Monitoring misinformation related interventions by Facebook, Twitter and YouTube: methods and illustration.

Shaden Shabayek[*], Héloïse Théro[†], Dana Almanla[‡], Emmanuel Vincent[§]

April 29, 2022

Abstract

There is growing pressure for mainstream platforms, such as Facebook, Twitter or YouTube, to fight misinformation by moderating the content that spreads on their site. We investigated the interventions of the platforms by collecting social media data via APIs and scraping. These interventions can be classified into three broad categories: (i) temporary or permanent suspension of users, (ii) displaying flags and information panels, and (iii) reducing the visibility of some content. We provide examples illustrating how researchers can monitor the interventions within each of the three categories for each platform. Finally, we discuss the restrictions to access data and the lack of transparency regarding misinformation related interventions, and how to help the academic community, NGOs and data journalists to successfully study online misinformation.

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-like 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

[*]Corresponding author: shaden.shabayek@sciencespo.fr, Sciences Po, médialab.

[†]thero.heloise@gmail.com, Sciences Po, médialab.

[‡]dana.almanla@cri-paris.org, CRI.

[§]emmanuel.vincent@sciencespo.fr, Sciences Po, médialab.

[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.

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 (see Fathaigh et al. (2021) [3] for a discussion for the case of the EU). 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 misinformation 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 misinformation phenomenon is still ill-defined, 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 when it goes viral. This gave rise to a set of heterogenous policies across mainstream platforms, mainly based on fact-checking but also content moderators and users' reporting. For example, Facebook has a substantial partnership program with Fact-checking partners certified by the non-partisan International Fact-Checking Network. The company 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

[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.

[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.

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 policies are generally 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. 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*". In particular, policies regarding misinformation are generally not part of the set of platform rules or community guidelines (as of July 2021).

Misinformation related 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 to date, 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 mainstream platforms' current actions regarding content with low credibility or false information with data mining. 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 [18], a webmining tool developed by the SciencesPo médialab, was often used, and we scripted our own data mining code when it was necessary.

	Application Programming Interface (API)	**Web Scrapping**
	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, in this article, we survey 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*) informing users with flags and notices, and (*iii*) reducing the visibility of some content. We compile in Table 2, 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

[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/.

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 content created prior to suspension remains available to the user and their followers. However, when the suspension is permanent, in most cases, followers or subscribers no longer have 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 Facebook, Twitter and YouTube, and we 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 historical data of deleted accounts, these reports cannot be verified by the academic community, NGOs and journalists.

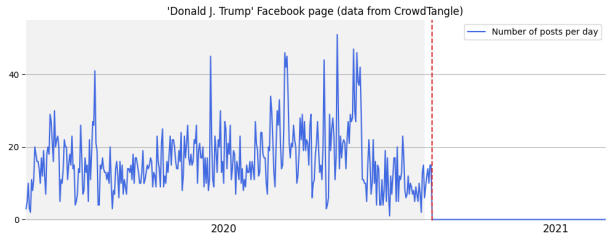

Figure 1: Number of Facebook posts published each day by the Facebook page *Donald J. Trump* between January 1, 2020 and June 15, 2021. The data corresponds to 6 083 posts retrieved from the CrowdTangle API using the *posts* endpoint.

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. Hence, we collected the 6 083 posts it had published between January 1, 2020 and

[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.

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

June 15, 2021 using the *posts* endpoint.[11] The minet Python library [18] was used 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 (see Table 2 in Appendix 6.1). 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.[12] The 12-hour account lock is hard to observe in the data, especially for users who do not have an over the clock regular tweeting activity.[13] 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[14]. In particular, this website has several failed fact-checks concerning the published articles, according to Media Bias Fact Check and open.feedback.org.[15]. 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 the left panel of 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.

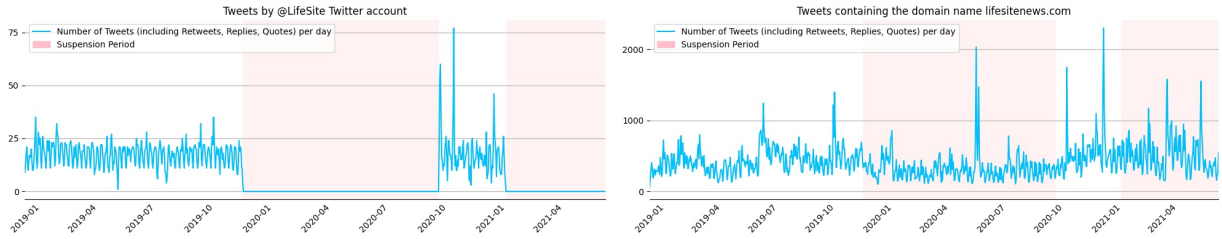

Figure 2: Left Panel: number of Tweets (including Retweets, Replies, Quotes) per day of the Twitter account @*Lifesite* linked to the website lifesitenews.com from January, 2019 until April 2021. Right Panel: number of Tweets (including Retweets, Replies, Quotes) per day that have shared a lifesitenews.com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

To further assess the impact of this double temporary suspension, we collect via the Twitter API v2, all the Tweets that have shared during the same period a url link containing lifesitenews.com.

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

[12] See https://en.wikipedia.org/wiki/Twitter_suspensions.

[13] The following tool https://makeadverbsgreatagain.org/allegedly/ provides an overview of the daily and hourly tweeting activity, including repetition of Tweets for any given Twitter account.

[14] 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.

[15] See https://mediabiasfactcheck.com/life-site-news/ and https://open.feedback.org/media/CM.

The right panel 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, Replies, Quotes) 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, Replies, Quotes) per day during the exact same period one year earlier.

Finally, suspending a Twitter account is an intervention which aims at penalizing users, in order to make them respect the Twitter rules. In the case of Twitter accounts linked to domain names, the activity of the account usually consists in spreading on social media articles published on their website. The right panel of Figure 2 points towards the limitations of suspending the Twitter account @$LifeSite$, because the articles published on lifesitenews.com were still being actively shared by other Twitter users.

2.3 YouTube

In this section, we turn to the suspension policy of YouTube. When a YouTube channel publishes a video that violates the community guidelines for the first time, it will usually receive a warning. But for the second violation, the channel receives its first strike. A strike means that users are not allowed to post content on their YouTube channel for one week. A strike remains associated to a channel for 90 days, and if a second strike occurs in the same 90-day period as the first strike, the suspension will then last for two weeks. A third strike results in the termination of the channel.[16] To illustrate the implementation of this policy, we use the following YouTube channels as two examples of one-week suspensions: One America news Network and Tony Heller.

The website of One America News Network has a "low" factual reporting score according to the website Media Bias Fact Check and several inaccurate articles according to open.feedback.org.[17] According to NBC news (2020) [4], the YouTube channel received a first strike on November 24, 2020 for the promotion of a false cure for COVID-19. We collected the number of video uploaded and the number of view of this channel using the YouTube API v3, between November 2020 and January 2021. For the video counts, we used the *playlists* 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, then via the *videos* endpoint we retrieved the view counts on June 2021.

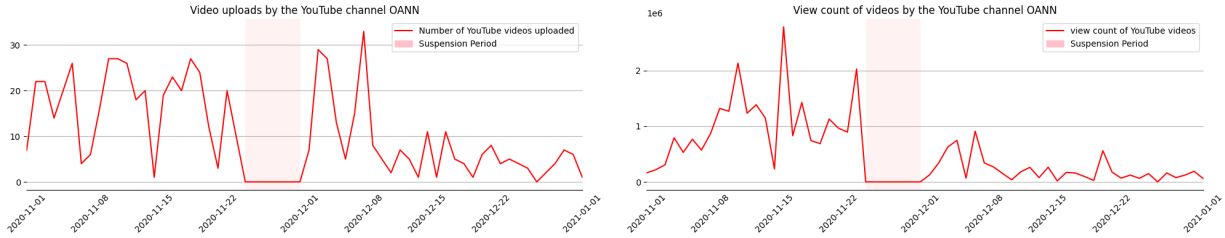

Figure 3: Left Panel: Number of YouTube videos uploaded each day by the youtube channel *One America news Network* November 1, 2020 and January 1, 2021. Right Panel: accumulated view counts for videos. The metrics correspond to the videos' publishing date and the data is retrieved from the YouTube API with the *playlists* and *videos* endpoints.

[16] https://support.google.com/youtube/answer/2802032.

[17] See mediabiasfactcheck.com/one-america-news-network/ and https://open.feedback.org/media/5M.

Figure 3 shows the daily number of videos uploaded by the channel, and its number of views. We can clearly see the suspension period in the data, as the number of published video is down to zero for one week after the strike on November 24, 2020. Because no videos were published during this period, the number of views were also nul. When comparing one month before and after the suspension, the view count decreased by -73%, indicating that the videos uploaded after the suspension were less popular. Furthermore, the activity of the channel was close to zero after March 2021, and OANN's Twitter account has announced their migration from YouTube to Rumble on March 17, 2021 (see the left panel of Figure 4).

Figure 4: Left Panel: Tweet announcing OANN moving to Rumble, Twitter ID 1372238828425998336. Right Panel: Tony Heller's tweet after getting temporally suspended from YouTube.

Tony Heller writes regularly blog posts on the website *realclimatescience.com*, which also has a "low" factual reporting score according to the website Media Bias Fact Check.[18] He tweeted on September 29, 2020 about his YouTube channel being 'shut down' because of a video about an anti-covid-lockdown doctor getting arrested (see the right panel of Figure 4). We applied the same methods as in the previous example to collect data. Again a one-week suspension period from September 29 until October 5 can be observed clearly in the daily number of uploaded videos and of views (see Figure 5). Comparing one month before and after the suspension, this channel also witnessed a drop of view counts by -70%.

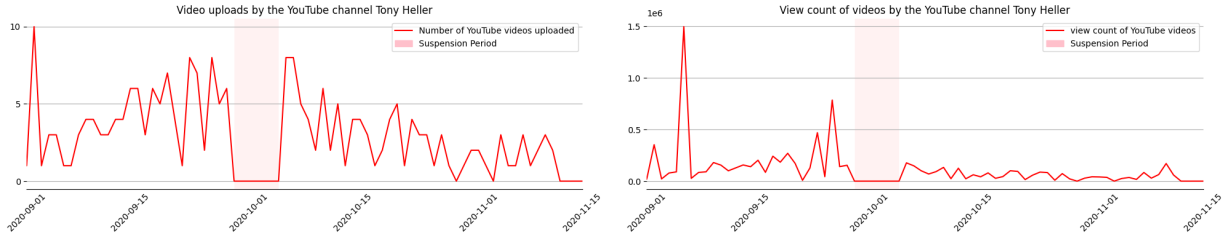

Figure 5: Left Panel: Number of YouTube videos uploaded each day by the YouTube channel *Tony Heller* between September 1, 2020 and November 15, 2020. Right Panel: accumulated view counts for videos uploaded by the same YouTube channel. The date corresponds to the videos' publishing date.

As most YouTube channels upload videos on a weekly or even a monthly basis, a one-week suspension does not appear very restrictive at first sight. But we observed that the intervention

[18] See mediabiasfactcheck.com/real-climate-science/

was followed by a reduction in the number of view for the two channels investigated, with one channel even stopping its YouTube activity in the following months. Although more research is needed, it is possible that the one-week suspension following a strike might reduce the reach of active YouTube channels.

3 Flags, Notices and information panels

Mainstream platforms such as Facebook, Twitter and YouTube can resort to providing more context to users regarding a specific post, Tweet or video. This type of intervention takes different formats according to the platform and also has different denominations. Facebook for example refers to this intervention by mentioning "flags", while Twitter refers to "notices" and "interstitials" and YouTube uses the term "information panels". In this section, we explain the specifics of this intervention for each platform and illustrate with examples.

3.1 Facebook

To the best of our knowledge, two types of flags can currently be displayed on Facebook posts: (*i*) information banners which do not refute the message in the post and provide a link which redirects to an authoritative source, such as "Visit the COVID-19 Information Centre for vaccine resources" and (*ii*) fact-check flags that assigns a "rating" for a text or a link in a given post, such as "False information Checked by independent fact-checkers" (see Figure 6). The fact-check flags can display one of the following ratings : "False", "Partly false", "Missing context", "False headline", "Altered media", "Opinion", "Satire", "Not eligible" and "True".[19]

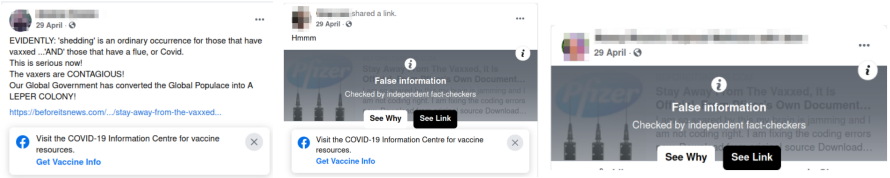

Figure 6: Examples of Facebook posts having shared a link fact-checked as False by one of Facebook's partners. Screenshots taken on July 8, 2021.

No information or available fields regarding the Facebook flags can be found on Buzzsumo or CrowdTangle, the two APIs we use to access Facebook data. The only way to verify Facebook's flagging policy is thus via scraping publicly available content.

We first searched for all the Facebook posts having shared a specific link[20] rated as 'False' by Science Feedback, a fact-checking organization which partners with Facebook. To do so, we used the 'search' endpoint in CrowdTangle API. Twenty Facebook posts were collected this way, and we scripted an adapted scraping code to verify whether they contained 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 informed Facebook that they have assigned the rating "False" for this link, we expected that all these posts would contain a fact-check flag, but we had no expectations for the information banner.

[19] See https://www.facebook.com/business/help/341102040382165.
[20] Click here to access the link and its fact-check by Health Feedback.

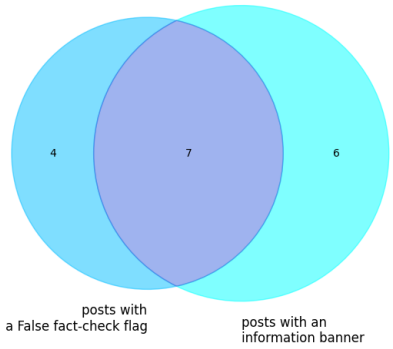

Figure 7: Count of different flags for 17 Facebook posts who have shared the exact same link, fact-checked False by one of Facebook's partners.

Surprisingly only 11 posts out of the remaining 17 had a "False information" flag (see Figure 7 and 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" flag if the link is not expanded.

We also observed that most of the posts sharing this "False" link (13 out of 17, Figure 7) 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 8). Interstitials are presented as a greyed box on top of a Tweet, which indicates 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 9). 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 contains a notice or not; while the interstitial "possibly sensitive" and "withheld" are both Tweet fields that can be recovered[21] from the Twitter API v2. Hence, in order to investigate the presence of notices of all types, we resorted to scrape publicly available Tweets, using minet [18].

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

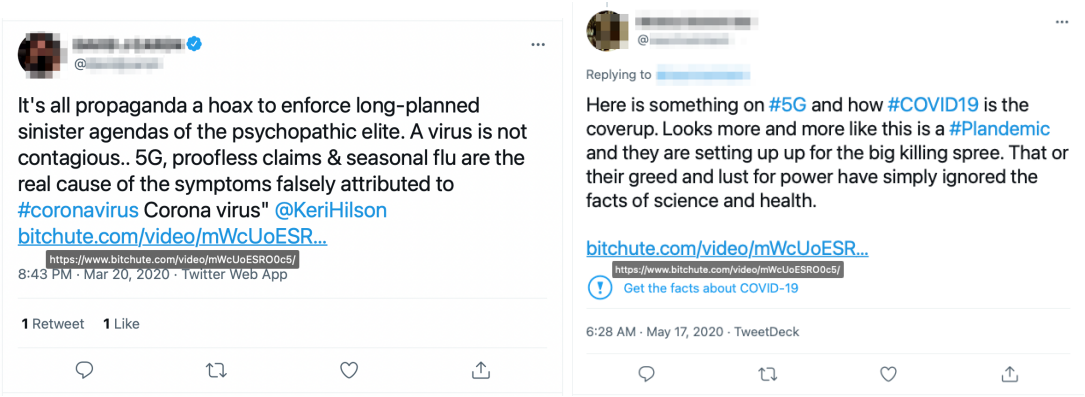

Figure 8: Two Tweets sharing the same link marked as "False" by Science Feedback, screenshots taken on June 20, 2021. Left Panel: Tweet without an information notice. Right Panel: Tweet containing an information notice.

Figure 9: 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 in the left panel. Screenshots taken on July 1, 2021.

This tool was recently enriched upon our request in order to capture whether a Tweet contains a notice or not, via the "minet twitter scrape" command.

In this section, we take a deeper look at how notices and interstitials are introduced by Twitter, to add context to potentially inaccurate or misleading content. To that end, we gathered a set of 3 094 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 all the Tweets that have shared at least one link out of the set of 3 094 links marked as *False*. This data collection resulted in 323 938 Tweets, excluding Retweets. Only 28 Tweets contained the notice *"Get the facts about COVID-19"*, 5 Tweets contained the notice *"Learn about US 2020 election security efforts"* and only 1 had the following notice *"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 notice *"Get the facts about COVID-19"* (see Figure 8). 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 9 344 (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[22] marked as "False", which also contains the interstitial *potentially sensitive content* (see Figure 9). Among the 32 Tweets that have shared the exact same link, only 5 tweets had the interstitial "potentially sensitive content". Again this points towards the non-automation of interstitials placement based on a specific URL. To date, no data is available to verify whether some interstitials or notices were placed based on: Tweet reporting by other users, or Twitter moderators, or algorithmic rules. Furthermore, notice that the 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*.

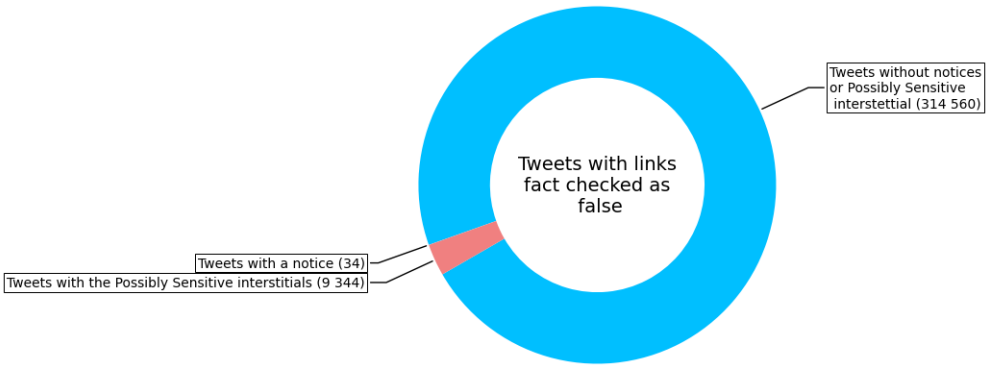

Figure 10: Summary of the number of Tweets among the set of 323 938 Tweets containing a link marked as False, by Science Feedback, which contain a notice or a Possibly sensitive interstitial or none.

Finally, the 9 344 Tweets containing the interstitial "potentially sensitive content" widely outnumber the 34 tweets which contain a notice (see Figure 10). 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 do not contain the label *"Get the facts about COVID-19"*. We also noticed that the 9 344 Tweets which contain the interstitial *"potentially sensitive content"* do not necessarily contain graphic violent content[23] but rather

[22]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.

[23]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

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 notices which provide more context to users.[24]

3.3 YouTube

YouTube provide information panel at the top of search results or under a video related to topics prone to misinformation. Information panels provide more context on a controversial matter, with a link to an independent, third-party partner's website such as Wikipedia or the World Health Organization. Importantly, YouTube states that these panels exist regardless of the point of view expressed in a given video. YouTube also specifies that information panels may not be available in all countries/regions and languages.[25]

To further study the assignment of information panels to videos, we compiled a list of 170 YouTube videos, fact-checked by Science Feedback between April 2019 and March 2021 and marked as containing inaccurate information. We collected the information panels, when present, under each video by scraping the content of the web page. Three types of information panels were found regarding COVID-19, COVID-19 Vaccine, and climate change, as shown in Figure 11.

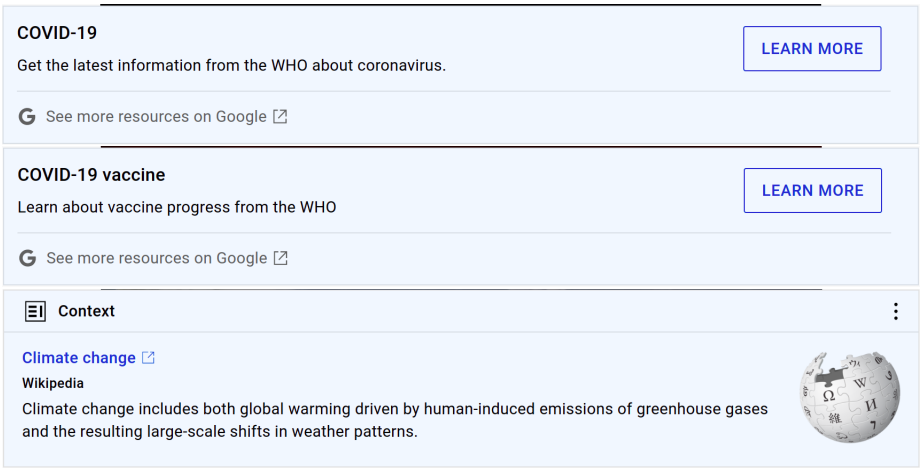

Figure 11: Examples of YouTube information panels displayed under Youtube videos respectively about: COVID-19, COVID-19 Vaccine and climate change. The videos were accessed on September 2, 2021.

Almost half (46 %) of the videos containing misinformation were shown without any information panel (Figure 12). For the rest of videos, the most frequent information panel concerned COVID-19 (42 %), followed by COVID-19 vaccine (9 %). It is possible that YouTube is currently focusing its efforts on the current 'hot' controversial topics with a potentially wide public impact, leaving room for misinformation related to other topics to spread, but more research is needed to formally compare misinformation videos about COVID-19 and about other subjects.

be aware that they will see sensitive media if they click through".

[24]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".

[25]https://support.google.com/youtube/answer/9004474

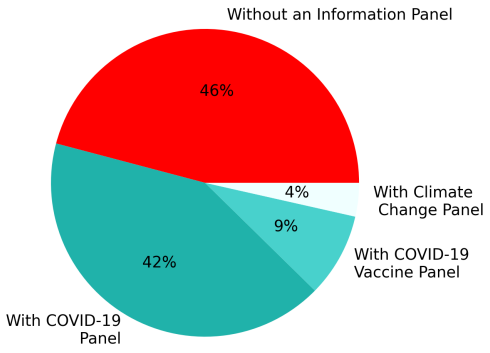

Figure 12: Summary of the number of YouTube videos which contain an information panel, among the set of videos marked as false by Science Feedback.

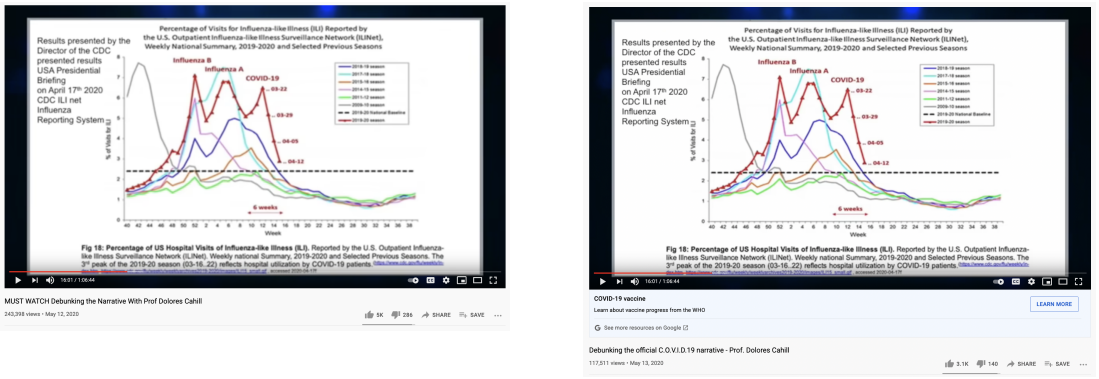

Figure 13: Screenshots showing two duplicates of the same video, taken on July 21, 2021. Left panel: a YouTube video about COVID-19 without an information panel. Right panel: a duplicate of the same video uploaded under a different title and with a 'COVID-19 vaccine' information panel.

We noticed that the same video was sometimes uploaded multiple times and thus appeared on different YouTube links. One YouTube link might contain an information panel while its duplicates did not (see Figure 13 for an example). In addition, we noticed that some keywords in the video titles were often associated with an information panel, such as 'COVID-19', 'coronavirus', 'pandemic' and 'testing', while some notational variants like 'C O V I D 19' or 'Cv19' were not associated with a panel. Therefore, we suspect that YouTube is automatically adding information panels based on the presence of certain keywords in the video title (and maybe in the video description), independently of the content of the video. Note that in Figure 13, the video with 'C.O.V.I.D.19' in its title had an information panel, hinting that YouTube might adapt its list of suspect keywords as new notational variants for COVID-19 appear, or that some information panels might be added manually (by content regulators or when a video is flagged by many users).

4 Reducing the visibility

Mainstream social media platforms can reduce the visibility of content created or shared by 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 indirect methods to explore how reducing the visibility of content works on Facebook, Twitter and YouTube.

4.1 Facebook

To tackle misinformation, Facebook can reduce the spread of misleading content through their built-in ranking system. More specifically, 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 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[26]. Furthermore, the factual reporting of $Infowars$ has been rated *very low* by the website Media Bias Fact Check and it has several failed fact-checks reported by open.feedback.org.[27]

On May 2, 2019, Facebook announced they would prohibit users from sharing Infowars content unless, they are explicitly condemning the material.[28] To verify the measure, we used the "/posts/search" endpoint[29] 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.[30]

The number of public posts sharing an $Infowars$ link remained globally stable throughout 2019 (see Figure 14 left panel). Thus the measure announced by Facebook does not seem to have prevented users from sharing an $Infowars$ link. Nevertheless, a clear drop in engagement was observed on May 2, 2019 (see Figure 14 right 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

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

[27] See https://mediabiasfactcheck.com/infowars-alex-jones/ and open.feedback.org https://open.feedback.org/media/B9.

[28] 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/.

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

[30] 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.

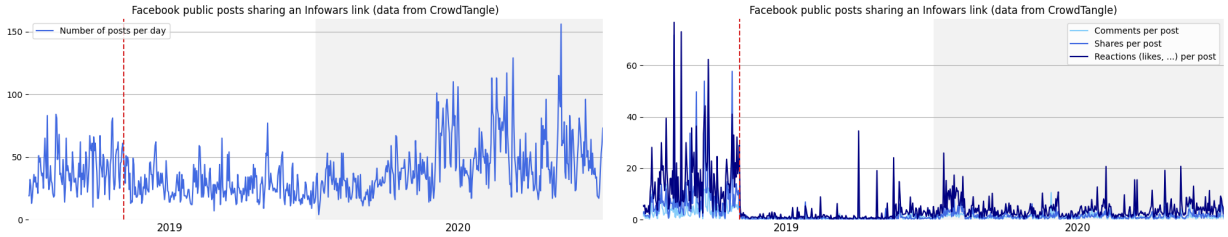

Figure 14: Public Facebook posts sharing an Infowars link in 2019 and 2020 collected from the CrowdTangle API. The red line marks the date of May 2, 2019, when Facebook has announced the ban regarding Infowars. Left panel: Number of daily posts. Right panel: Engagement metrics: average number of reactions, shares and comments per post.

again by the end of 2019 / beginning of 2020, suggesting that the *reduce* measure may have been lifted a few months after its implementation.

Furthermore, Facebook has also announced on December 20, 2019 that "The BL is now banned from Facebook".[31] The Beauty of life (thebl.com/) is a US-based media company operated by The Epoch Times that shares pro-Trump views and conspiracy theories such as QAnon.[32] To verify Facebook's ban, 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[33]. 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.

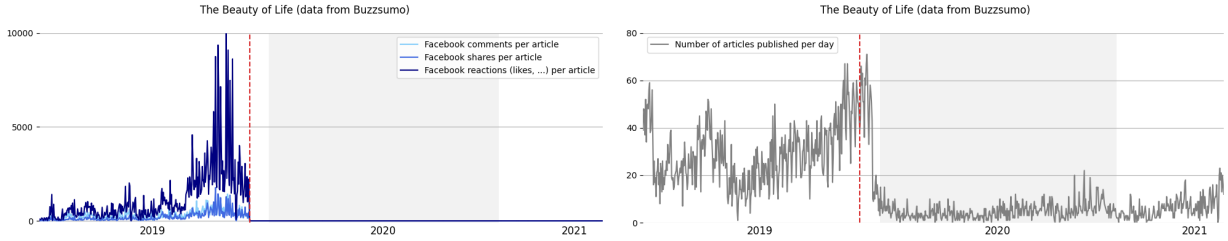

Figure 15: Articles from The Beauty of Life website (thebl.com) published between January 1, 2019 and June 15, 2021 and gathered from the Buzzsumo API. Left panel: Facebook engagement metrics (average number of reactions, shares and comments per article). Right panel: Number of articles published per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

The number of Facebook reactions, shares and comments dropped to zero for TheBL's articles published after December 1, 2019 (see Figure 15 left panel), indicating the start of the ban. We note that although the ban was communicated in an article published on December 20, 2019, it

[31] `https://about.fb.com/news/2019/12/removing-coordinated-inauthentic-behavior-from-georgia-vietnam-and-the-us/`

[32] See wikipedia article.

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

seems to have actually started on December 1, 2019. The communication around the ban appeared 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 15 right 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[34] 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 Media Bias Fact Check, a website which provides a measure of factual reporting of several media outlets based on available fact-checks.[35]

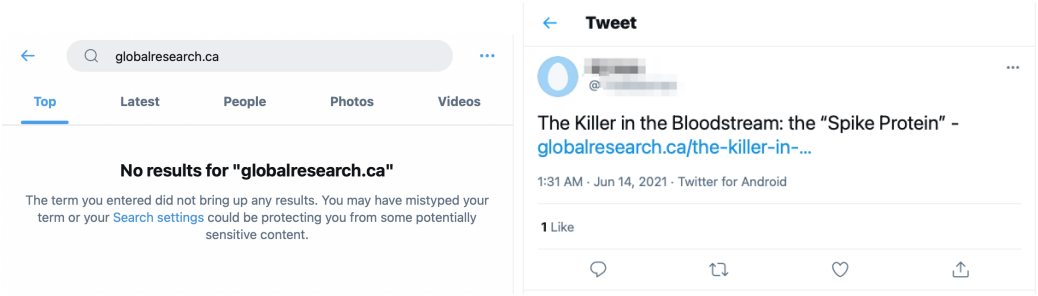

Figure 16: Screenshots taken on June 14, 2021. Left Panel: screenshot that shows that no results can be found when searching for articles linked to the domain name *globalresearch.ca* via the Twitter search box. Right Panel: example of a Tweet posted on June 14 containing the domain name *globalresearch.ca*.

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 16, taken on June 14, 2021. To further investigate the possible implementation of a reduced visibility measure, we searched via the Twitter API v2 for Tweets, excluding Retweets, containing the query *globalresearch.ca* from April 15, 2021 until August 15, 2021. As shown in the left panel in Figure 17, we find a strictly positive number of Tweets containing the domain name *globalresearch.ca* after the date on which the screenshot was taken (left panel of Figure 16). 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 (see right panel of Figure 16). Furthermore, those collected Tweets have strictly positive engagement metrics as shown in the right panel of Figure 17. Hence, the users who post Tweets including articles from the *globalresearch.ca* website receive Tweet level engagement from their own followers.

Finally, the website *globalresearch.ca* is linked to the Twitter account @*CRG_CRM*, which was suspended from Twitter on May 25, 2021 as shown in a message about the account's suspension

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

[35] See https://mediabiasfactcheck.com/global-research/.

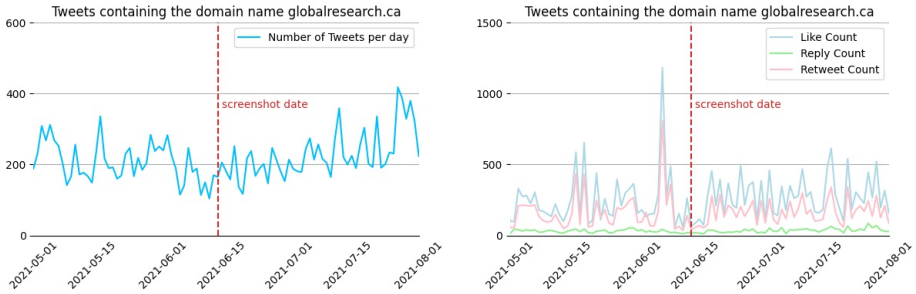

Figure 17: Left Panel: daily number of Tweets, excluding Retweets, containing the query *globalresearch.ca* from January 1, 2021 until June 30, 2021. Right Panel: engagement metrics of Tweets containing the query *globalresearch.ca* from January 1, 2021 until June 30, 2021. Data collected via the Twitter API v2 on August 17, 2021.

(see the right panel in Figure 18). 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 18). 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.

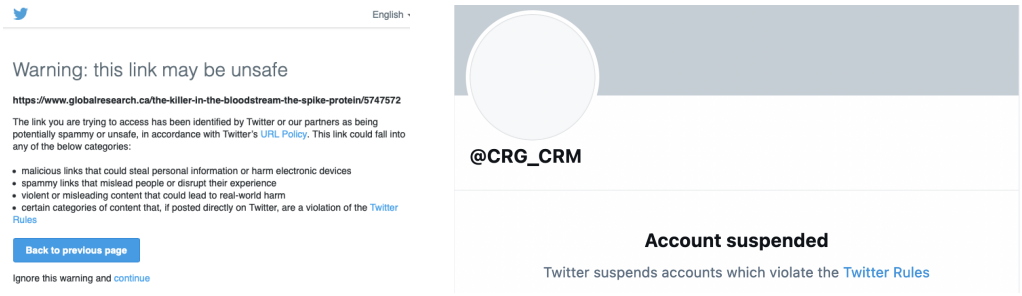

Figure 18: Screenshots taken on June 14, 2021. Left Panel: warning message that appears when a user attempts to click on a link which contains the domain name *globalresearch.ca*. Right Panel: account suspension message of @*CRG_CRM* linked to *globalresearch.ca*.

4.3 YouTube

YouTube can reduce the visibility of certain videos via its recommendation system, namely by recommending content from authoritative sources. The platform usually recommends content to the users based on the watch history and searched queries in google and YouTube; which can both be cleared via account settings. A recommendation system based on previous preferences has a downside, namely when a user repeatedly watches videos that might be misleading or that promote conspiracy theories and gets recommended similar videos. YouTube states that they set

out to prevent their systems from serving up content that could misinform users in a harmful way, particularly in domains that rely on veracity, such as science, medicine, news, or historical events. YouTube identifies misinformation by using external human evaluators and experts to examine if a given video is promoting misleading information or conspiracy theories, and then use machine learning systems to discard misinformation from their recommendations.[36]

To examine this effect, we designed an experiment that simulates a user's behavior on YouTube. We used a list of 45 videos marked as *False* by Science Feedback and classified under the category "health". The experiment starts with one video marked as "False" from our list of videos and a clean watch history. For each video, our automated program collects around[37] 20 recommendations and clicks on the top 10 videos in sequential order. After that, for every video in the set of the top 10 recommendations, we collect their own top 20 recommended videos. In addition, to simulate a user's behavior, the automated program watches each video for a few minutes[38]. Doing so, we collected 10 549 YouTube recommended video entries, among which 3 895 were unique. To analyse the collected data, we look for misinformation: (i) at the channel level and (ii) at the video level.

The 3 895 unique videos came from 1 562 different channels, and we counted the number of times every channel got recommended. Figure 19 shows the top 10 recommended channels, whose videos represent 30% of the recommendations. Fox News and Sky News Australia, which both have a "mixed" factual reporting level according to the website Media Bias Fact Check and have several failed fact-checks,[39] are at the first and third place respectively. PragerU is also among the list of the 10 most recommended channels, with a "low" level of factual reporting according to the website Media Bias Fact check.

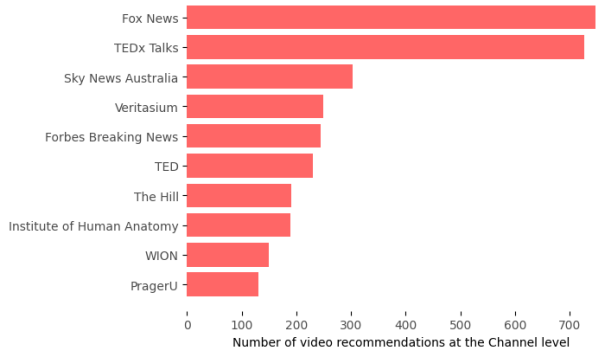

Figure 19: Top 10 channels recommended by YouTube when we simulated a user that starts by watching a video fact-checked as False and then follows the top recommendations.

We then ranked the videos based on the number of times they got recommended, and asked Science Feedback to fact-check the top 40 recommended videos (those videos were recommended 1 735 times, which represents 16% of the recommendations). Among the 40 examined videos,

[36] See the page "How does YouTube address misinformation?"

[37] When we visit a video, we get a number of recommendations that vary between 18 videos and 22 videos.

[38] If the video duration is ten minutes or less, the automated program watches the whole video. If the video duration is strictly higher than ten minutes, then the automated program watches only five minutes.

[39] See https://mediabiasfactcheck.com/fox-news-bias/ and https://mediabiasfactcheck.com/sky-news-australia/.

four videos were marked as *misleading* and they were recommended 132 times out of the 1 735 recommendations (that is 7.7% of the recommendations). Four other videos were marked as *hyper-partisan* and they were recommended 184 times (10.7%).

As we collected the recommendations of misinformation videos, we expected to find a high proportion of videos of questionable content. But only 18.4% of the recommendations corresponded to *misleading* or *hyper-partisan* videos. This relatively small percentage hints towards that idea that YouTube is probably reducing the visibility of videos containing misinformation. However, when we investigated misinformation at the channel level, we found that videos belonging to channels that are not highly authoritative were being often recommended. It appears that the recommendation system of YouTube is reducing the visibility at the video level but not necessarily at the channel level.

5 Discussion

In this article we have assessed three broad categories of interventions to tackle misinformation by Facebook, Twitter and YouTube: (*i*) suspension of accounts, (*ii*) introduction of flags, labels and information panels and (*iii*) reducing the visibility of content.[40] Furthermore, we have provided a methodology that can be useful for third-party monitoring. The code needed to collect the data and create the figures is available on the following public GitHub repository: github.com/medialab/webclim_misinformation_related_interventions, and we hope it can be a useful ressource for researchers, NGOs and journalists addressing online misinformation.

Mainstream platforms such as Facebook, Twitter and Youtube are increasingly working towards higher data transparency. We invite the reader to see links that redirect to the respective transparency centers in Appendix 6.1. Yet the present research sheds light on the restrictions to access data and its limits, which impede the academic community from studying successfully online misinformation and assessing the impact of platforms' interventions regarding this phenomenon. Hence, few comments are in order regarding the limitations to access data, the lack of transparency vis-à-vis misinformation related interventions and the lack of interventions designed specifically for misinformation.

A need for more complete data. When Facebook, Twitter or YouTube intervene by suspending an account, the data can no longer be accessed via scraping or via the official APIs (Twitter, YouTube, CrowdTangle) for the purposes of scientific research. This hampers the investigation of how platforms moderate content. To further illustrate this point, the original list of YouTube videos with failed fact-checks (summarized in Figure 12), had over 200 YouTube videos available in March 2021. However, by June 2021, thirty videos had disappeared from the YouTube API for policy violation. Similarly, in a related research project, Théro and Vincent (2020) [23] searched on CrowdTangle for 94 public accounts sharing specific content associated with misinformation in November 2020. They tried to collect the posts from those 94 pages in January 2021, to only discover that 11 pages had disappeared from the CrowdTangle API. Researchers' work could be simplified, if metadata for content posted by removed accounts could still be accessible via the APIs.

Although it is currently impossible to verify permanent deletions using available social media data, their effects can be observed using an adapted methodology. Following the Capitol riots

[40] For a detailed list of interventions of multiple platforms, we invite the reader to navigate through the following airtable compiled by Slatz and Leibowicz (2021) [21].

in January 2021, Twitter has announced the suspension of more than 70 000 accounts related to QAnon.[41] A journalist has recently investigated the effects of this measure by counting the number of times hashtags related to the QAnon movement have been used over time, such as #QAnon or #WWG1WGA. A drastic reduction in hashtag use was found when comparing the first quarter of 2021 with the first quarter of 2020.[42] Thus, counting specific hashtag use over time allows to indirectly observe permanent suspensions on Twitter and there is a need for similar methodologies to be experimented on data from other platforms. However, a point of concern is that data can also disappear because of other factors independent of platforms' content moderation. To provide an example, McMinn et al. (2012) [11] collected 120 million Tweets related to general topics. The same published set of Tweet IDs was then used by Mazoyer et al. (2020) [10], to run the same collection in 2018. Mazoyer et al. (2020) [10] were only able to retrieve 67 millions of Tweets (55%). The authors argue that this drop is probably due to Tweets having been removed by users and Twitter accounts being shut by the users themselves. As data can disappear from platforms over time, it is hard to distinguish whether platforms or users are the ones responsible for data removal. Hence in the special case of Twitter accounts related to QAnon, it is hard to identify how much of the drop in QAnon related hashtag use is due to the Twitter intervention or to accounts migrating to other platforms out of fear of having their accounts suspended by the platform or simply a drop in using the hashtag for other unidentified factors.

Aside from the disappearance of data due to content moderation, some essential metrics for the study of misinformation are not accessible via main APIs. In particular, the *reach* of posts on Facebook, which is the number of people who actually see a given post, cannot be recovered from the CrowdTangle API. [43] Hence, researchers can only build proxies for this metric based on available data (see section 4.1 for an example). Furthermore, to the best of our knowledge, fields related to flags, notices and information panels are not available on the official APIs, with the exception of the *withheld* and *possibly sensitive content* interstitials in the Twitter API v2. For the purposes of the present research we thus had to scrape this information from publicly available content on Facebook, Twitter and YouTube. Again, this disrupts the monitoring of the platforms' interventions and the assessment of their impact (Pasquetto, Swire-Thompson, Amazeen et al. (2020) [14]).

A last concern regards the growing restrictions to access APIs (Perriam et al.(2019) [17]). In April 2018, likely as a response to the Cambridge Analytica scandal, Facebook severely limited access to their official APIs[44] and CrowdTangle is today a reliable source for researchers to access Facebook data. However, it should be noted that CrowdTangle has been recently used by journalists to show that Facebook is disproportionally favoring highly partisan information[45] and a general suspicion is growing that its access might be shut down in a near future to journalists and researchers.[46] Regarding Twitter and YouTube, it is still possible and relatively easy to access their

[41] https://blog.twitter.com/en_us/topics/company/2021/protecting–the-conversation-following-the-riots-in-washington–.

[42] See https://twitter.com/Shayan86/status/1394742784683298818.

[43] See https://help.crowdtangle.com/en/articles/4558716-understanding-and-citing-crowdtangle-data.

[44] See https://about.fb.com/news/2018/04/restricting-data-access/ and http://thepoliticsofsystems.net/2018/08//facebooks-app-review-and-how-independent-research-just-got-a-lot-harder/.

[45] See the news article https://www.economist.com/graphic-detail/2020/09/10/facebook-offers-a-distorted-view-of-american-news and the Tweets from https://twitter.com/facebookstop10. The last link is a Twitter account showing each day the ten most shared link on Facebook, most of them coming from extreme right and misinformation sources. Facebook has reacted to this here: https://about.fb.com/news/2020/11/what-do-people-actually-see-on-facebook-in-the-us/.

[46] See the news articles https://www.nytimes.com/2021/07/14/technology/facebook-data.html and https://www.bloomberg.com/news/articles/2021-08-03/facebook-disables-accounts-tied-to-nyu-research-project.

official APIs.

A need for more transparency. Mainstream platforms generally lack transparency regarding their misinformation related interventions. The broad lines of policies regarding misinformation are communicated via blog posts and are not typically part of the set of rules or community guidelines of mainstream platforms (see Appendix 6.1). Yet, when navigating through the categories of policy areas, to the best of our knowledge, misinformation is absent from the data available on the transparency centers of Facebook, Twitter and YouTube. Platforms' official communication about misinformation interventions is scarce and the academic community, NGOs and data journalists, usually discover interventions related to misinformation via monitoring social media accounts related to domain names with several failed fact-checks or via articles in News outlets. For example, while conducting the present research, it was brought to our attention that the Twitter account related to the domain name *globalresearch.ca* was suspended. But to the best of our knowledge, there was no official communication by the Twitter Safety team to explain the reasons of this suspension (see section 4.2).

We are aware that full transparency regarding operational aspects of misinformation related interventions can backfire. This is because the aimed users might find ways to circumvent the interventions. We noticed that some YouTube channels are trying to avoid the likely automated COVID-19 information panels, by using notational variants such as "C.O.V.I.D" or "C O V I D". Similarly, we came across notational variants on Twitter of the domain name *off-guardian.org* [47] such as *off-guardian. org*, *off-guardian .org*, *off-guardian./org*, *off-guardian*org*.[48] Furthermore there is little information about how algorithms are practically used as a tool for content moderation; since they can be used to reduce the visibility of some content or even prevent problematic content from being recommended. Hence researchers have to resort to proxy measures and simulated experiments to have a better understanding the impact of algorithmic design. Based on a sample of 23 YouTube studies, Yesilada and Lewandowsky (2022) [25] find that YouTube recommender system could lead users to problematic content, but highlight the limitations of this conclusion since there is an incomplete understanding of the recommendation system and models built might not reflect the actual functioning of YouTube recommendation algorithm. Hence, a balance is needed between: (*i*) fully transparent communication about misinformation related interventions so that both users and researchers can understand and investigate how the platforms are regulating content, and (*ii*) not explicitly giving too much information which would allow the policies to be bypassed.

In certain cases, the platform interventions were not motivated by content moderation to limit misinformation, but rather by other rule violations. For example, the official reason for the suspension of Donald Trump's Twitter account was not misinformation[49] but the violation of Twitter's Glorification of Violence Policy, with his last Tweets encouraging the Capitol's riot.[50] Similarly, when Facebook removed four pages related to the InfoWars website, the cited reason was repeated violations of Community Standards: "While much of the discussion around Infowars has been related to false news, which is a serious issue that we are working to address [...], none of the violations

[47] Here is an example of a failed fact-check of one of the articles of the website *off-guardian.org* : politifact.com/factchecks/2020/jul/07/blog-posting/covid-19-tests-are-not-scientifically-meaningless/

[48] For this case, we failed to understand precisely what kind of intervention Twitter users were circumventing, since they could share in a Tweet an article redirecting to the website *off-guardian.org* with the correct notation and only a warning message would appear when one clicks on it.

[49] Although Donald Trump may have spread many misinformation during his governance, see for example: http://allianceforscience.cornell.edu/blog/2020/10/what-drove-the-covid-misinformation-infodemic/.

[50] See https://blog.twitter.com/en_us/topics/company/2020/suspension.

that spurred today's removals were related to this".[51] Regarding The Beauty of Life website, its ban was officially due to coordinated inauthentic behavior,[52] which includes using fake accounts that misrepresent one's identity or using methods to artificially boost the popularity of content. According to Facebook, coordinated inauthentic behavior is a distinct phenomenon from disinformation, as "most of the content shared by coordinated manipulation campaigns isn't probably false".[53] Nevertheless, both misinformation and coordinated inauthentic behavior were attested for the Beauty of Life, and reported to Facebook by the fact-checking organization Snopes.[54] In the above examples, the compagnies' communication was very clear on the reason of the interventions (although one can suspect that the official reason given can depart from the real motivation behind an intervention). However, very often no official communication on behalf of the platforms can be found for a specific interventions observed in collected data. Because misinformation is often associated with creating fake accounts to speed its spread, and incitation to violence, it is sometimes unclear to distinguish between interventions due to misinformation and interventions due to other guideline violations, and more transparent communication on platforms' actions would help monitor the interventions specifically related to misinformation.

A need for more consistent and adapted policies. Misinformation on the web is a problem that is increasingly discussed since the 2016 American elections and in the years that followed, platforms may have decided on a case-by-case basis which actions to enforce. Infowars and The Beauty of Life domains have both several failed fact-checks[55] and have both violated Facebook's community guidelines (incitation to violence and coordinated unauthentic behavior). Yet one's distribution was reduced for only half of a year, while the other one's links cannot be shared anymore on Facebook since the end of 2019, and it is not clear why such different enforcements were applied to these comparable websites. It is also possible that platforms usually follow specific rules on how to fight misinformation, but that sometimes the sudden media coverage of specific events leads them to take exceptional actions, such as with the Capitol riots and Trump's suspension from Facebook, Twitter and YouTube. We would argue that policies should be clearly stated and carefully designed, rather than implementing interventions on a case-by-case basis and in response to the latest news events.

Furthermore, it is likely that some interventions used to tackle online misinformation correspond to interventions that were designed for earlier policy areas. To illustrate, YouTube explicitly states [56] that: "Several policies in our Community Guidelines are directly applicable to misinformation", such as the deceptive practices, impersonation and hate speech policies. Hence interventions such as content deletion or account suspension, might be efficient for policy areas for which they were specifically designed, such as the policy area *Adult Nudity & Sexual Activity*. Nevertheless, tackling online misinformation might need adapted interventions. Namely, account suspensions or content deletion motivated by the spread of inaccurate information can be perceived as censorship by the targeted users. In particular, such interventions can generate platform migrations (see Rogers (2020) [19]) towards new alternative social media platforms (e.g. Telegram, Gab, Minds).

Finally, the means to tackle online misinformation and understanding the phenomenon itself,

[51] See the article https://about.fb.com/news/2018/08/enforcing-our-community-standards/.

[52] See https://about.fb.com/news/2019/12/removing-coordinated-inauthentic-behavior-from-georgia-vietnam-and-the-us/.

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

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

[55] See https://mediabiasfactcheck.com/infowars-alex-jones/ and https://mediabiasfactcheck.com/the-bl/.

[56] See What policies exist to fight misinformation on Youtube? (Last visited September 2021).

is an active area of research. Developing pertinent interventions against misinformation depends on understanding the interventions currently used by platforms (e.g. for labeling, see Morrow et al. (2021) [12]) along with the psychology of interaction with social media (Pennycook and Rand (2021) [16]). Adding fact-checking labels to controversial content is certainly the intervention whose effectiveness was the most studied so far. It was shown that credibility indicators do reduce misperceptions and sharing, although their impact varied vastly depending on the presentation of the warnings (see Yaqub (2020) [24] and Sharevski (2021) [22]). Moreover, a fact-checking label can lead to unpredicted consequences depending on the perception of the person reading it. On Facebook, Twitter and YouTube, fact-checking messages are currently applied only to some content, while the rest of the content on a given platform remains naturally unlabelled. Such sparse application of warnings can lead to an "implied truth" effect, where users may assume that content without warnings have actually been verified and shown to be true (Pennycook et al. (2020) [15]). Using a different approach, Mosleh et al. (2021) [13] identified Twitter users sharing false news and replied to their false Tweets with links redirecting to fact-checking websites. They observed a decrease in the quality and accuracy of the content subsequently shared by the corrected user. These research papers highlight the necessity to investigate potential unintended effects of well-meaning interventions. Furthermore, misinformation related interventions can be poorly misunderstood. Based on survey data, Saltz et al. (2021) [20] show that 40% of the survey respondents (wrongly) believed that content is mostly or all fact-checked. To date, research about misinformation related interventions by platforms is still scarce and addressing the specific needs in terms of data for auditing and monitoring tasks can greatly help in improving our understanding of these interventions and study their impact. The methods presented in this paper can help to monitor not only the platforms' actions against misinformation, but also their subsequent consequences in real-life settings.

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, Sandra Crouse Quinn, and John W Ayers. Twitter and facebook posts about covid-19 are less likely to spread misinformation compared to other health topics. *PLoS ONE*, 17(1), 2022.

[3] Ronan O Fathaigh, Natali Helberger, and Naomi Appelman. The perils of legally defining disinformation. *Internet Policy Review*, 10(4), 2021.

[4] 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.

[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] Béatrice Mazoyer, Julia Cagé, Nicolas Hervé, and Céline Hudelot. A french corpus for event detection on twitter. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6220–6227, 2020.

[11] Andrew J McMinn, Yashar Moshfeghi, and Joemon M Jose. Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 409–418, 2013.

[12] Garret Morrow, Briony Swire-Thompson, Jessica Polny, Matthew Kopec, and John Whihbey. The emerging science of content labeling: Contextualizing social media content moderation. *mimeo*, 2021.

[13] Mohsen Mosleh, Cameron Martel, Dean Eckles, and David Rand. Perverse downstream consequences of debunking: Being corrected by another user for posting false political news increases subsequent sharing of low quality, partisan, and toxic content in a twitter field experiment. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

[14] Irene V. Pasquetto, Briony Swire-Thompson, Michelle A. Amazeen, Fabrício Benevenuto, Nadia M. Brashier, Robert M. Bond, Lia C. Bozarth, Ceren Budak, Ullrich K. H. Ecker, Lisa K. Fazio, Emilio Ferrara, Andrew J. Flanagin, Alessandro Flammini, Deen Freelon, Nir Grinberg, Ralph Hertwig, Kathleen Hall Jamieson, Kenneth Joseph, Jason J. Jones, R. Kelly Garrett, Daniel Kreiss, Shannon McGregor, Jasmine McNealy, Drew Margolin, Alice Marwick, FiIippo Menczer, Miriam J. Metzger, Seungahn Nah, Stephan Lewandowsky, Philipp Lorenz-Spreen, Pablo Ortellado, Gordon Pennycook, Ethan Porter, David G. Rand, Ronald E. Robertson, Francesca Tripodi, Soroush Vosoughi, Chris Vargo, Onur Varol, Brian E. Weeks, John Wihbey, Thomas J. Wood, and Kai-Cheng Yang. Tackling misinformation: What researchers could do with social media data. *Harvard Kennedy School Misinformation Review*, 1(8), December 2020.

[15] Gordon Pennycook, Adam Bear, Evan T Collins, and David G Rand. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, 66(11):4944–4957, 2020.

[16] Gordon Pennycook and David G Rand. The psychology of fake news. *Trends in cognitive sciences*, 2021.

[17] Jessamy Perriam, Andreas Birkbak, and Andy Freeman. Digital methods in a post-api environment. *International Journal of Social Research Methodology*, 2019.

[18] 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.

[19] Richard Rogers. Deplatforming: Following extreme internet celebrities to telegram and alternative social media deplatforming: Following extreme internet celebrities to telegram and alternative social media deplatforming: Following extreme internet celebreties to telegram and alternative social media. *European Journal of Communication*, 35(3):213–229, 2020.

[20] E. Saltz, S. Barari, C.R. Leibowicz, and C. Wardle. Misinformation interventions are common, divisive, and poorly understood. *Harvard Kennedy School (HKS) Misinformation Review*, 2(5), 2021.

[21] Emily Saltz and Claire Leibowicz. Shadow bans, fact-checks, info hubs: The big guide to how platforms are handling misinformation in 2021. *NiemanLab 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.

[22] Filipo Sharevski, Raniem Alsaadi, Peter Jachim, and Emma Pieroni. Misinformation warning labels: Twitter's soft moderation effects on covid-19 vaccine belief echoes. *arXiv preprint arXiv:2104.00779*, 2021.

[23] Héloïse Théro and Emmanuel Vincent. Investigating facebook's interventions against accounts that repeatedly share misinformation. *Information Processing and Management*, 59(2), 2022.

[24] Waheeb Yaqub, Otari Kakhidze, Morgan L Brockman, Nasir Memon, and Sameer Patil. Effects of credibility indicators on social media news sharing intent. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–14, 2020.

[25] Muhsin Yesilada and Stephan Lewandowsky. Systematic review: Youtube recommendations and problematic content. *Internet Policy Review*, 11(1), 2022.

6 Appendix

6.1 Quick access to community guidelines and policies

Rules	f	facebook.com/communitystandards/recentupdates/
	t	help.twitter.com/en/rules-and-policies/twitter-rules
	y	youtube.com/intl/en_us/howyoutubeworks/policies/community-guidelines/
Rules enforcement	f	transparency.fb.com/data/community-standards-enforcement/
	t	transparency.twitter.com/en/reports/rules-enforcement.html
	y	transparencyreport.google.com/youtube-policy/
Transparency center	f	transparency.fb.com/data/ https://law.yale.edu/yls-today/news/facebook-data-transparency-advisory-group-releases-final-report
	t	transparency.twitter.com/en/reports.html
	y	transparencyreport.google.com/?hl=en
Policy regarding Covid-19	f	https://www.facebook.com/help/230764881494641/
	t	help.twitter.com/en/rules-and-policies/medical-misinformation-policy blog.twitter.com/en_us/topics/company/2021/updates-to-our-work-on-covid-19-vaccine-misinformation blog.twitter.com/en_us/topics/company/2020/covid-19
	y	support.google.com/youtube/answer/9891785
Fact-checking policy	f	facebook.com/journalismproject/programs/third-party-fact-checking/how-it-works
	t	x
	y	support.google.com/youtube/answer/9229632
Fighting misinformation	f	facebook.com/formedia/blog/working-to-stop-misinformation-and-false-news about.fb.com/news/2018/05/hard-questions-false-news/
	t	
	y	youtube.com/intl/en_us/howyoutubeworks/our-commitments/fighting-misinformation/#policies blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/
Strike System	f	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?* about.fb.com/news/2018/08/enforcing-our-community-standards/ https://transparency.fb.com/en-gb/enforcement/taking-action/counting-strikes/ (updated July 29,2021)
	t	See in the following links the section: *Account locks and permanent suspension* help.twitter.com/en/rules-and-policies/election-integrity-policy help.twitter.com/en/rules-and-policies/medical-misinformation-policy
	y	https://support.google.com/youtube/answer/2802032?hl=en. Accessed 21 6 2021
Account suspension	f	
	t	https://help.twitter.com/en/managing-your-account/suspended-twitter-accounts
	y	
Flags, Notice and Information Panels	f	https://www.facebook.com/business/help/341102040382165
	t	https://help.twitter.com/en/rules-and-policies/notices-on-twitter
	y	support.google.com/youtube/answer/9004474?hl=en

Table 2: Summary of ressources, last accessed on July 5, 2021.

Acknowledgments
We thank our colleagues at Sciences Po médialab for their feedback and suggestions.

Funding
This research was supported by the Make Our Planet Great Again' French state aid managed by the Agence Nationale de la Recherche under the 'Investissements d'avenir' program with the reference ANR-19-MPGA-0005.

Competing interests
We have no conflicts of interest to disclose.

Ethics
The data collection and processing complied with the EU General Data Protection Regulation (GDPR).