

Persistence of online misinformation: Investigating Facebook’s actions against “repeat offenders”

Anonymous TTO submission

Abstract

Like most web platforms, Facebook is under pressure to regulate misinformation. According to the company, pages that repeatedly share misinformation (“repeat offenders”) will have their distribution reduced, but little is known about the implementation or the efficacy of this measure. In a first study, combining data from a fact-checking organization and CrowdTangle, we did not find that Facebook groups repeatedly sharing misinformation had reduced engagement metrics, but a drastic drop was observed around June 9, 2020. No public information was given by Facebook about this sudden decrease. In a second study, we gathered Facebook pages that shared a ‘reduced distribution’ notification from Facebook, and found a decrease in engagement following the notification. Facebook’s reduced distribution policy against repeat offenders thus appears to be applied only on Facebook pages, and not on Facebook groups.

1 Introduction

With an ever-increasing proportion of the public getting their information online, mainly through search engines, social media and video platforms (Mitchell et al., 2016), the spread of misinformation through these platforms has received growing attention. Recent studies and the political context of January 2021 show how the presence of misinformation online can contribute to negative societal consequences by fueling false beliefs, such as the idea that massive voter fraud occurred during the US 2020 presidential election, which contributed to the January 6, 2021 insurrection at the U.S. Capitol (Benkler et al., 2020) and other false stories about presidential candidates (Allcott and Gentzkow, 2017). Misinformation has also confused the public about the reality of climate change (Brulle, 2018; Porter et al.,

2019) and stoked skepticism about vaccine safety among the public (Featherstone and Zhang, 2020; Lahouati et al., 2020). In April 2020, a questionnaire from the Reuters Institute found that people in the UK use online sources more often 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 seven days (Fletcher et al., 2020). The structural importance of Facebook to 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 25% of the cases, second to Google’s referral volume accounting for 54% of traffic¹.

Lawmakers and regulators are increasingly pressuring platforms to limit the spread of disinformation. In the US, the House of Representatives organized hearings and convoked representatives of the main platforms to shed light on how they are being weaponized to spread “misinformation and conspiracy theories online” (Donovan et al., 2020). In Europe, the European Commission has established a ‘Code of Practice on Disinformation’² that enjoins platforms to voluntarily comply with a set of commitments (Heldt, 2019). However, there is little data available and few established processes to monitor the implementation of these measures and quantify their actual impact. This is what we propose to tackle in this paper by offering a methodology to monitor Facebook’s implementation of one of its core policies against misinformation. We chose to focus on Facebook as it is the biggest social media platform with more than 2 billion users worldwide.

¹<https://www.parse.ly/resources/data-studies/referrer-dashboard>, accessed on 2021-07-08.

²<https://ec.europa.eu/digital-single-market/en/code-practice-disinformation>.

Facebook announced a three-part policy to fight against ‘misleading or harmful content’: they claim to *remove* harmful information, *reduce* the spread of misinformation and *inform* people with additional context³. Facebook has developed the most extensive third-party fact-checking program with dozens of partner institution to assist the company in this endeavour⁴. When a fact-checking partner flags a URL, a post or a video as misinformation, Facebook claims to display the posts marked as “False” or “Partly False” further down in users’ feed, further reducing the virality of these posts. Facebook also informs page or group owners when published posts on pages or groups that they manage are marked as misinformation, inviting them to correct the posts. Facebook’s *reduce* policy is not only applied to individual posts, but also to organizations and communities that often publish posts containing misinformation, as indicated by this statement in their publishers’ help center⁵:

Pages and websites that repeatedly share misinformation rated False or Altered will have some restrictions, including having their distribution reduced.

So far Facebook has not provided data showing how their reduce policy is implemented, which would allow researchers to quantify its impact on misinformation circulation. To the best of our knowledge, the impact of the reduce policy has not yet been audited directly. It is in this way that the present research paper distinguishes itself from the articles that measured overall levels of misinformation on the platform (Allcott et al., 2019; Kornbluh et al., 2020; Resnick et al., 2018).

We used CrowdTangle, a public insights tool owned and operated by Facebook, to access Facebook data (Team, 2021). CrowdTangle exclusively tracks public content, and provides access to engagement metrics (such as number of likes, shares and comments), but not to the reach (number of views) of content⁶. We first investigated how

Facebook enforces its ‘reduce’ policy by combining data from a Facebook fact-checking partner identifying URLs sharing misinformation and tracking engagement metrics of the Facebook accounts that repeatedly share such misinformation. We then further investigated the effects of Facebook’s policy on engagement metrics of a set of Facebook pages claiming to be under reduced distribution.

2 Investigating the ‘reduce’ policy on Facebook groups repeatedly sharing misinformation

To investigate the effect of fact-checking on Facebook accounts that repeatedly share misinformation, we first used data from Science Feedback, which is part of Facebook’s third-party fact-checking program⁷.

2.1 Methods

Science Feedback is a fact-checking organization, in which academics are verifying the credibility of science-related claims and articles. Out of the 4,000+ URLs labeled by Science Feedback, we relied on the 2,452 URLs labeled as ‘False’, which we call “false news links”. Were excluded the URLs labeled as ‘Partly False’, ‘Missing Context’, ‘False headlines’ or ‘True’, as well as the URLs marked as ‘False’ but ‘corrected to True’ by the publisher, since these labels do not contribute to the repeat offender status according to Facebook’s guidelines. The list of false news links was obtained on January 4, 2021 and cover links flagged in 2019 and 2020.

Using the ‘/links’ endpoint from the CrowdTangle API, we gathered the public Facebook groups and pages that shared at least one false news link between January 1, 2019 and December 31, 2020. Due to the API limitations, if a URL was shared in more than 1000 posts, we collected only the 1000 posts that received the highest number of interactions⁸. We focused on the accounts that spread the most misinformation, and chose a threshold of 24 different false news links shared over the past two years.

The corresponding 307 Facebook accounts (289 Facebook groups and 18 Facebook pages) are named ‘repeat offenders accounts’. All the posts

³<https://about.fb.com/news/2018/05/inside-feed-reduce-remove-inform/>

⁴<https://www.facebook.com/business/help/341102040382165>

⁵<https://www.facebook.com/business/help/2593586717571940>
<https://www.facebook.com/business/help/297022994952764>

⁶<https://help.crowdtangle.com/en/articles/3192685-citing-crowdtangle-data>, <https://help.crowdtangle.com/en/articles/4558716-understanding-and-citing-crowdtangle-data>

⁷<https://sciencefeedback.co/science-feedback-partnering-with-facebook-in-fight-against-misinformation/>

⁸<https://github.com/CrowdTangle/API/wiki/Links>

they published between January 1, 2019 and December 31, 2020 were collected using the '/posts' endpoint. We calculated the engagement per post by summing the number of comments, shares and reactions (such as 'like', 'love', 'favorite', 'haha', 'wow', 'sad' and 'angry' reactions) that each post received.

'Repeat offenders' accounts are supposed to have their distribution reduced, according to Facebook's official communication, but the precise rule Facebook uses to classify an account as 'repeat offenders' is not specified. An undisclosed source obtained by a journalist indicated 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"⁹.

Based on this 'two strikes in 90 days' rule and the list of strike dates known by Science Feedback, we inferred periods during which each account must have been under repeat offender status. If a post sharing a misinformation link was published after the corresponding fact-check, we used the date of the post as the strike date. If the account first shared a link, which was then fact-checked as 'False', the fact-check publication date was used as the strike date. Any given time in which an account shared two or more false news links over the past 90 days is defined as a repeat offender period.

2.2 Results

Figure 1's displays the engagement metrics for one 'repeat offender' example group named 'Australian Climate Sceptics Group'. The known strike dates are shown as red lines at the bottom, and the inferred 'repeat offender' periods are shaded in red. The average engagement per post vary throughout the measuring period, but they do not appear to be related with the alternance of 'repeat offender' and 'no strike' periods (see Figure 1). When comparing the average engagement metrics between the 'repeat offender' and the 'no strike' periods, we actually observe a 61 % increase in engagement during the 'repeat offender' periods.

To generalize, we calculate the percentage change between the 'repeat offender' and the 'no strike' periods for each of the 256 Facebook accounts that published at least one post during each

⁹<https://www.nbcnews.com/tech/tech-news/sensitive-claims-bias-facebook-relaxed-misinformation-rules-conservative-pages-n1236182>

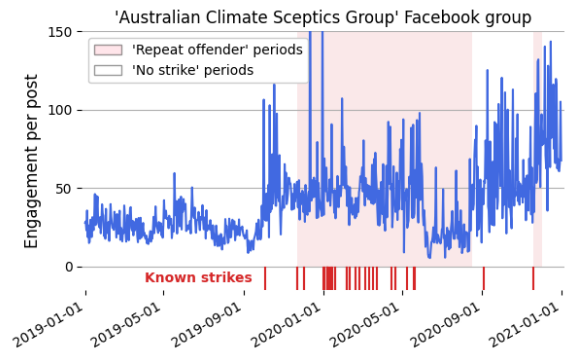


Figure 1: Average engagement (the sum of comments, shares, likes, ...) per post for the 'Australian Climate Sceptics Group' Facebook group for each day in 2019 and 2020. Each red line at the bottom represents the date of a known strike for this group, and the areas shaded in red represent the 'repeat offender' periods as defined by the 'two strikes in 90 days' rule.

period (see Figure 2)¹⁰. The average percentage change is 7%, and the median -6%, both values being very close to zero. A Wilcoxon test confirmed that these percentage changes are not significantly different from zero ($W = 16051$, $p\text{-value} = 0.74$).

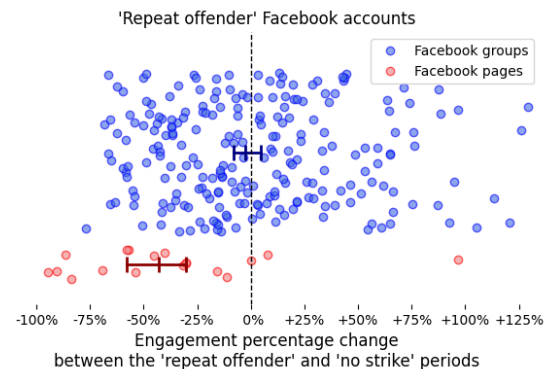


Figure 2: Percentage changes between the average engagement per post during the 'repeat offender' periods and the 'no strike' periods. Each blue dot represents a Facebook group and each red dot a Facebook page. The bars shows the median and its 90% confidence interval.

There appeared to be a difference between Facebook groups and pages, as the median percentage change for Facebook groups was -3 %, while the median for Facebook pages was -43 %. When applied only on the Facebook pages' percentage changes, the Wilcoxon test was significant ($W =$

¹⁰The percentage changes were calculated on the periods between January 1, 2019 and June 8, 2020. Because of the drop in engagement described further down, the second semester of 2020 was excluded for its vastly diminished and not representative engagement level (see Figure 3).

21, p -value = 0.0034). It should be noted that our sample contained only 18 Facebook pages, which is not enough to draw firm conclusions.

To see whether the strikes would otherwise influence an account's distribution over time, we also plotted the average engagement per post for each day of the 2019-2020 period aggregated over the 307 'repeat offender' accounts (Figure 3 Top panel). The engagement per post remained rather constant until June 9, 2020, where it suddenly experienced a massive drop that brings their levels back to a half of the early 2019 values.

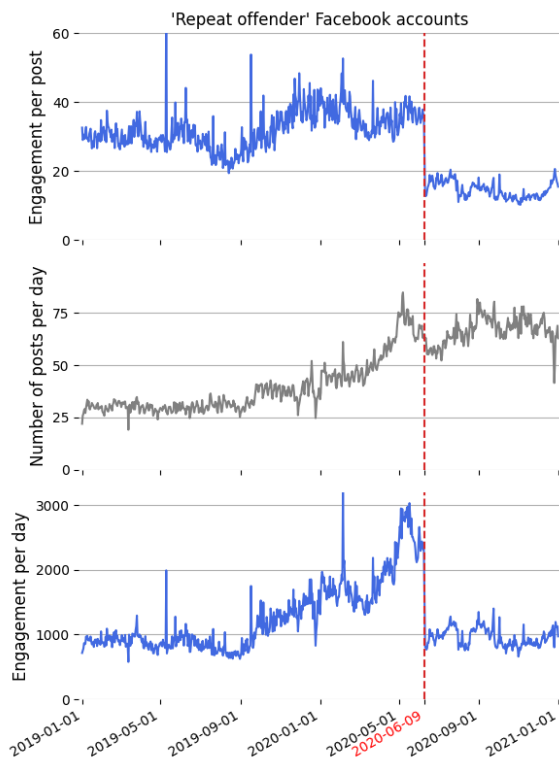


Figure 3: Metrics aggregated over the 307 Facebook accounts that repeatedly shared false news links. (**Top panel**) Average engagement per post. (**Middel panel**) Number of posts per day. (**Bottom panel**) Engagement per day. The dotted line marks the date of June 9, 2020, when a sudden drop in engagement is observed.

To contextualize this decrease in engagement, we also plotted the average number of posts published each day, and the total engagement per day generated by these accounts. The engagement per day (Figure 3 Bottom panel) can be decomposed as the number of posts (Middle panel) times the engagement per post (Top panel). The number of posts per day has more than doubled between September 2019 and June 2020, showing an increase in activity for the observed misinformation

accounts. As the engagement per post stayed stable, the total engagement they generated per day has also doubled during the same period. In a context where the overall engagement for these misinformation accounts was rising, the massive decrease observed in the engagement per post on June 9, 2020 actually brought the engagement level back to its early 2019 values, instead of decreasing it.

We looked at the accounts individually, and calculated the percentage change in engagement per post for each account between 30 days before June 9, 2020 and 30 days after (Figure 4). The average percentage change was -21%, and the median -43%. Most of the accounts (219 out of 289) experienced a decrease in engagement¹¹. A Wilcoxon test revealed these percentage changes to be very significantly different than zero ($W = 9012$, p -value = 4.6×10^{-17}).

It appears that the Facebook pages were less affected by this decrease than Facebook groups, as their median percentage change was -5% ($n = 18$), while the median for the Facebook groups was -45% ($n = 271$). When tested separately, the Facebook pages' percentage changes were not different than zero ($W = 73$, p -value = 0.61). The Facebook pages may have not been concerned by the measure taken on June 9, 2020.

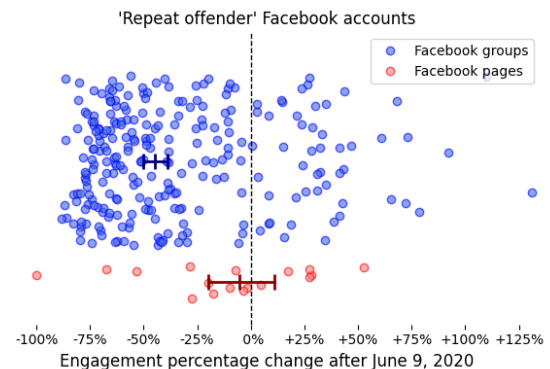


Figure 4: Percentage changes between the average engagement per post 30 days before June 9, 2020 and 30 days after. Each blue dot represents a Facebook group and each red dot a Facebook page. The bars show the median and its 90% confidence interval.

We can only explain such a massive change by a modification in how Facebook's algorithm promoted the content from these groups in June 9,

¹¹ A decrease in engagement on June 9, 2020 can be seen for the 'Australian Climate Sceptics Group' in Figure 1 (the percentage change was -60% for this example).

2020. Despite this one-shot measure, the results showed a global lack of correlation between the fact-checks' dates and a decrease in engagement for 'repeat offenders' Facebook groups. However we only took into account the links labelled as 'False' by one fact-checking organization (Science Feedback), while Facebook partners with over 60 fact-checking organizations¹². The true repeat offender periods could thus be longer than the ones inferred here. Moreover, although we observed no effect on Facebook groups, Facebook pages did appear to have a reduced engagement per post during the 'repeat offender' periods. But further data is needed to confirm this hypothesis. Thus, in the following section we rather focused on Facebook pages, and used another methodology to infer their 'repeat offender' status.

3 Investigating the 'reduce' policy on self-declared 'repeat offenders' Facebook pages

3.1 Methods

We analyzed a set of pages that have publicly shared a message complaining about being placed under 'repeat offender' status and including a screenshot as evidence. We had observed this behaviour in two popular pages ("Mark Levin" and "100 Percent FED Up") and sought more of such examples. To assemble a list of self-declared repeat offenders, we searched Facebook via the CrowdTangle API, using the '/posts/search' endpoint of the API on November 25, 2020, for posts published since January 1, 2020 with the following keywords:

- 'reduced distribution' AND ('restricted' OR 'censored' OR 'silenced')
- 'Your page has reduced distribution'

We manually opened the hundreds of resulting posts, and kept the posts we found to meet the following criteria (see Figure 5 Top panel for an example):

- they shared a screenshot of the Facebook message received
- the Facebook message indicated 'Your page has reduced distribution and other restrictions because of repeatedly sharing of false news.'

¹²<https://about.fb.com/news/2020/04/covid-19-misinfo-update/>

- the page name was visible on the screenshot message

We obtained a list of 94 pages associated with one of these posts. We found only Facebook pages in this case, and no groups. A search using the terms 'Your group has reduced distribution' did not yield any result.

To verify whether Facebook applied any restriction to these pages, we collected all the posts that these 94 pages published between January 1, 2019 and December 31, 2020 from the CrowdTangle API using the '/posts' endpoint. The collection was run on January 11, 2021. We were only able to collect data from 83 of these pages, as 11 were deleted from the CrowdTangle database since our search in November. This highlights an important issue when studying misinformation trends on Facebook: some data disappears as accounts are deleted or changed to 'private'.

The date of the last violation notification was used as the inferred start date of reduced distribution when the date was visible in the screenshot. When the screenshot did not include the date of the last violation notification, we used the date of the post as the inferred start date of reduced distribution.

3.2 Results

Figure 5 shows the reduced distribution Facebook notification shared by the 'I Love Carbon Dioxide' page on April 28, 2020, and the average engagement per post of that page over the past two years. The engagement do not appear to be reduced after April 28, 2020. If we compare the engagement between 30 days before and 30 days after April 28, 2020, the percentage change is only 2%, indicating almost no change between these two periods.

We then calculated the percentage change 30 days before and after the reduced distribution start date for each of the 82 Facebook pages that published at least one post during these periods (see Figure 6). The average percentage change was -16%, the median -24%, and a Wilcoxon test revealed the percentage changes to be significantly different than zero ($W = 911$, $p\text{-value} = 0.00026$). We thus found a small decrease in engagement following the 'reduced distribution' notification, that varied greatly across the different Facebook pages.

We further investigated whether an important drop in engagement also occurred on this set of

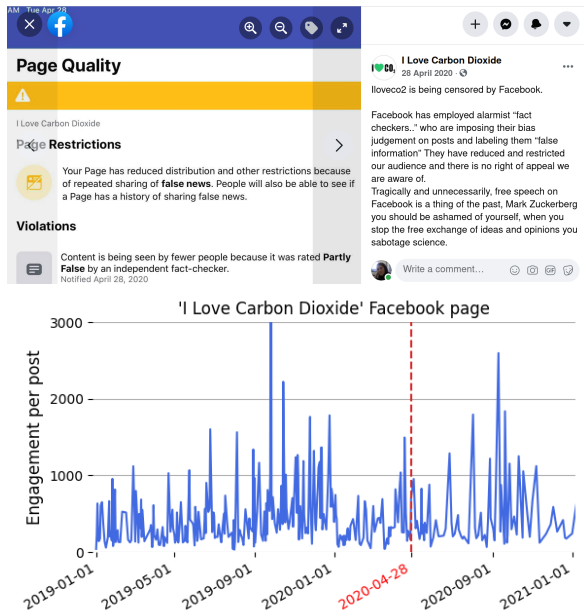


Figure 5: (Top panel) Screenshot of a post from the 'I Love Carbon Dioxide' Facebook page sharing a reduced distribution notification from Facebook. (Bottom panel) Average engagement per post for the "I Love Carbon Dioxide" page for each day in 2019 and 2020, with the reduced distribution start date shown in red.

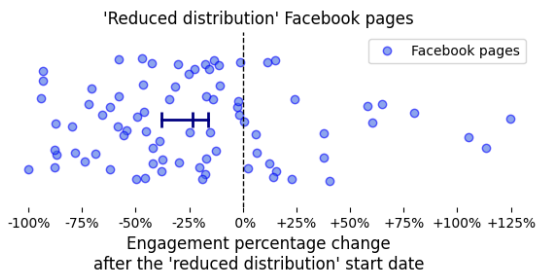


Figure 6: Percentage changes between the average engagement per post 30 days before the 'reduced distribution' start date and 30 days after. Each dot represents a Facebook page. The bars show the median and its 90% confidence interval.

Facebook pages. When we compared 30 days before and after June 9, 2020, the median percentage change was 3%, and no significantly different than zero ($W = 1093$, $p\text{-value} = 0.055$). It confirms that Facebook pages appeared to be not affected by the reduce measure implemented on the Facebook groups investigated in the previous section.

4 Discussion

Facebook, the most widely used social media platform in the world, has announced a series of measures to curb the spread of misinformation, no-

tably by reducing the visibility of 'repeat offenders': accounts that repeatedly share false information. However, the effects of platforms' diverse policies to tackle misinformation remains understudied (Pasquetto et al., 2020). Our study aims to address this knowledge gap by verifying the application and measuring the consequences of Facebook's 'reduce' policy on the targeted accounts' engagement metrics.

As a first step, we investigated the reach of 307 Facebook accounts (mainly groups) having repeatedly shared false information using a fact-checker's dataset. The sharing of two false news links over a three-month period is supposed to be penalized by reduced distribution of their content. However, we find no compelling evidence that this policy of Facebook is actually leading to a significant reduction in engagement for Facebook groups, which would be an expected outcome of reduced distribution. However, a decrease in engagement was visible for the 18 Facebook pages in the sample. As a second step, we gathered 83 Facebook pages that shared a notification they received from Facebook announcing their account was under reduced distribution. The pages' engagement metrics were significantly lower after the notification, suggesting that the 'reduced distribution' measure was indeed implemented on the pages. It should be noted that no group was found when searching for accounts sharing a reduced distribution notification, which reinforces the hypothesis that the 'repeat offender' policy is applied only on Facebook pages, and not on groups.

By analyzing the evolution of the repeat offenders' engagement metrics, we also discovered a sudden drop affecting Facebook groups around June 9, 2020. For many Facebook groups, the decrease was drastic: around 70% - 80%. Interestingly, the 18 Facebook pages collected in the first sample were not concerned by this decrease, as the 83 pages from the second sample. This 'June drop' does not correspond to any official communication by Facebook on the matter. It indicates that Facebook has very likely taken internal decisions that heavily impact the organic reach of repeat offenders' groups, in ways that differ from its stated policy. More transparency from Facebook would be needed to understand the nature and origin of this change. It would also bring clarity on how rules aimed at limiting the spread of misinformation are being enforced.

On the difference between pages and groups. Why implement different reduce measures on groups and pages? Why a sudden reduce measure, and has there been other similar measures implemented on other times?

Online misinformation introduces various threats to societies around the world, and the role of platforms in its distribution and regulation has been the subject of intense debate over the past few years (Rogers, 2020; De Gregorio and Stremlau, 2020). Consistent with other studies finding that misinformation persists at high levels on Facebook and other platforms (Kornbluh et al., 2020; Resnick et al., 2018), we find no evidence that Facebook is reducing the distribution of content from known spreaders of misinformation. One step Facebook could take to reduce misinformation on its platform would be to more rigorously enforce their repeat offenders policies. In the meantime, we advocate for scientists to pay greater attention to monitoring the actions, or lack thereof, by platforms against misinformation.

References

- Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36.
- Hunt Allcott, Matthew Gentzkow, and Chuan Yu. 2019. Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2):2053168019848554.
- Yochai Benkler, Casey Tilton, Bruce Etling, Hal Roberts, Justin Clark, Robert Faris, Jonas Kaiser, and Carolyn Schmitt. 2020. Mail-in voter fraud: Anatomy of a disinformation campaign. *Available at SSRN*.
- R Brulle. 2018. 30 years ago global warming became front-page news—and both republicans and democrats took it seriously. *The Conversation*.
- Giovanni De Gregorio and Nicole Stremlau. 2020. Internet shutdowns in africa—internet shutdowns and the limits of law. *International Journal of Communication*, 14:20.
- J. Donovan, N. Jankowicz, C. Otis, and M. Smith. 2020. House intelligence committee open virtual hearing: “misinformation, conspiracy theories, and ‘infodemics’: Stopping the spread online”. <https://intelligence.house.gov/news/documentsingle.aspx?DocumentID=1092>.
- Jieyu Ding Featherstone and Jingwen Zhang. 2020. Feeling angry: the effects of vaccine misinformation and refutational messages on negative emotions and vaccination attitude. *Journal of Health Communication*, 25(9):692–702.
- Richard Fletcher, Antonis Kalogeropoulos, Felix M Simon, and Rasmus Kleis Nielsen. 2020. Information inequality in the uk coronavirus communications crisis. *Reuters Institute for the Study of Journalism*.
- Amélie Heldt. 2019. Let’s meet halfway: Sharing new responsibilities in a digital age. *Journal of Information Policy*, 9:336–369.
- K Kornbluh, A Goldstein, and E Weiner. 2020. New study by digital new deal finds engagement with deceptive outlets higher on facebook today than run-up to 2016 election. gmf the german marshall fund of the united states. viitattu 16.12. 2020.
- Marin Lahouati, Antoine De Coucy, Jean Sarlangue, and Charles Cazanave. 2020. Spread of vaccine hesitancy in france: What about youtube™? *Vaccine*, 38(36):5779–5782.
- Paul Mena. 2020. Cleaning up social media: The effect of warning labels on likelihood of sharing false news on facebook. *Policy & internet*, 12(2):165–183.
- Amy Mitchell, Jeffrey Gottfried, Michael Barthel, and Elisa Shearer. 2016. The modern news consumer: News attitudes and practices in the digital era. *Pew Research Center*.
- Irene V Pasquetto, Briony Swire-Thompson, Michelle A Amazeen, Fabrício Benevenuto, Nadia M Brashier, Robert M Bond, Lia C Bozarth, Ceren Budak, Ullrich KH Ecker, Lisa K Fazio, et al. 2020. Tackling misinformation: What researchers could do with social media data. *The Harvard Kennedy School Misinformation Review*.
- Ethan Porter, Thomas J Wood, and Babak Bahador. 2019. Can presidential misinformation on climate change be corrected? evidence from internet and phone experiments. *Research & Politics*, 6(3):2053168019864784.
- Adrian Rauchfleisch and Jonas Kaiser. 2021. Deplatforming the far-right: An analysis of youtube and bitchute. *Available at SSRN*.
- Paul Resnick, Aviv Ovadya, and Garlin Gilchrist. 2018. Iffy quotient: A platform health metric for misinformation. *Center for Social Media Responsibility*, 17.
- Richard Rogers. 2020. Deplatforming: Following extreme internet celebrities to telegram and alternative social media. *European Journal of Communication*, 35(3):213–229.
- CrowdTangle Team. 2021. Crowdtangle. *Facebook, Menlo Park, California, United States*.