Investigating Facebook's policies to tackle misinformation during the COVID-19 pandemic

Héloïse Théro
heloise.thero@sciencespo.fr
médialab - Sciences Po
Paris, France

Emmanuel Vincent
emmanuel.vincent@sciencespo.fr
médialab - Sciences Po
Paris, France

ABSTRACT

Using a dataset of URLs marked as 'False' by Science Feedback, an organization verifying the credibility of science-related viral information, we investigated the reach of groups and pages sharing misinformation on Facebook during 2019 and 2020, and Facebook's actions to curb its spread. We found that, consistently with what the company publicly announced, Facebook removed large QAnon conspiracy accounts in August 2020, which were instrumental in spreading science misinformation. We also investigated Facebook's 'repeat offenders' policy and found that only some accounts that repeatedly share misinformation display periods in which the popularity of their posts was temporarily reduced. Despite these measures, we have witnessed that most Facebook accounts spreading scientific misinformation have increased their reach at the beginning of the COVID-19 pandemic (March - June 2020), and that a drastic drop in their reach suddenly occurred around the 9th of June, 2020. Meanwhile the reach of a set of mainstream news accounts remained stable across the entire period. No public information was given by Facebook about this sudden decrease.

KEYWORDS

Disinformation, Fact-checking, Social media analysis, Algorithmic bias and transparency

ACM Reference Format:
Héloïse Théro and Emmanuel Vincent. 2021. Investigating Facebook's policies to tackle misinformation during the COVID-19 pandemic. In *The Web Conference, April 19–23, 2021.* ACM, New York, NY, USA, 6 pages.

1 INTRODUCTION

While the new coronavirus keeps propagating, so are misleading or false information regarding the pandemic. It has led the director of the World Health Organization to declare: "we're not just fighting an epidemic; we're fighting an infodemic." [10] The term 'infodemic' has been used to express the threat related to the massive spread of misinformation, as it can undermine public health measures to contain the spread of the virus [23].

Today, an increasing proportion of the public get their information online [16], mainly through search engines, social media and video platforms. In April 2020, a questionnaire from the Reuters Institute in the UK found that people use more online than offline sources when looking for information about the coronavirus. Among social media platforms, Facebook was the most widely used with 24% of the respondents saying they used Facebook to access COVID-19 information in the last 7 days [9]. The predominance of Facebook in the media landscape is confirmed by Parse.ly's dashboard, showing that the visitors to their 2500+ online media sites are referred by Facebook in 26% of the cases (only Google had a higher referral volume with 52% of the traffic) [5].

These platforms have come under heavy criticism over the past few years for enabling the circulation of misinformation on massive scales [14, 18]. Facebook has publicly announced a three-part policy to fight against 'misleading or harmful content' : they *remove* harmful information, *reduce* the spread of misinformation and *inform* people with additional context [15]. Facebook is indeed communicating regularly about deleting accounts spreading hate speech or incitation to violence, and about promoting official sources of information about elections, health or climate change [1, 6, 11, 12].

Regarding how they handle misinformation, Facebook has announced more specific measures:

- first to identify potential false news and seek the expertise of fact-checkers to review content and clearly label misinformation, and
- second to ensure that fewer people see misinformation, and to take action against repeat offenders by reducing their distribution [3].

The application of the first measure can be transparently verified on Facebook, with some posts being publicly labeled as misinformation. However, there is little data available to measure the efficiency and track the impact of the latter measure.

A study analysed the reach of a set of websites identified as sources of false stories on Facebook and Twitter from January 2015 to July 2018. They found that during the 2016 American elections, total engagement on Facebook and Twitter for these sites had more than doubled compared to pre-election levels. Following the election, however, Facebook engagements fell sharply, while Twitter shares continued to increase for these sites, suggesting that Facebook might have taken measures to contain misinformation [8].

A more recent article measured the level of interactions on Facebook with articles from outlets that repeatedly publish false content from 2016 to 2020, and found results contrasting with Allcott and colleagues'. Although they did observe a decrease in the first and second quarter of 2017, interactions with 'deceptive' outlets have increased since, and are now 242 percent higher on Facebook than during the run-up to the 2016 election [13].

The Web Conference, April 19–23, 2021, Ljubljana, Slovenia

Table 1: List of QAnon related groups or pages that have shared 10 misinformation links or more found in the data collected on June 2 and August 31, 2020, with their number of followers in parentheses. The number of followers was retrieved at the time of the CrowdTangle request, which explains the differences between the June and August columns.

QAnon related accounts 10+ misinformation links June 2, 2020	QAnon related accounts 10+ misinformation links August 31, 2020
OFFICIAL Q / QANON (167,223)	Q The Greatest Story Ever Told (27,384)
We Are Q (44,838)	Real Qanon Follow The White Rabbit (16,495)
QAnon -Posts by Q (35,821)	Q ANON QUÉBEC (12,008)
WWG1 WGA...Q (29,977)	QAnon Arts & Info Pub WWG1WGA (2,162)
QANON PIZZAGATE FULL DISCLOSURE GROUP (28,102)	Qanon Suomi (2,144)
Q The Greatest Story Ever Told (23,539)	
Q The Great Awakening (19,462)	
Real Qanon Follow The White Rabbit (13,417)	
Q ANON QUÉBEC (10,958)	
Truth Bomb News! #Qanon #The Great Awakening (9,278)	
MemeLab / QAnon / Q-Map (9,168)	
QAnon 8ch Uncensored Research (7,727)	
12 accounts **399,510 total followers**	**5 accounts** **60,193 total followers**

Our analysis is based on Science Feedback's fact-checking dataset. Science Feedback is a scientific organization verifying the credibility of science-related claims and articles for a general audience. The organization tracks the most influential press articles or social media posts, invites scientists with domain expertise to evaluate their credibility and is one of Facebook's third-party fact-checkers [22]. We used the URLs marked as 'False' by Science Feedback (misinformation links) to investigate their reach on Facebook and Facebook's efforts to reduce the distribution of 'repeat offenders'.

2 FACEBOOK REMOVING QANON ACCOUNTS, POTENT MISINFORMATION SPREADERS

On August 19, 2020, Facebook publicly announced having removed over 790 groups and 100 pages tied to QAnon, a growing right-wing conspiracy theory, for "hav[ing] demonstrated significant risks to public safety" [7], consistently with Facebook's Dangerous Individuals and Organizations policy. Twitter and Youtube also announced to take action on activity associated with QAnon [20, 21]. In this section, we verify the impact of Facebook's policy in our dataset.

To that end, we rely on the 1,290 URLs labeled as 'False' by Science Feedback (i.e., we excluded the URLs labeled as 'Partly False', 'Missing Context', 'False headlines' or 'True'). The list was obtained on June 2, 2020. From the CrowdTangle API (using the `/links` endpoint) and using the `minet` library [4], we gathered the Facebook pages and groups that publicly shared at least one 'False' URL over a 12 months period (between June 2, 2019 and June 1, 2020). Due to the API limitations at the time, if a URL was shared in more than 100 posts, we collected only the 100 posts that received the highest number of interactions.

On August 27, 2020, we repeated the above operation with the then 1,691 URLs labeled 'False' collecting all the Facebook accounts that shared at least one of the links over a 12 months period (between September 1, 2019 and August 31, 2020).

Table 2: Number of QAnon related accounts that have shared 3 misinformation links or more and their total follower number, found in the data collected on June 2 and August 31, 2020.

QAnon related accounts 3+ misinformation links June 2, 2020	QAnon related accounts 3+ misinformation links August 31, 2020
35 accounts **723,742 total followers**	**46 accounts** **371,288 total followers**

We first filtered the Facebook groups and pages to keep only the ones that shared at least 10 different 'False' URLs (misinformation links), and manually identified the groups and pages linked to the QAnon theory. In the dataset collected in June, we found 12 accounts whose name was associated with Q or QAnon among the total of 204 accounts having shared 10 misinformation links or more. In the dataset collected in August, we found only 5 accounts linked to the conspiracy theory among the total of 185 accounts (see Table 1 for the full lists of names). The total number of followers of these accounts went from almost 400,000 in June to only 60,000 in August.

From the lists shown in Table 1, we can see that all the QAnon related groups can be automatically identified by searching groups containing the letter 'Q' in their name. To capture a broader set of accounts, we ran a 'Q' search on all the accounts having shared at least 3 misinformation links on June (1,384 accounts) and August (1,670 accounts). The list was then manually sorted to remove accounts with a 'Q' in their name that were not related to the QAnon theory (such as 'AMERICANS Against Excessive Quarantine!' or 'Info Pro-Trump Québec').

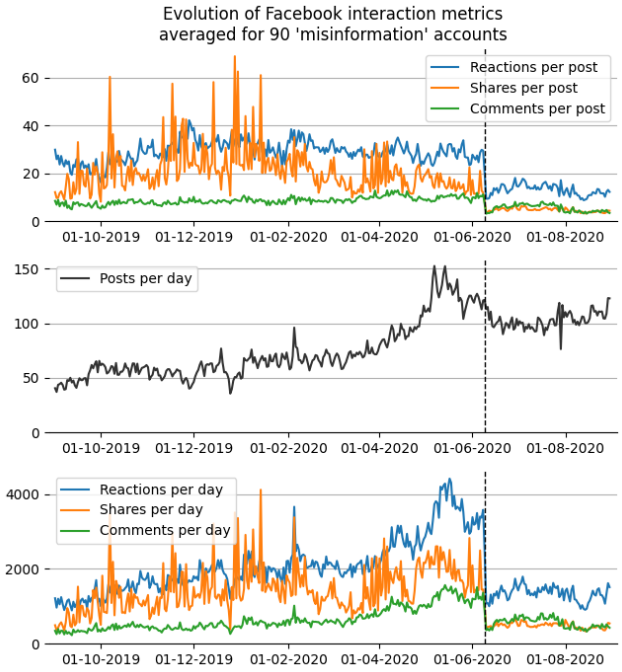

Figure 1: Temporal evolution of Facebook interaction metrics averaged for 75 Facebook groups and 15 pages repeatedly sharing 'False' URLs. The dotted line marks the date of June 9, 2020, when a sudden drop in reactions, shares and comments can be observed.

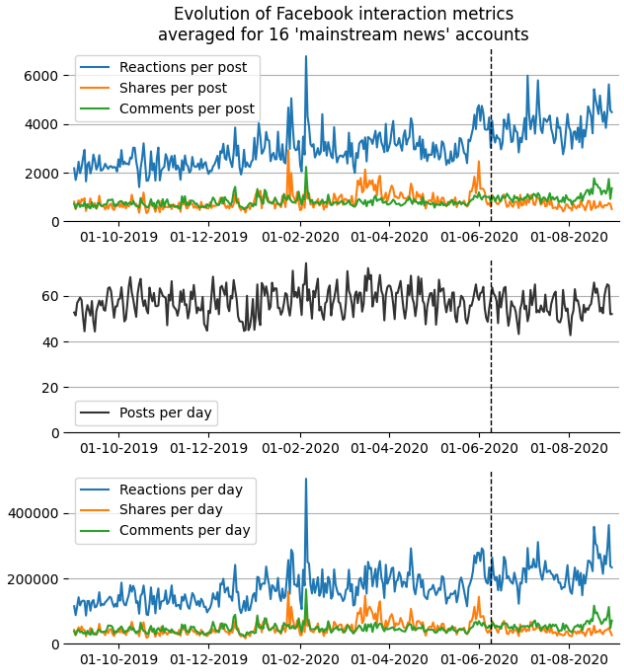

Figure 2: Temporal evolution of Facebook interaction metrics averaged for 6 Facebook groups and 10 pages associated with 'mainstream news' outlets. The dotted line marks the date of June 9, 2020.

For this broader set of Facebook accounts, the number of QAnon related accounts has actually increased from 35 to 46 between June and August 2020, but their total follower count has decreased by half (Table 2). This suggests that Facebook has mainly removed popular QAnon accounts, while many new accounts have been created over the three months period.

Our results confirmed that Facebook suppressed many popular accounts linked to QAnon, some of which have been instrumental in spreading misinformation. Despite this policy, we still found 46 QAnon groups sharing 3 misinformation links or more after August 19, 2020. In a recent update to their official communication [7], Facebook announced they would act more firmly: "Starting today [October 6, 2020], we will remove any Facebook Pages, Groups and Instagram accounts representing QAnon, even if they contain no violent content." Further data collection will be needed to track the impact of this stated policy.

3 A SUDDEN DROP IN THE REACH OF MOST MISINFORMATION ACCOUNTS

In this section, we turn to the overall evolution in the interactions generated by posts in groups repeatedly sharing misinformation during the COVID-19 pandemic.

To that end, we focused on the 90 Facebook accounts (75 groups and 15 pages) that have shared 15 or more different misinformation links in the data collected in August. From the CrowdTangle API (using the `/posts` endpoint), we collected all the posts they published between September 1, 2019 and August 31, 2020. The collection was performed between the September 1 and 7, 2020. We computed for the 90 accounts their mean daily number of posts, as well as the mean daily number of comments, shares and reactions (aggregating the 'like', 'love', 'favorite', 'ahah', 'wow', 'sad' and 'angry' reactions) per post (Figure 1).

During the period from March to June 2020, we observe a doubling in the number of daily posts, from 60 to a peak at 120 posts per day (Figure 1 middle panel), while the number of reactions, shares and comments per post vary both upwards and downwards alternatively (Figure 1 top panel). As a result, the number of reactions, comments and shares per day has risen from March to June 2020 (Figure 1 bottom panel), leading to an increase in total engagement for these accounts.

Surprisingly, we observe a sudden and large decrease in all three engagement metrics around June 9, 2020, with a decrease of 42% in the number of reactions per post, of 50% in the number of shares and of 46% for comments between June 8 and 10, 2020.

To verify whether this drop could be driven by the sudden disappearance of one (or many) popular group, we computed the same metrics aggregated over the 64 Facebook accounts that did not either appear or disappear between September 1, 2019 and August 31, 2020. We still observed a drop in engagement at the same date with the number of reactions per post declining by 42%, by 46%

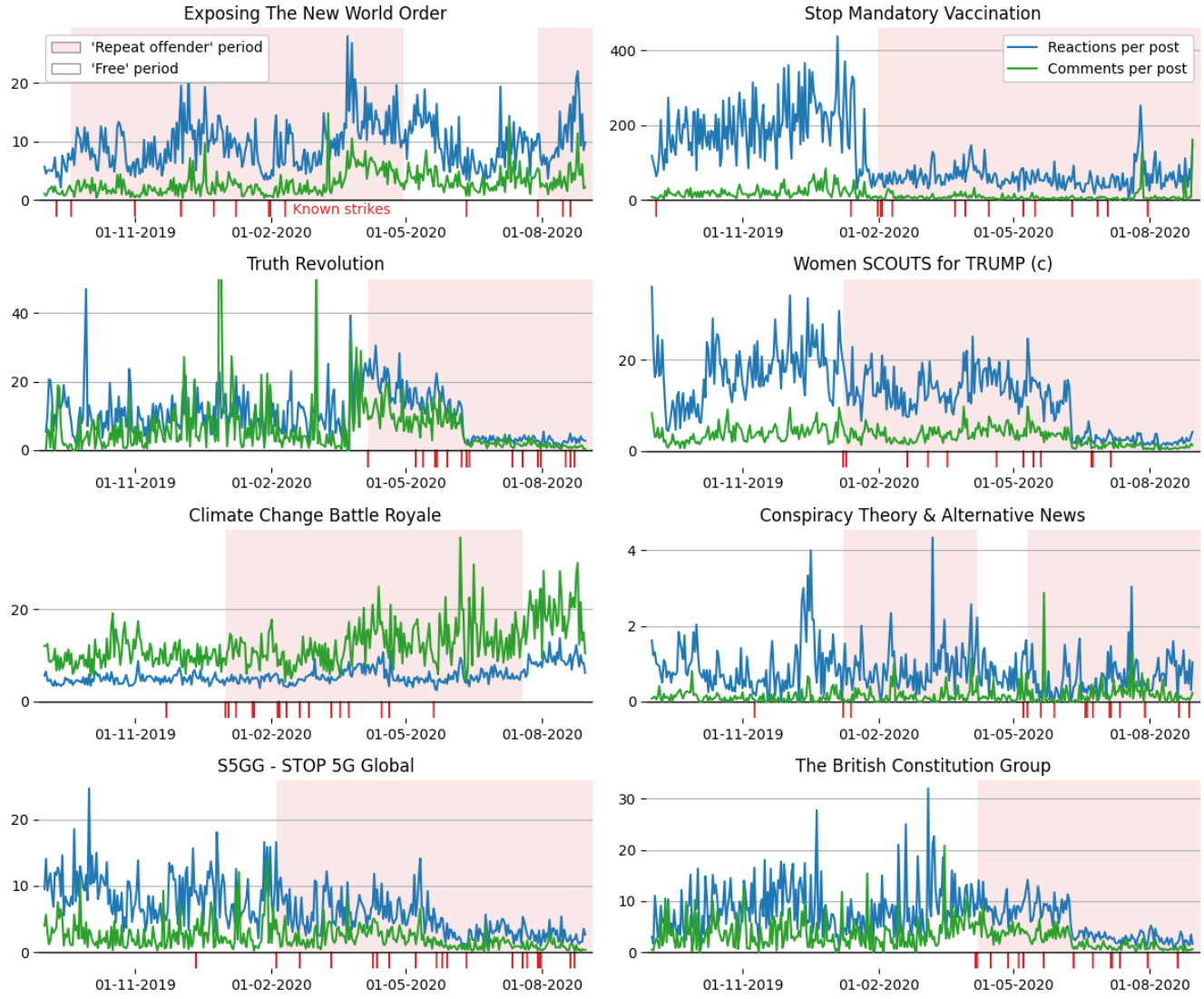

Figure 3: Mean number of reactions and comments per post over the past year for 8 Facebook accounts among the 90 spreading misinformation repeatedly. Each red line at the bottom of a subplot represents the date of a known strike (a shared link labelled 'False' by Science Feedback), and the areas shaded in red represent the 'repeat offender' periods as defined by the '2 strikes in less than 90 days' rule. The areas not shaded in red thus correspond to 'free' periods (i.e. not 'repeat offender' periods) for these accounts, as far as we know.

for the shares and 46% for the comments. We can thus reject this hypothesis.

To compare the Facebook pages and groups sharing misinformation repeatedly with a control set of accounts, we gathered Facebook pages and groups associated with 'mainstream news' outlets. A report from NewsWhip [2] was used to identify the 10 media that communicated the most about COVID-19 online, i.e. NBC, The Daily Mail, CNN, Fox News, The Independent, BBC, The New York Times, The Washington Post, Yahoo and The New York Post. We searched on Facebook for the name of the outlets, and listed 10 pages and 6 groups with the verified 'blue check'. Using CrowdTangle, we collected all the posts published by these 16 accounts between September 1, 2019 and August 31, 2020. The collection was performed on the September 7, 2020.

Figure 2 displays the same statistics as for the 'misinformation accounts'. Contrary to what we observe for the 'misinformation accounts', there is no drop in reach around June 9, 2020 (dotted line on Figure 2) for 'mainstream news accounts'. As for trends, their daily number of posts remained stable throughout the period, while the number of engagements per post increased by roughly 100%. As a result, the total engagement for these accounts increased with the

same proportion. We note that a post on a mainstream news account is generating about two orders of magnitude more reactions, shares and comments per day than a post on a misinformation account.

To investigate whether the 'June drop' could be explained by Facebook taking action against accounts that repeatedly shared misinformation, we correlated the magnitude of the drop with the number of misinformation links shared by a page and contrast this with other account characteristics such as the number of followers and the mean engagement (reactions + comments + shares) per post. The drop was measured for each account by the negative growth rate in the sum of reactions, comments and shares between June 8 and 10, 2020. We found no correlation between the drop and the number of followers (Pearson correlation coefficient: $r = 0.11$), the mean engagement per post ($r = 0.09$) or the number of shared misinformation links ($r = 0.16$). We thus cannot provide an explanation for how Facebook selected the reduced reach groups.

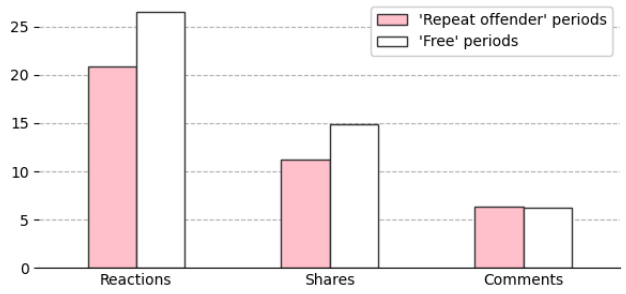

Figure 4: Comparison between the mean number of reactions, shares and comments per post during the 'repeat offender' periods and the rest of the time (called here 'free' periods.

4 INVESTIGATING THE REACH OF ACCOUNTS SHARING MISINFORMATION

According to Facebook, *"Pages and websites that repeatedly share misinformation rated False or Altered will have some restrictions, including having their distribution reduced."* [3]. Accounts repeatedly sharing misinformation are labeled 'repeat offenders', but in the official communication it is not clear whether a group is classified as such after sharing 2 misinformation links or more. A non-official source precised that "The company operates on a 'strike' basis, meaning a page can post inaccurate information and receive a one-strike warning before the platform takes action. Two strikes in 90 days places an account into 'repeat offender' status" [19].

To verify this policy, we first computed the daily number of reactions and comments per post for each of the 90 accounts spreading the most misinformation (Figure 3). The number of shares per post was excluded as it was more prone to extreme values, and thus less reliable to spot reduced reach periods. The 'June' drop was apparent for some accounts such as 'Truth Revolution', 'Women SCOUTS for TRUMP (c)' or 'The British Consitution Group'. (Figure 3).

We computed the 'repeat offender' periods for each group based on the list of strike dates known by Science Feedback. If the post sharing a 'False' URL got published after the fact-check for the corresponding link, we used the date of the post as the date of the strike. In the case when the account first shared a link, which was then fact-checked as 'False', the fact-check publication date was used as the strike date. The strike dates are shown as red lines on Figure 3. From the list of strike dates, we computed the 'repeat offender' periods for each account using the following rule: at any given time, if an account has shared two or more misinformation links over the past 90 days, it was defined as a 'repeat offender' (periods shaded in red on Figure 3).

We can see that most accounts display stable or even increasing interaction metrics during the periods shaded in red, although one group ('Stop Mandatory Vaccination') has a decrease in the number of reactions and comments per post that could coincide with a 'repeat offender' period.

We then compared the average number of reactions, comments and shares per post during the 'repeat offender' periods with the same metrics during the 'free' periods (Figure 4). To keep only representative data, we kept the accounts that have published at least 100 posts during the 'repeat offender' period and 100 posts during the 'free' periods. We also excluded the interaction metrics after June 9, 2020, as the 'June' drop may bias this comparison.

We found the interaction metrics to be globally similar between the 'repeat offender' and the 'free' periods. Paired Wilcoxon tests revealed no differences to be significant using a 0.01 threshold (Reactions: $t = 1488$, $p = 0.4$; Shares: $t = 1113$, $p = 0.01$; Comments: $t = 1474$, $p = 0.4$). Given information in our possession, we found no evidence that repeatedly sharing misinformation had actual consequences on the accounts' reach.

5 CONCLUSION

Misinformation on social media is a potential threat to the fight against the COVID-19 pandemic. Using a fact-checking organization dataset, we investigated the reach of a set of accounts that repeatedly shared misinformation on Facebook over the past year, and tested whether Facebook's stated policies to tackle misinformation were applied to these accounts.

We compared the temporal evolution of engagement metrics of Facebook accounts spreading misinformation to accounts of 'mainstream news' outlets. While the daily engagement metrics of 'misinformation' accounts doubled during the first half of 2020, mainly because of an increase in posting, the 'mainstream news' accounts' activity remained stable and the total engagement with their posts increased.

This observation draws parallels with another study finding that between 2015 and 2017, Facebook interactions for major and small news sites have remained relatively stable with a modest upward trend, while interactions for websites known to publish false information peaked during the 2016 election [8]. A more recent study showed that during the second quarter of 2020, Facebook interaction levels with articles from all news sites (deceptive and legitimate) have overall increased. However, interactions with articles from 'deceptive sites' remained high during the third quarter of 2020, while interactions with articles from legitimate journalistic outlets fell close to their pre-pandemic levels [13].

Regarding Facebook's actions against misinformation, some of our findings are consistent with Facebook's public announcements:

we did witness that QAnon groups with a large number of followers have disappeared after August 19, 2020 [7], and that certain Facebook accounts instrumental in spreading misinformation had their reach momentarily reduced for several weeks.

However, while Facebook publicly announced to apply reduced reach to 'repeat offenders' [3], we did not observe a reduction in the number of reactions or comments per day for accounts that repeatedly share misinformation links, nor did we observe a correlation between the number of 'False' flag received by a page and the proportion of time spent with reduced reach.

We also observed on June 9, 2020 a large drop in engagement per post metrics for most misinformation accounts, but this 'June drop' does not correspond to any official communication on that matter. Further investigations will be needed to understand the nature and origin of this change.

One limitation of this study is that we only took into account the URLs labelled as 'False' by one fact-checking organization (Science Feedback), while Facebook partners with over 60 fact-checking organizations [17]. Data from other fact-checkers would help to further investigate the relationships between 'False' flags and consequences for a Facebook account.

We hope these results can be useful to those working on addressing online misinformation. The code used to collect and clean the data, and to plot the figures is available on GitHub (the URL will be given later because the name of the GitHub account and of the contributors would impair the double-blind procedure). We have listed the URLs of the Facebook groups and pages collected in the GitHub repository so that the figures could be as easy to reproduce as possible.

ACKNOWLEDGMENTS

To Guillaume Plique, Benjamin Ooghe-Tabanou and all the médialab technical team for their help. This research was supported by the Programme d'Investissements d'Avenir (ANR-19-MPGA-0005).

REFERENCES

[1] 2020. Additional Steps to Protect Myanmar's 2020 Election. (2020). Retrieved October, 2020 from https://about.fb.com/news/2020/09/additional-steps-to-protect-myanmars-2020-election/
[2] 2020. Coverage of the Coronavirus on Web and Social. (2020). Retrieved July, 2020 from https://go.newswhip.com/2020_03_Covid-19_LP.html
[3] 2020. Fact-Checking on Facebook. (2020). Retrieved September, 2020 from https://www.facebook.com/business/help/2593586717571940?locale=en_GB
[4] 2020. Minet: a webmining library and command line tool written in python. (2020). https://github.com/medialab/minet
[5] 2020. Parse.ly's Network Referrer Dashboard. (2020). Retrieved September 30, 2020 from https://www.parse.ly/resources/data-studies/referrer-dashboard
[6] 2020. Stepping Up the Fight Against Climate Change. (2020). Retrieved October 19, 2020 from https://about.fb.com/news/2020/09/stepping-up-the-fight-against-climate-change/
[7] 2020. An Update to How We Address Movements and Organizations Tied to Violence. (2020). Retrieved October 16, 2020 from https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/
[8] Hunt Allcott, Matthew Gentzkow, and Chuan Yu. 2019. Trends in the diffusion of misinformation on social media. *Research & Politics* 6, 2 (2019), 205316801984855 4.
[9] Richard Fletcher, Antonis Kalogeropoulos, Felix Simon, and Rasmus Kleis Nielsen. 2020. Information inequality in the UK coronavirus communications crisis. (2020). Retrieved October 16, 2020 from https://reutersinstitute.politics.ox.ac.uk/information-inequality-uk-coronavirus-communications-crisis
[10] Tedros Adhanom Ghebreyesus. 2020. Munich Security Conference. (2020). Retrieved October 16, 2020 from https://www.who.int/dg/speeches/detail/munich-security-conference
[11] Nathaniel Gleicher. 2020. Removing Coordinated Inauthentic Behavior. (2020). Retrieved October, 2020 from https://about.fb.com/news/2020/07/removing-political-coordinated-inauthentic-behavior/
[12] Kang-Xing Jin. 2020. Keeping People Safe and Informed About the Coronavirus. (2020). Retrieved October 19, 2020 from https://about.fb.com/news/2020/10/coronavirus/
[13] Karen Kornbluh, Adrienne Goldstein, and Eli Weiner. 2020. New Study by Digital New Deal Finds Engagement with Deceptive Outlets Higher on Facebook Today Than Run-up to 2016 Election. (2020). Retrieved October 17, 2020 from https://www.gmfus.org/blog/2020/10/12/new-study-digital-new-deal-finds-engagement-deceptive-outlets-higher-facebook-today
[14] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.
[15] Tessa Lyons. 2018. The Three-Part Recipe for Cleaning up Your News Feed. (2018). Retrieved July, 2020 from https://about.fb.com/news/2018/05/inside-feed-reduce-remove-inform/
[16] Amy Mitchell, Jeffrey Gottfried, Michael Barthel, and Elisa Shearer. 2016. Pathways to news. (2016). Retrieved October 16, 2020 from https://www.journalism.org/2016/07/07/pathways-to-news/
[17] Guy Rosen. 2020. An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19. (2020). Retrieved September, 2020 from https://about.fb.com/news/2020/04/covid-19-misinfo-update/
[18] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications* 9, 1 (2018), 1–9.
[19] Olivia Solon. 2020. Sensitive to claims of bias, Facebook relaxed misinformation rules for conservative pages. (2020). Retrieved September, 2020 from https://www.nbcnews.com/tech/tech-news/sensitive-claims-bias-facebook-relaxed-misinformation-rules-conservative-pages-n1236182
[20] The YouTube Team. 2020. Managing harmful conspiracy theories on YouTube. (2020). Retrieved October 19, 2020 from https://blog.youtube/news-and-events/harmful-conspiracy-theories-youtube/
[21] TwitterSafety. 2020. (2020). Retrieved October 19, 2020 from https://twitter.com/TwitterSafety/status/1285726277719199746
[22] Emmanuel M. Vincent. 2020. Science Feedback partnering with Facebook in fight against misinformation. (2020). Retrieved October, 2020 from https://sciencefeedback.co/science-feedback-partnering-with-facebook-in-fight-against-misinformation/
[23] John Zarocostas. 2020. How to fight an infodemic. *The Lancet* 395, 10225 (2020), 676.