Preliminary Draft

July 15, 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

1 Introduction

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) [2] and Broniatowski et al. (2020) [1]). Yet, this seemingly small subset is generating great concern in traditional media and in society in a broader sense. [1]

Section 230 in the United States Communications Decency Act[2] provides immunity for website platforms against the content created by users. Nevertheless, there is growing pressure for Mainstream Social Media Platforms (hereafter MSMP), such as Facebook, Twitter or Youtube, to moderate the available content. In particular, platforms seem to take explicit actions when content is in violation of local laws in different jurisdictions, e.g. laws regarding defamation of a racial nature, dissemination of symbols from unconstitutional organizations, privacy protection, digital security, electoral laws. Facebook reports having implemented a total of 64.7 thousand content restrictions based on local law across all countries in 2020.[3] Google reports a total of 26 thousand government requests to remove content from July 2020 to December 2020, among which 11.4 thousand concerned Youtube.[4] Twitter reports having received 42.2 thousand legal demands from third-parties from January to June 2020, and has responded by withholding 82 thousand accounts and 3.1 thousand tweets.[5]

Furthermore, MSMP 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

[1] 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."

[2] Similar regulation exists in the European Union, see articles 12 and 15 of the E-commerce Directive (2000).

[3] 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.

[4] See Google's Transparency report, government requests to remove content: transparencyreport.google.com/government-removals/overview.

[5] See Twitter Transparency website, Removal requests: transparency.twitter.com/en/reports/removal-requests.html#2020-jan-jun.

authenticity; which include violence, terrorism, child sexual exploitation, abuse, harassment, hateful conduct, suicide or self-harm, illegal or regulated goods and services, platform manipulation and spam.[6] 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.

In this article, we focus on MSMP's policies and actions regarding content with low credibility or false information, commonly referred to as *Fake News*.[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 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. [8] Facebook use a number of signals and machine learning models to predict misinformation and surface it to fact-checkers.[9] Twitter seems to have a different approach where they focus on providing context rather than fact-checking[10] and the platform is testing a new system based on the wisdom of the crowds to tackle misinformation.[11]

Note, put here something along the following lines: cite misinformation review article about support for fact-checking + hard to correct beliefs once it goes viral. cite Pennycook / Rand? + where to put or cite this ?

During the COVID19 global health pandemic platforms have upgraded their guidelines to include a set of rules to tackle the propagation of potentially harmful content.[12] 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[13], suspending accounts, implementing strike systems, reducing the visibility of

[6] For an exhaustive overview of the community standards of Facebook, see: facebook.com/communitystandards/. For the Twitter Rules see: help.twitter.com/en/rules-and-policies/twitter-rules, and for Youtube community guidelines see: youtube.com/intl/en_us/howyoutubeworks/policies/communityguidelines/.

[7] For an overview on the *concept* of *Fake News*, we refer the reader to the article The science of fake news by Lazer et al. (2018) [4].

[8] For an overview of Facebook's third-party fact checking program see: https://www.facebook.com/journalismproject/programs/third-party-fact-checking/how-it-works.

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

[10] 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.

[11] See Twitter Birdwatch: https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.

[12] For Facebook, Twitter and Youtube see respectively the following updates concerning COVID19 related misinformation policies: https://support.google.com/youtube/answer/9891785, help.twitter.com/en/rules-and-policies/medical-misinformation-policy and support.google.com/youtube/answer/9891785.

[13] put here the list of Facebook and Twitter terms about covid.

	Application Programming Interface (API)	Web Scraping
	CrowdTangle API and Buzzsumo API	
	Twitter API V2	Minet tw scrape
	Youtube API V3	

Table 1: Data collection

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 MSMP 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[14]. The 2019 report of the Facebook Data Transparency Advisory Group[15] (DTAG) states clearly 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.*". Facebook's strategy to tackle "False News" is three fold : Remove, Reduce, Inform. It is explained via a blog post on the facebook Newsroom.[16]. Similarly, Twitter communicates about actions related to misinformation via their Twitter Safety Blog/account[17]. In a post on Youtube Official Blog[18], the platform explained its "Four Rs of Responsibility" and how it raises authoritative content reduces borderline content and harmful misinformation.

Note, put here something along the following line: make point here about the fact that in the transparency centers we find things for the violations in the platforms rules but not misinformation.... needed so that the academic community can study the reach and impact of this phenomena and provide guidelines for adequate actions... + where to put this airtable ?

In the present article, we will explain how to verify with data mining MSMP's actions regarding content with low credibility or false information, through a series of examples for different actions and platforms. For the purpose of clarity, we only 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.[19] We further choose Twitter because it is a social networking platform with the most news-focused users, according to the Pew Research center (2019) [3]. More specifically, we survey a number of common policies used in order to tackle misinformation, across the three above cited MSMP, Facebook, Twitter and Youtube: temporary or permanent suspension of users, reducing the visibility of some content, introducing flags and notices. We do not provide an exhaustive list of methods on how to investigate the platforms' policies. We rather provide a methodology to investigate key policies, that can be useful to researchers or journalists interested in implementing external monitoring. We summarize in table 1 the tools used to collect the data from Facebook, Twitter and Youtube, that we use in multiple examples throughout the present article. Finally, we discuss how an increased effort of transparency regarding specific content can

[14] As of July 2021.

[15] To read the full report, see: https://law.yale.edu/yls-today/news/facebook-data-transparency-advisory-group-releases-final-report?fbclid=IwAR2xMZr5GdD1GaNpjsXR3_yeeIR4H9iFASfrni5HKcJVAO5oWA52bvwcZxU.

[16] See: https://about.fb.com/news/2018/05/hard-questions-false-news/.

[17] See https://blog.twitter.com/en_us/authors.TwitterSafety.

[18] See https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/.

[19] 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/.

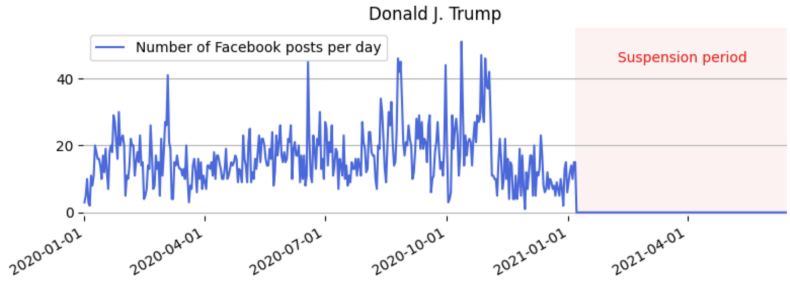

Figure 1: Number of Facebook posts published each day by the Facebook page *Donald J. Trump* between January 1, 2020 and June 15, 2021. The data corresponds to 6 083 posts retrieved from the CrowdTangle API using the *posts* endpoint.

help the community of researchers study and assess the impact of platforms' policies regarding misinformation.

2 Policies

2.1 Temporary suspension and 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.[20] 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.

2.1.1 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.[21] 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.[22] But as long as external persons do not have access to deleted accounts data, these reports cannot be verified by independent researchers or journalists.

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

[20] A list of notable Twitter temporary and permanent suspensions can be found on wikipedia: https://en.wikipedia.org/wiki/Twitter_suspensions.

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

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

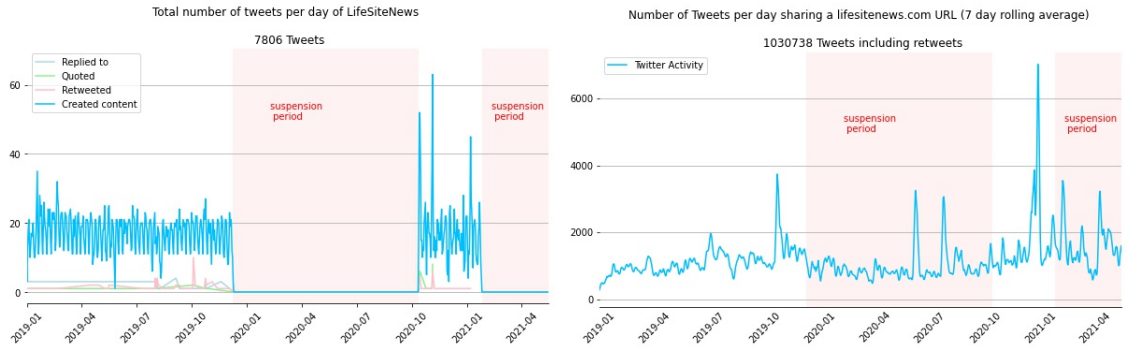

Figure 2: Panel (a): number of Tweets per day of the Twitter account @*Lifesite* linked to the website lifesitenews.com from January, 2019 until April 2021. Panel (b): number of Tweets per day that have shared a lifesitenews.com URL link from January, 2019 until April 2021.

following the Capitol attack on January 6, 2021.[23] 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.[24] We used the minet command line tool [5] to collect data.[25] 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.

2.1.2 Twitter

Twitter has implemented a strike system as part of their Civi Integrity Policy[26] and their COVID19 misleading information policy.[27] 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. 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[28] 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[29]. In particular, this website has several failed fact-checks concerning

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

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

[25] The exact command can be found here.

[26] See help.twitter.com/en/rules-and-policies/election-integrity-policy

[27] See help.twitter.com/en/rules-and-policies/medical-misinformation-policy and blog.twitter.com/en_us/topics/company/2021/updates-to-our-work-on-covid-19-vaccine-misinformation

[28] See the official documentation on the Twitter's Help Center regarding account suspension: https://help.twitter.com/en/managing-your-account/suspended-twitter-accounts .

[29] 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.

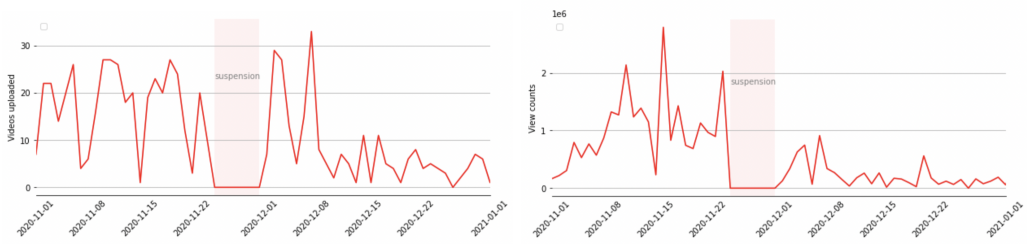

Figure 3: panel (a): Number of Youtube videos uploaded each day by the youtube channel *One America news Network* November 1, 2020 and January 1, 2021. Panel (b): accumulated view counts for videos. The metrics correspond to the videos' publishing date and the data is retrieved from the youtube API with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

the published articles, according to Iffy.news.[30]. We collected the activity (tweets, replies, quotes, retweets) on their Twitter account via the Twitter API, using the historical search endpoint.[31] We then plotted the number of Tweets, Retweets, Quotes and Replies per day, as shown in panel a 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.

To further assess the impact of this double temporary suspension, we collect via Minet Command Line Tool [5], 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.

2.1.3 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.[32] 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.

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

[31] See the documentation: https://developer.twitter.com/en/docs/twitter-api/tweets/search/api-reference/get-tweets-search-all.

[32] See the "Community Guidelines strike basics". Youtube help, Google Developers, https://support.google.com/youtube/answer/2802032?hl=en. Accessed 21 6 2021.

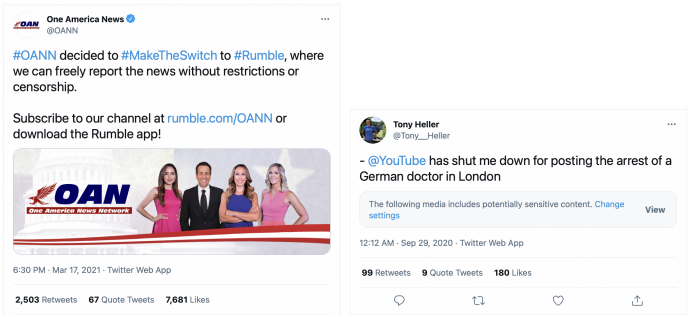

Figure 4: Panel (a): Tweet announcing moving to rumble by OANN (Twitter), Twitter ID 1372238828425998336. Panel (b): Tony Heller's tweet after getting suspended from Youtube, Twitter ID 1310703852769796097.

First, we investigate the temporary suspension case 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 COVID19.[33] 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.[34]

In addition, as shown in figure 3 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 4) 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 4). 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 4 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.

[33] See nbcnews, YouTube suspends OANN for violating its Covid-19 policy nbcnews, Ahiza García-Hodges, 24 11 2020.

[34] See the Google documentation https://developers.google.com/youtube/v3/docs/videos/list and https://developers.google.com/youtube/v3/docs/playlists/list

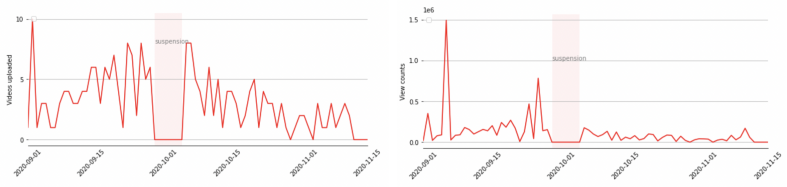

Figure 5: Panel (a): Number of Youtube videos uploaded each day by the Youtube channel *Tony Heller* between September 1, 2020 and November 15, 2020. Panel (b): accumulated view counts for videos uploaded by the same Youtube channel. The date corresponds to the videos' publishing date.

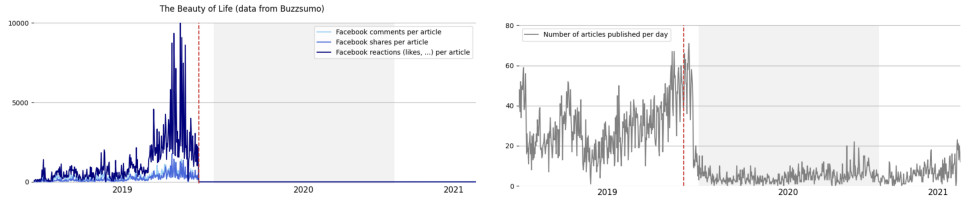

Figure 6: Articles from The Beauty of Life website (thebl.com) published between January 1, 2019 and June 15, 2021 and gathered from the Buzzsumo API. (Top) Facebook engagement metrics (average number of reactions, shares and comments per article). (Bottom) Number of articles published per day. The red line marks the date of December 1, 2019.

2.2 Blocking links

Another measure that mainstream social media platforms can apply is to prevent users from sharing specific types of content, in this example URLs coming from a specific domain name.

2.2.1 Facebook

The Beauty of life (thebl.com/) is a US-based media company that shares pro-Trump views and conspiracy theories such as QAnon.[35] Facebook has announced on December 20, 2019 that "The BL is now banned from Facebook" for coordinated inauthentic behavior[36], 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".[37] 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 and other various public articles[38].

[35] See wikipedia article.

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

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

[38] 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/

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[39]. 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.[40]

The number of Facebook reactions, shares and comments dropped to zero for TheBL's articles published after December 1, 2019 (see figure 6 top panel), indicating the start of the ban. We can note that although the ban was communicated in an article[41] published on December 20, 2019, it 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 6 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.

2.2.2 Twitter

link

2.3 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.

2.3.1 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.[42] This measure can be verified by looking at the number of views (reach) of a post, but this metric is not available via the CrowdTangle API or on 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 via the CrowdTangle API and Buzzsumo. If a post reaches less users because it has a

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

[40] The command can be found in the following Github repository: https://github.com/medialab/truth-and-trust-online-2021/blob/master/code/collect_facebook_buzzsumo_thebl_data.py.

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

[42] Lyons, T. (2018, May 22). The three-part recipe for cleaning up your news feed. Facebook Newsroom. about.fb.com/news/2018/05/inside-feed-reduce-remove-inform/.

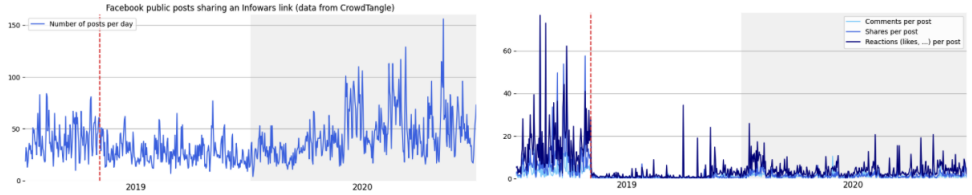

Figure 7: Public Facebook posts sharing an Infowars link in 2019 and 2020 collected from the CrowdTangle API. The red line marks the date of May 2, 2019, when Facebook has announced the ban regarding Infowars. (Top panel) Number of daily posts. (Bottom panel) Engagement metrics: average number of reactions, shares and comments per post.

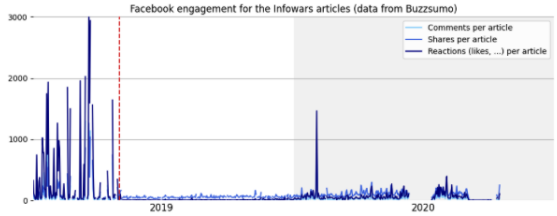

Figure 8: Facebook engagement metrics (average number of reactions, shares and comments per article) for the Infowars articles published in 2019 – 2020 and gathered from the Buzzsumo API. The red line marks the date of May 2, 2019, when Facebook has announced the ban regarding Infowars.

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[43]. Furthermore, the factual reporting of $Infowars$ has been rated *very low* by the Media Bias / Fact Check resource of Iffy.news.[44]

On May 2, 2019, Facebook announced they would prohibit users from sharing Infowars content unless, they are explicitly condemning the material.[45] To verify the measure, we used the "/posts/search" endpoint[46] 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.[47] The command used can be found in the following Github repository: link. The number of public posts sharing an $Infowars$ link remained globally stable throughout 2019 (see figure 7 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

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

[44]See the Iffy.news page: https://mediabiasfactcheck.com/infowars-alex-jones/.

[45]See https://www.wired.com/story/facebook-bans-alex-jones-extremists/ and about.fb.com/news/2018/08/enforcing-our-community-standards/.

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

[47]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.

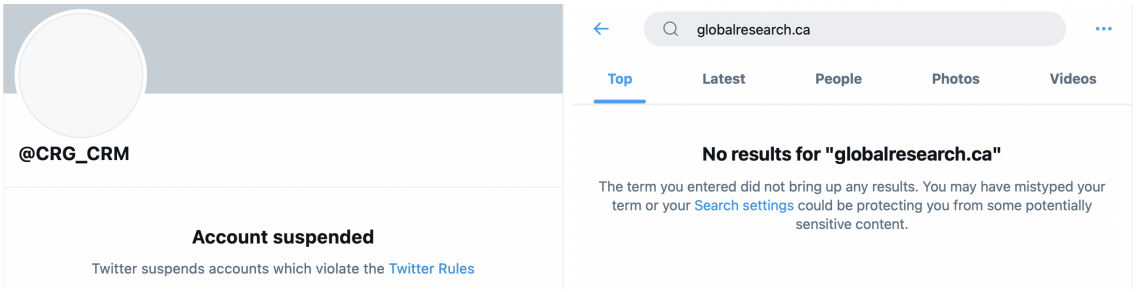

Figure 9: Screenshots taken on June 14, 2021. Top panel: screenshot that shows the account of *globalresearch.ca* suspended on Twitter. Bottom panel: screenshot that shows that no results can be found when searching for *globalresearch.ca*.

May 2, 2019 (see figure 7 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.

As CrowdTangle is tracking posts only from certain public groups and pages, we also used the "/search/articles" endpoint of the Buzzsumo API, to gather a richer Facebook dataset. We collected the engagement data for the 14 232 articles crawled by Buzzsumo from the $Infowars$ website between January 1, 2019 and December 31, 2020. [48] We observe that the articles published after May 2, 2019 received less Facebook engagement than the ones published before (see figure 8), with a percentage change of -97% for the reactions, -59% for the shares and -97% for the comments. An increase in engagement was also observed in 2020. It reinforces the hypothesis that Facebook reduced the reach of posts sharing Infowars links only during a few months in 2019.

2.3.2 Twitter

Twitter can take action against a tweet which violates the Twitter rules[49], by limiting its visibility 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 iffy.news - a website which provides a database of websites with low factual reporting levels.[50]

The website *globalresearch.ca* is linked to the Twitter account $@CRG\_CRM$; which was recently suspended.[51] When a user searches via the twitter search-box for any URL link of this

[48] The command can be found: here.

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

[50] For *globalresearch.ca* see https://mediabiasfactcheck.com/global-research/.

[51] We noticed the message about the account suspension on May 25, 2021. But 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 anytime between April 15, 2021 and May 25, 2021 (see the suspension screenshot in panel (*a*) of figure 9). Furthermore, *globalresearch.ca* has multiple failed fact-checks, see https://mediabiasfactcheck.com/global-research/

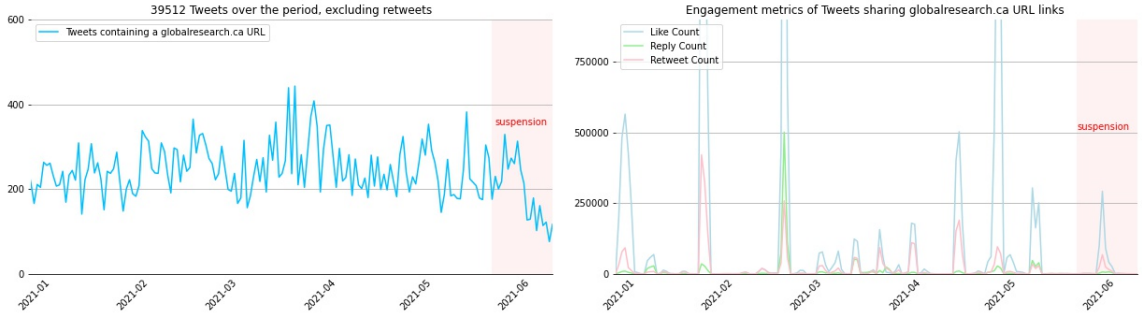

Figure 10: Top panel: daily number of Tweets, excluding retweets, containing the query *globalresearch.ca* from January 1, 2021 until June 10, 2021. Bottom panel: engagement metrics of Tweets containing the query *globalresearch.ca* from January 1, 2021 until June 10, 2021. Data collected via the Twitter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

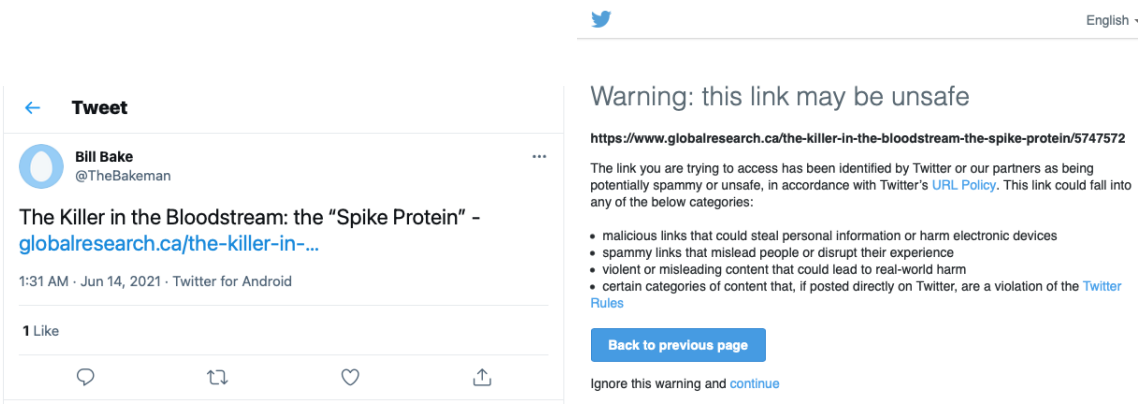

Figure 11

website, no results appear as shown in the screenshot in panel (*b*) of figure 9, taken on June 14, 2021. To further investigate the possible 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 panel (a) in figure 10, we find a strictly positive number of tweets containing the URL link *globalresearch.ca* throughout May 2021 and the first week of June 2021. Hence, the visibility of tweets containing this URL link has been reduced because users can no longer access tweets containing the URL link *globalresearch.ca* via the search box. Nevertheless users are not restrained from posting tweets containing this URL, as shown in the screenshot in panel (*a*) of figure 11, 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 10. Hence, the users who tweet articles from the *globalresearch.ca* website receive tweet level engagement from their own followers. Finally, when a user attempts to click on the URL link *globalresearch.ca* contained in the Tweet, a warning message appears and indicates that the link may be unsafe (see screenshot in panel (b) in figure 11).

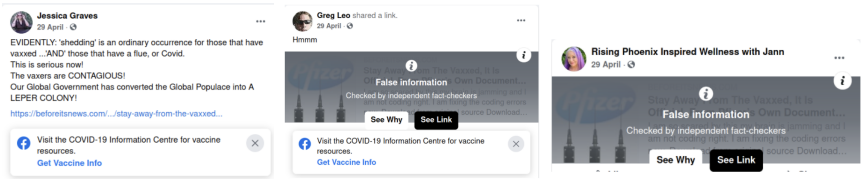

Figure 12: Examples of the Facebook posts having shared a link fact-checked False by one of Facebook's partners (Screenshots taken on July 8, 2021: https://www.facebook.com/groups/1220117708132394/permalink/2386760061468147, https://www.facebook.com/groups/473809623000471/permalink/1333200093728082, https://www.facebook.com/691911990845585/posts/3835629366473816).

2.3.3 Youtube

What about something about recommendations ? authoritative content

2.4 Flags, Notices and labels

2.4.1 Facebook

To the best of our knowledge, two types of flags can currently be seen on Facebook posts, videos and pictures: 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 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 12). 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.

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[52]. 20 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. 3 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.

Surprinsingly only 11 posts out of the remaining 17 had a 'False information' flag (see Table 2). 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 12 for examples). As the 'False information' flag is applied on the link banner,

[52]link: https://beforeitsnews.com/eu/2021/04/stay-away-from-the-vaxxed-it-is-official-from-pfizers-own-documents-2671454.html and its fact-check: https://healthfeedback.org/claimreview/insufficient-evidence-to-claim-covid-19-vaccines-cause-menstrual-irregularities-in-vaccinated-women-vaccinated-people-arent-making-unvaccinated-people-ill/).

Number of posts			
with no 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

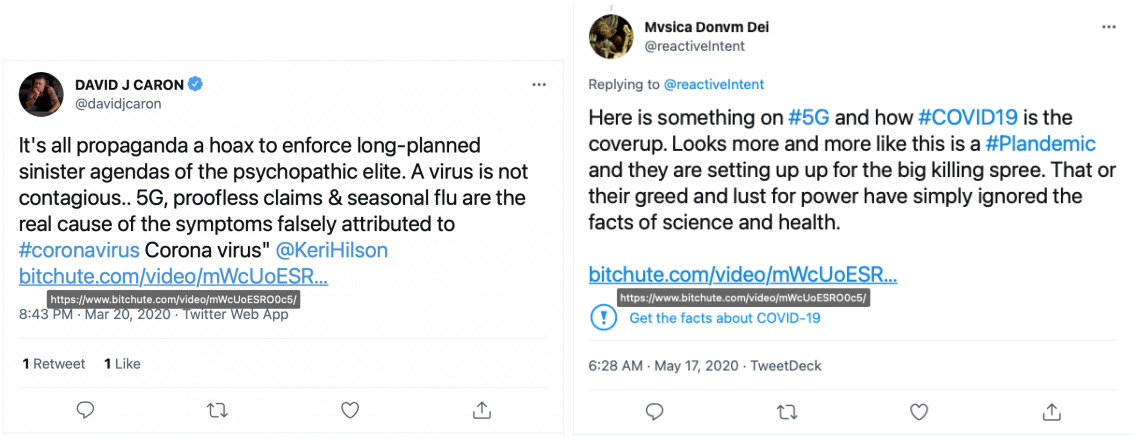

Figure 13: Two tweets sharing the same URL link marked as False by a Fact-Checker, screenshots taken on June 20, 2021. Panel (a): Tweet ID 1241088065462026242 without a label. Panel (b): Tweet ID 1261876171584745472 containing a label.

and not on the link itself, a user could share a link without the 'False information' banner as in the first example of Figure 12.

We also observed that most of the posts (13 out of 17, see Table 2) had an information banner saying 'Visit the COVID-19 Information Centre for vaccine resources' (see the first two examples of Figure 12). 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 right panel of Figure 12.

2.4.2 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.[53] At the tweet level, notices take the form of a label or an interstitial. Labels are context specific (e.g. COVID19 or presidential elections) and redirect users to a URL link to get more context (see figure 13 for an example). 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 (see figure 14 for an example). At the account level, notices can indicate whether an account has been temporarily or permanently suspended.

[53]See Notices on Twitter and what they mean: https://help.twitter.com/en/rules-and-policies/notices-on-twitter.

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[54] from the API. Hence, in order to investigate the presence of labels, 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 URL links of articles which were marked as $False$ 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 [5] all the tweets that have shared a URL 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 URL link, among the set of collected tweets. More specifically, exactly 657 tweets had shared a URL 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 13). This points towards the non-automation of the tweet labelling process and that it might be that only 3 tweets got labelled after being reported by a user.

We further examine the placement of interstitials that indicate a possibly sensitive content. We find that only 2.97% out of 323 938 tweets containing a URL marked as False, have an interstitial "potentially sensitive content". In particular, many speak about COVID19 and do not contain a label to provide users with more context from authoritative sources. Figure 14 provides an example of a tweet sharing a URL marked as False by a Fact-checker and which contains an interstitial "potentially sensitive content". We find 32 other Tweets, among our set of collected tweets, who share the exact same URL link as in the previous example. Among those 32 Tweets, only 5 tweets had the interstitial "potentially sensitive content". Again this points towards the non-automation of the interstitials placement.

Finally, the Twitter Safety announced in a Tweet[55] 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".

2.4.3 Youtube

Youtube may provide an information panel for videos with topics that are prone to misinformation like COVID19, moon landing, and climate change.[56] Information panels provide ressources concerning a potentially controversial topic from independent third party partners 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.

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

[55] See Tweet ID : 1379515615954620418

[56] See the section "Information panel giving topical context" on Youtube Help, Google accessed on June 28, 2021: support.google.com/youtube/answer/9004474?hl=en.

Figure 14: Panel (a): Tweet containing a URL link hidden behind an interstitial. Panel (b): when clicking on *view*, to view the content hidden behind the interstitial. Tweet ID 1285866521533861888. Screenshots taken on July 1, 2021.

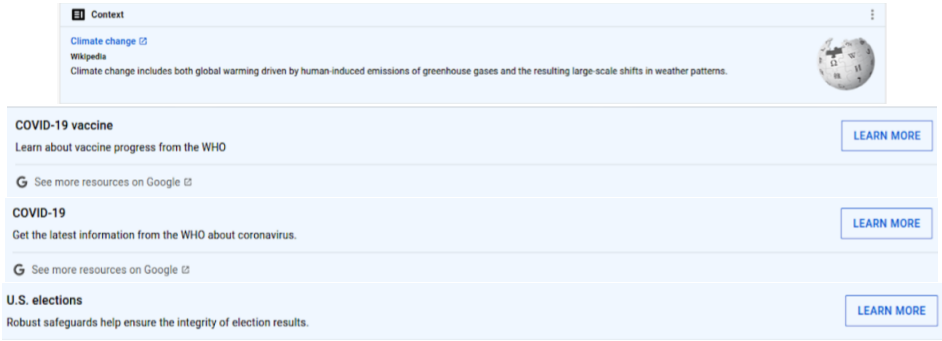

Figure 15: Top to buttom respectively: information panel displayed under some videos related to climate change, COVID19 vaccine, COVID19 and elections.

Category	Number of Youtube videos with an information panel	Number of Youtube videos without an information panel
General Health	0 (0%)	11 (100%)
Vaccine	21 (75%)	7 (25%)
COVID19	73 (63%)	43 (37%)
Climate change	5 (28%)	13 (72%)

Table 3

To further study the assignment of information panels to videos, we implemented the following exercice. We compiled a list of 171 Youtube videos, checked by independent fact-checkers[57] and marked as containing false or inaccurate information. We collected the information panels, when present, in each video by scrapping the content of the web and connecting to a US server. Four types of information panels were found in the list of visited videos as shown in figure15: information about COVID19, Vaccines, Climate change and US elections.

We classified the list of Youtube videos by topic based on its content as shown in table 3. We recorded the number of videos containing an information panel within each category. For COVID19 and vaccine related videos more than half of the videos contained an information panel. Nevertheless, we noticed that duplicates of the same Youtube video can get uploaded under different video titles and that the information panel appears under the some duplicates, but not for all. In addition, we noticed for COVID19 related videos, when the video title did not include keywords like (Testing, Pandemic, COVID, coronavirus), the video might 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 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 since many of the videos that didn't have a panel had the word COVID or coronavirus or vaccine mentioned in the video itself.

For climate change we only examined 18 misleading videos and we found that 28% did not display an information panel. Lastly , for the general health category, we included the videos that were not related to COVID19 but contained misinformation concerning cures for cancer, abortion, and viruses; no information panels were displayed for that set of Youtube videos. Therefore, it is likely that Youtube add information panels below videos concerning controversial topics that can have misleading information like the climate change, COVID19, flat earth and vaccine.

3 Discussion

Sporadic points for the discussion, no structure yet.

[57] Is it science Feedback ? if so add.

- For a previous research project[58], 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 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...

[58] reference?

- 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.

References

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.

4 Appendix

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		https://transparency.fb.com/data/
		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 blog.twitter.com/en_us/topics/company/2021/updates-to-our-work-on-covid-19-vaccine-misinf 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 about.fb.com/news/2018/05/hard-questions-false-news/
		youtube.com/intl/en_us/howyoutubeworks/our-commitments/fighting-misinformation/#p blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/

Table 4: Summary of ressources, last accessed on July 5, 2021.