-and-the-us/*, December 2019.

[5] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on twitter during the 2016 u.s. presidential election. *Science*, 263(274), 2019.

[6] Adam Hughes and Stefan Wojicik. 10 facts about americans and twitter. *Pew Research Center*, 2019.

[7] David Lazer, Matthew Baum, Yochai Benkler, Adam Berinsky, Kelly Greenhill, Filippo Menczer, Miriam Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven Sloman, Cass Sunstein, Emily Thorson, Duncan Watts, and Jonathan Zittrain. The science of fake news. *Science*, 359(6380), 2018.

[8] Tessa Lyons. The three-part recipe for cleaning up your news feed. *Facebook Newsroom https://about.fb.com/news/2018/05/inside-feed-reduce-remove-inform/*, May 2018.

[9] Paris Martineau. Facebook bans alex jones, other extremists - but not as planned. *Wired https://www.wired.com/story/facebook-bans-alex-jones-extremists/*, February 2019.

[10] Guillaume Plique, Pauline Breteau, Jules Farjas, Héloïse Théro, and Jean Descamps. Minet, a webmining cli tool and library for python zenodo. http://doi.org/10.5281/zenodo.4564399. 2019.

[11] Emily Saltz and Claire Leibowicz. Shadow bans, fact-checks, info hubs: The big guide to how platforms are handling misinformation in 2021. *Nieman-Lab https://www.niemanlab.org/2021/06/shadow-bans-fact-checks-info-hubs-the-big-guide-to-how-platforms-are-handling-misinformation-in-2021/*, June 2021.

# 6 Appendix

## 6.1 Quick access to community guidelines and policies

| | | |
|---|---|---|
| Rules | | facebook.com/communitystandards/recentupdates/ |
| | | help.twitter.com/en/rules-and-policies/twitter-rules |
| | | youtube.com/intl/en_us/howyoutubeworks/policies/community-guidelines/ |
| Rules enforcement | | transparency.fb.com/data/community-standards-enforcement/ |
| | | transparency.twitter.com/en/reports/rules-enforcement.html |
| | | transparencyreport.google.com/youtube-policy/ |
| Transparency center | | transparency.fb.com/data/<br>https://law.yale.edu/yls-today/news/facebook-data-transparency-advisory-group-releases-final-report |
| | | transparency.twitter.com/en/reports.html |
| | | transparencyreport.google.com/?hl=en |
| Policy regarding Covid-19 | | https://www.facebook.com/help/230764881494641/ |
| | | help.twitter.com/en/rules-and-policies/medical-misinformation-policy<br>blog.twitter.com/en_us/topics/company/2021/updates-to-our-work-on-covid-19-vaccine-misinformation<br>blog.twitter.com/en_us/topics/company/2020/covid-19 |
| | | support.google.com/youtube/answer/9891785 |
| Fact-checking policy | | facebook.com/journalismproject/programs/third-party-fact-checking/how-it-works |
| | | x |
| | | support.google.com/youtube/answer/9229632 |
| Fighting misinformation | | facebook.com/formedia/blog/working-to-stop-misinformation-and-false-news<br>about.fb.com/news/2018/05/hard-questions-false-news/ |
| | | |
| | | youtube.com/intl/en_us/howyoutubeworks/our-commitments/fighting-misinformation/#policies<br>blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/ |
| Strike System | | See in the following link the section *What is the number of strikes a person or Page has to get to before you ban them?*<br>about.fb.com/news/2018/08/enforcing-our-community-standards/ |
| | | See in the following links the section: *Account locks and permanent suspension*<br>help.twitter.com/en/rules-and-policies/election-integrity-policy<br>help.twitter.com/en/rules-and-policies/medical-misinformation-policy |
| | | https://support.google.com/youtube/answer/2802032?hl=en. Accessed 21 6 2021 |
| Account suspension | | |
| | | https://help.twitter.com/en/managing-your-account/suspended-twitter-accounts |
| | | |
| Flags, Notice and Information Panels | | |
| | | https://help.twitter.com/en/rules-and-policies/notices-on-twitter |
| | | support.google.com/youtube/answer/9004474?hl=en |

**Table 5:** Summary of ressources, last accessed on July 5, 2021.