Digital Platforms' Governance: missing data & information to monitor, audit & investigate platforms' misinformation interventions

Shaden Shabayek, Emmanuel Vincent, Héloïse Théro

DE FACTO
Observatory
of Information

DIGITAL PLATFORMS' GOVERNANCE

Missing data & information to monitor,
audit & investigate platforms' misinformation interventions

Shaden Shabayek, Emmanuel Vincent and Héloïse Théro

Sciences Po médialab - June 2022 Draft version

Digital Platforms' Governance

Missing data & information to monitor, audit & investigate platforms' misinformation interventions

1. What's the issue?

There is a growing concern in society about the spread of misinformation on online platforms and its potential impact on democratic debates and public health. To address this concern, online platforms have been expanding their rules in order to tackle the spread of misleading information. During the COVID-19 global health pandemic, platforms have shown a willingness to ensure the access to reliable health information by implementing further new policies. Moreover, regulators on a national and European level are making progress on the design of a legal framework specifically tailored to tackle disinformation[1]. Namely large platforms in operation have signed the "EU Code of Practice on Disinformation" (2018). This code lists a number of actions that large platforms have agreed to implement, such as to "reduce revenues of the purveyors of disinformation", "prioritize relevant, authentic, and accurate and authoritative information" or "dilute the visibility of disinformation by improving the findability of trustworthy content". Since then, several new regulatory guidelines have been adopted at the level of the European Union or are awaiting entry into force; such as the Strengthened Code of Practice on Disinformation (June 2022) and the Digital Services Act[2] (hereafter the DSA) which includes an obligation for very large platforms to give "access to data that are necessary to monitor and assess compliance with this Regulation" (see Article 31 of the DSA proposal) to vetted researchers[3] according to specified requirements by the act. Along similar lines, Trans-atlantic initiatives have emerged, such as the Platform Accountability and Transparency Act (PATA), a bipartisan bill introduced by US senators Coons, Portman and Klobuchar (December 2021). This bill requires that platforms make certain key information available to independent researchers[4]. At the European level, the DSA adds up to the GDPR applied since May 2018, which offers further guarantees for the respect of privacy and the ethical use of data for research purposes. In that context, a variety of actors are reflecting and organizing the practicalities of such legal frameworks to meet up with ethical concerns related to data circulation between platforms and many members of the society. In particular, the provisions of Article 40 of the GDPR encourage the drawing up of codes of conduct. The working group on Platform-to-Researcher Data Access of the European Digital Media Observatory (EDMO) has recently (May 2022) drafted such a code of conduct so that data circulation between platforms and researchers can be organized in practice. At a national level, regulators such as the ARCOM in France are gathering[5] information about the type of data that would be needed so that researchers can effectively investigate the impact of digital platforms on our informational ecosystem.

Now from the perspective of researchers, assessing regularly the impact and pertinence of misinformation related interventions by online platforms and monitoring their implementation, as well as a careful investigation of the phenomenon of misinformation itself, are necessary safe-guards for democratic societies with growing digital spheres. Since their early

[1] See the Chronology of EU's action against disinformation from 2015 to 2021.

[2] To date (01/06/2022), the text still needs to be finalized at technical level, it is awaiting formal approval from Parliament and council. See the press release: https://www.europarl.europa.eu/news/en/press-room/20220412IPR27111/digital-services-act-agreement-for-a-transparent-and-safe-online-environment

[3] For example, the requirements for vetted researchers are given in paragraph 4 of Article 31 of the Proposal for a Regulation on a Single Market for Digital Services (Digital Services Act), 15 December 2020.

[4] For an update on this discussion, watch the May 4, 2022 Subcommittee Hearing, presiding Chair Coons on "Platform Transparency: Understanding the Impact of Social Media" judiciary.senate.gov/meetings/platform-transparency-understanding-the-impact-of-social-media

[5] See the call https://www.arcom.fr/consultations-publiques/consultation-publique-sur-lacces-aux-donnees-des-plateformes-en-ligne-pour-la-recherche

days, online platforms have emerged as digital spaces where information and opinions can circulate freely. The task of ensuring a balance between freedom of expression and access to reliable information regarding political life or public health, is tremendously intricate. In spite of transparency efforts by digital platforms, a number of issues still remain. There is limited access to specific data and information which would enable the academic community, NGOs, the civil society and data journalists to successfully study online misinformation along with the related interventions. In what follows, we provide illustrations of ways to monitor most common misinformation related interventions with currently available data, which precisely demonstrate the scarcity of pertinent data.

We further lay out a list of missing data and information that would enable more effective monitoring, auditing and investigation of misinformation-related interventions by platforms, along with the breadth of misinformation. Clearly, this list intersects items which are present in the above mentioned legal Acts (e.g. Article 30 of the DSA on Additional online advertising transparency). However, our list is meant to not only enumerate missing data with precision but also propose a differential level of access to this data, ranging from exclusive access to vetted researchers within a legal framework when it comes to sensitive data, to a broader access by enriching the fields in currently available APIs. Designing different levels of access to different types of data is meant to attain two goals: (1) preserve privacy of users and address potential concerns of digital platforms regarding legal immunity when giving extended access to their data or information which might put at stake the functioning of their business model (2) provide access to richer data to a wider set of actors when ethical concerns are not at stake, because the task at hand is considerable and combining the results of a variety of actors using different tools of analysis and perspectives, ranging from vetted researchers to journalists and NGOs, can yield a richer understanding of the functioning of platforms and their impact on our informational ecosystem.

2. How do platforms action content related to misinformation?

Over the past years, growing concerns about the spread of misinformation have encouraged platforms to action content deemed as misleading. Qualifying a piece of content as misleading or false can be a challenging algorithmic task, and a human intervention is generally necessary, via fact-checking organizations, moderators, or users' reporting. For example, Facebook (Meta) is working with over 80 fact-checking partner organizations and, according to the International Fact-Checking Network, the company represents the main source of revenue for many of these organizations. Twitter has a different approach where they focus on providing context rather than fact-checking[6] and the platform is testing a new crowd-based moderation system called "Birdwatch". As for YouTube (Google), this platform utilizes the schema.org ClaimReview markup, where fact-checking articles created by eligible publishers can appear on information panels next to the related content.

During the COVID-19 global health pandemic, platforms have upgraded their guidelines to include a set of rules to tackle the propagation of potentially harmful content. Those policies are enforced via existing actions such as: labeling content to

[6] To the best of our knowledge, Twitter does not have a page which summarizes its fact-checking strategy, if any. 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.

provide more context or indicate falsehood, publishing a list of claims that will be flagged or removed, suspending accounts, implementing strike systems, reducing the visibility of content, etc. Facebook's strategy to tackle misinformation is three fold : Remove, Reduce, Inform[7]. Twitter communicates about actions related to misinformation via their Twitter Safety account, such as testing a new feature to report Tweets that seem misleading to users or starting the automated labeling of Tweets that may contain misleading information about COVID19 vaccination[8]. On the YouTube Official Blog, the platform explains its "Four Rs of Responsibility" and how it raises authoritative content, reduces borderline content and harmful misinformation[9]. As each platform is a private company, those policies are not coordinated and are implemented in different ways across platforms. Nevertheless, there are common interventions against misinformation used by large digital platforms, which could be classified into three broad categories[10]: (*i*) informing users with flags, notices and information panels; (*ii*) reducing the visibility of some content, either at the post level or at the account level: and (*iii*) deleting content and suspending users temporarily or permanently in case of multiple rule violations.

Platforms make data available via official APIs (e.g. CrowdTangle, Twitter API V2, YouTube API V3), recent academic partnership programs (e.g. Social Science One[11]) and transparency centers. But specific data and information to investigate misinformation interventions and their impact are scarce - such as data about whether a piece of content has a notice or an information panel, or fields indicating the specific policy violation when an account, page or group is suspended. Furthermore, when navigating through the categories of policy areas, to the best of our knowledge, misinformation is absent from the data available on the transparency centers of key platforms, such as Facebook, Twitter and YouTube. Finally, platforms' official communication about misinformation interventions is rare and the academic community, NGOs and data journalists, usually discover interventions related to misinformation via monitoring social media accounts related to domain names with several failed fact-checks or via articles in news outlets.

3. Can we investigate misinformation related interventions with currently available data?

a. Account suspensions and content deletion

To date, investigating removed content and account suspensions, due to policy violations, is a burdensome task. This is because when a platform deletes a piece of content violating its rules or invites a user to delete a piece of content to regain access to the functionalities of the platform, the data disappears from official APIs[12] and naturally is no longer visible on the platform. Hence the deleted content can no longer be investigated. Indirect methods could be designed to study suspensions and content deletion linked to misinformation[13]. For

[7] See about.fb.com/news/2019/04/remove-reduce-inform-new-steps/ (last accessed May 16, 2022).

[8] See twitter.com/TwitterSafety/status/1379515615954620418 and twitter.com/TwitterSafety/status/1483076718730649607 (last accessed May 16, 2022).

[9] See blog.youtube/inside-youtube/the-four-rs-of-responsibility-remove/ (last accessed May 16, 2022).

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

[11] Social Science One is an academic organization based at Harvard University and founded by Facebook in 2018. It aims at facilitating partnerships between the academic sphere and private companies who own informative data about people and society. Namely, this partnership allowed the release of a very large Facebook URLs Dataset and facilitated access to this dataset for the purposes of academic research.

[12] Application Programming Interface.

[13] To the best of our knowledge, aggregate figures of account suspensions and content deletion directly related to misinformation rarely exist, with the exception of COVID-19 misleading information on the Twitter transparency center. Transparency reports of other

instance, in order to study the volume of suspensions or deleted content, one can re-collect data of pre-existing lists or collections from previous research of Facebook pages, groups or posts, YouTube channels or videos, or Twitter accounts or Tweets, and then investigate which accounts/content remain available and which have disappeared (if any). Similarly, using a cross-platform approach, researchers can gather some information about deleted content. Precisely, some users can deplore being targeted by a given platform's intervention (e.g. YouTube) via an account on a different platform (e.g. Twitter), or when a piece of content is simply removed from one platform (e.g. YouTube video) but information about that specific content remains on other platforms (e.g. title of a YouTube video along with the (inactive) link redirecting to it on YouTube). However, it should be noted that even when a platform displays a message indicating that a piece of content was removed for policy violation or that an account was suspended, to the best of our knowledge the specific policy violation is rarely indicated. Another indirect method consists in regularly monitoring and collecting data about accounts, pages, groups and channels who have previously shared content identified by fact-checkers as misleading, to be able to go back in time and analyze the content in case of an account suspension and investigate the pertinence and effectiveness of such intervention. Hence, access to this data needs to be provided in order to get a better understanding of the impact of content deletion directly linked to misinformation interventions and study potential indirect effects (e.g. users' migration to new platforms). Namely to address ethical and privacy concerns, access can be provided to vetted researchers for a fixed period of time, before the permanent deletion of the related metadata by platforms.

b. Reducing content visibility

Similarly, investigating how platforms reduce the visibility of problematic content via algorithmic recommendation systems can only be achieved with the design of models and experiments to simulate our understanding of how algorithms work. This method can yield incomplete and biased results, since there is no clear overview of how platforms integrate in their algorithms signals or variables to downrank or make problematic content less visible. Furthermore, very basic metrics such as the "reach" of posts on Facebook or Twitter (the number of actual viewers of a piece of content) are absent from official APIs and researchers have to resort to proxy measures, such as looking at the engagement received by pages or groups or accounts having shared multiple times a piece of content fact-checked as False, to then try to infer whether their visibility has been reduced. However, this approximation lacks precision. This is because the engagement rate (e.g. likes or comments numbers) reflects the intensity of the discussions on a given page or groups, but not the actual audience (reach) of a given piece of content. For example, pages linked to mainstream media outlets often have a very high audience and a low engagement rate, while highly politicized pages can have very high engagement rates without having a large audience. Finally, note that part of this information can be recovered from the Privacy-Protected Full URLs Data Set (see description in the next section), namely users' interactions on Facebook (e.g. views or clicks) with URLs that were "publicly" shared over 100 times. However, with this data, one cannot study the evolution over time of the reach (views or clicks) of a given Facebook public page or group that have repeatedly shared misleading content, especially those with a limited audience, because not all posts contain URLs, not all posts have been shared "publicly" over 100 times and noise added to the engagement count can heavily bias the analysis especially for low values.

platforms can be accessed via the Twitter Transparency center, see the section "industry transparency reports" (last accessed May 10, 2022).

Box 1: Systematic review: YouTube recommendations and problematic content.

This paper provides a systematic review of studies that have investigated the YouTube recommender system in relation with problematic content, such as radicalisation, conspiracy theories, misinformation and disinformation. Exactly 23 papers, which meet a set of eligibility criteria of inclusion, are analyzed. The authors find that 14 studies out of the 23 suggest that the YouTube recommender system is facilitating problematic content pathways, that is individuals are exposed to content via the recommender system that they might have not otherwise encountered. The remaining studies either produce mixed results (7 studies) or do not suggest that the YouTube recommender system is facilitating exposure to problematic content (2 studies). Finally, the authors discuss the limitations of these studies, in terms of methods and modeling choices. These limitations stem from the lack of a clear understanding of how YouTube recommender systems actually work, since such information is not fully disclosed, but also from limited access to relevant data.

Yesilada, M. & Lewandowsky, S. (2022). Systematic review: YouTube recommendations and problematic content. *Internet Policy Review*, *11*(1).

C. Informing with banners, flags and notices

By contrast, information banners, flags and notices, can be studied since they are visible when searching posts or videos. This intervention was particularly adopted by many large platforms in direct relation with the diffusion of misleading information and is usually based on fact-checked content, moderators' intervention or users' reporting. In what follows, we design several exercises to show how this intervention consisting in informing users could be studied or monitored. We take a set of fact-checked links as a starting point and study the kind of interventions applied to them.

Privacy-Protected Full URLs Data Set

We use a dataset extracted from the June 2021 version of the Facebook Privacy-Protected Full URLs Data Set, (Messing et al., 2021), which contains a column that indicates whether a rating was attributed to a given URL by a third-party fact-checking partner (see the field "tpfc_rating"). This very large data set constitutes a great resource for researchers because Facebook has the most substantial fact-checking program relative to other platforms and hence it provides aggregated data about which links contained in posts have been subject to a fact-check. It can be used not only to study misinformation on Facebook, but also on other platforms such as YouTube and Twitter, because of cross-platform traffic (e.g. sharing a YouTube video via a Facebook post or a Tweet). However it should be noted that the dataset is updated with a delay. Hence it cannot be used in real time to investigate ongoingevents. Furthermore in order to protect users' privacy, Facebook limited access to the dataset to URLs that have been shared publicly[14] over 100 times and this might create biases when trying to investigate the volume of misinformation.

[14] This means that for a link to be included in the dataset, it must have been shared at least 100 times by users who chose the option "public" in their privacy settings (as opposed to "share to friends" for example).

Figure 1 displays the breakdown of links that were reviewed by third-party fact-checking partners, by rating categories: fact checked as false, not eligible, fact checked as true, etc. Figure 2 shows the number of links fact checked as false grouped by the country where the link was most shared. We restrict the present investigation to the set of 2109 links flagged as False since January 1, 2020 and that were most shared in the ten European countries[15] for which we have over 100 links in the dataset. As of April 5, 2022 only 1436 among the 2109 links are still active (status 200).

Box 2: Examining potential bias in large-scale censored data

This paper conducts a cross-dataset analysis over the month of December 2018, to compare findings based on (i) the Facebook Privacy-Protected Full URLs Data set and (ii) data from a nationally representative desktop web panel from the company Nielsen. They found that 10% of clicks went to fake news domains in that month when using the Facebook dataset, against 2.5% of clicks when using the Nielsen dataset. By matching URLs between both datasets, along with a CrowdTangle investigation and an internal Facebook investigation, the authors show that this overestimation (4X difference) is due to the 100-public-share threshold introduced in the Facebook dataset for privacy-protective procedures. Hence, they argue that censoring part of this large Facebook dataset can dramatically affect conclusions about misinformation drawn from the data.

Allen, J., Mobius, M., Rothschild D. M., & Watts, D. J. (2021). Research note: Examining potential bias in large-scale censored data. *Harvard Kennedy School (HKS) Misinformation Review*.

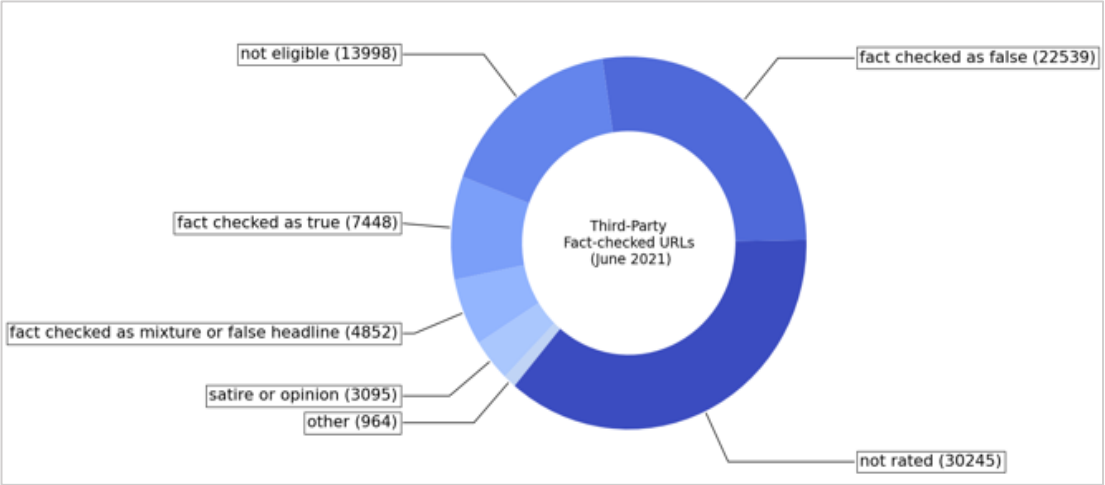


Figure 1: Third Party Fact-Checked URLs in the dataset broken down by rating

We show the volume per rating category of links for which the field “tpfc_rating” in the data set is not empty. The category “other” includes exactly three rating categories: ‘fact checked as missing context’ (821 links), ‘prank generator’ (142 links) and ‘fact checked as altered media’ (1 link).

[15] The ten countries are: France, Germany, Greece, Spain, Poland, Hungary, Great-Britain, Italy, Netherlands, Lithuania.

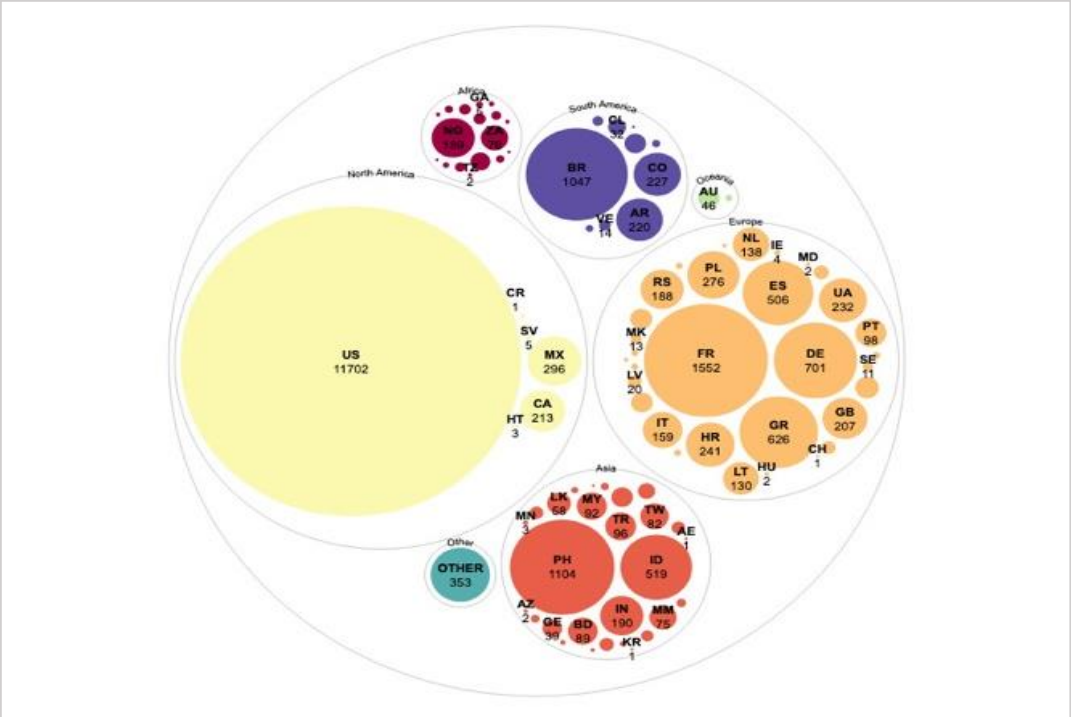


Figure 2: Links fact-checked as false, by country where it was most shared
(See the field “public_shares_top_country” in the dataset)

Case Study 1: YouTube videos fact-checked as False

We found 293 links that correspond to YouTube videos, among the 2109 links rated as False by a third party fact-checking partner, from ten EU countries where those links were most shared. We wanted to know whether these videos have been actioned by YouTube. To that end, we searched on the YouTube API v3 for these videos and found that only 76 videos (~26%) were still available. However, we were unable to find on the YouTube API v3, a field indicating whether the available video contained an information panel or any metadata related to the videos that were no longer available. Hence, we searched for this information by scraping the content of information panels when the videos were active and any message that appears when clicking on the links of unavailable videos (see Figure 3).

As summarized in Figure 4, we found that only 20 videos (out of 76 active videos) contained an information panel (about COVID19) and 12% of the remaining videos without an information panel had the comments functionality deactivated. Furthermore, among the 217 inactive videos around ~20% were removed and the most common reason (86% of those videos) was “for violating YouTube Community Guidelines” (see panel a, Figure 3), without clearly indicating which specific policy violation. The remaining 80% of inactive videos displayed a message saying that the video is unavailable and for ~53% of these videos no reason was provided (see panel b, Figure 3). Finally, for active videos, we found that the average view and like count of videos was higher for videos without an information panel, relative to videos containing an information panel (see panel a, Figure 5). However, the median view and like count of videos containing an information panel was higher than those without an information panel (see panel b, Figure 5). This indicates that for videos containing an information panel, there is a long tail of relatively “low” view and like count bringing the mean down more than the median.

Panel a ***Panel b***

Figure 3: Messages displayed for inactive videos fact-checked as False
(Screenshots taken on May 11, 2022)

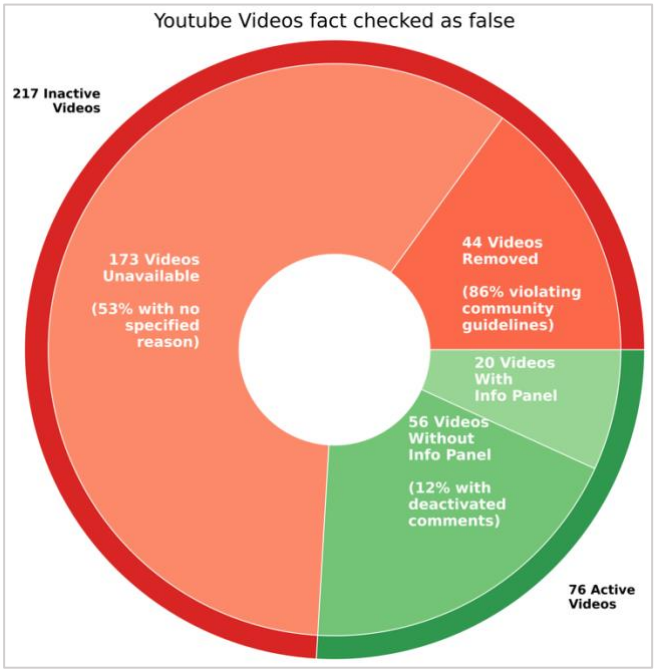
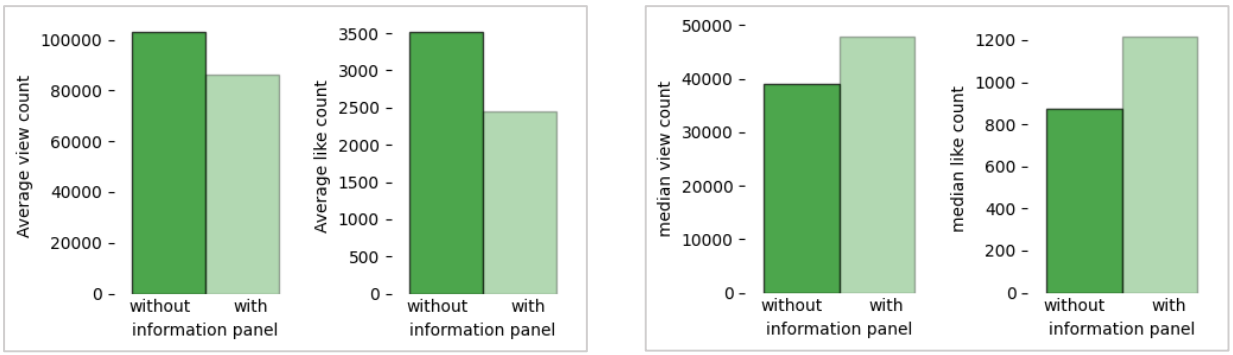


Figure 4: YouTube videos fact-checked as False



Panel a: Average view & like count **Panel b: Median view & like count**

Figure 5: View and like count of active YouTube videos "fact-checked as False"

Case Study 2: Tweets containing links fact-checked as False

For this second case study, we take the 2109 links "fact-checked as False" since January 1, 2020 and that were most shared in ten European countries (see the subsection Privacy-Protected Full URLs Data Set), and collect Tweets which have shared at least one of these links. We aimed to see whether those tweets contained a notice (see panel a, Figure 6) or an interstitial (see panel b, Figure 6). On the Twitter API v2, to the best of our knowledge, there does not exist a field which indicates whether a notice was attributed to a Tweet; only fields related to the interstitials "possibly sensitive" and "withheld" can be recovered from the Twitter API v2[16].

Hence, we collected Tweets using minet, a web mining command line tool[17], in order to be able to recover Twitter "notices" whenever they exist. We found ~72k Tweets (excluding retweets) having shared at least one of the 2109 links flagged as False (see Figure 7). Among these Tweets, only 11 contained a notice (e.g. "Stay informed") and 2910 Tweets had the interstitial "possibly sensitive". According to the Twitter rules, the interstitial "possibly sensitive" is used to advise viewers "that they will see sensitive media, like adult content or graphic violence". Manually checking Tweets marked as "possibly sensitive" we found no adult content nor graphic violence (see example in panels b and c of Figure 6) - but either links (from our initial list) or quoted tweets of other users[18]. Inspecting the ten Tweets that contained a notice (e.g., "Stay informed"), we found that: 1) there exists a big variation in terms of follower count of the eleven users linked with the eleven Tweets[19], 2) there exists other Tweets without a notice, containing the exact same link and some keywords as the Tweets containing a notice. Both remarks suggest that this process might not be automated and might be the result of users' reporting.

In addition, we compared engagement (sum of likes, retweets and replies) for Tweets containing an interstitial "possibly sensitive" and Tweets without an interstitial (see Figure 8). We found that that the average engagement for Tweets with an interstitial "possibly sensitive" were lower (around 2) when compared to tweets without an interstitial (around 6), suggesting that attributing an interstitial to a Tweet might contribute to lowering its visibility and hence engagement. However, it should be noted that in absolute values, the engagement for the collected tweets containing a link fact-checked as false is low; the median of total engagement for those Tweets is zero and the 75th percentile is equal to one.

Finally, we used a cross-platform approach and studied engagement for Tweets having shared a YouTube video marked as False by a fact-checking partner (see Case Study 1). We found that the average engagement (likes, replies, retweets) on Twitter for YouTube videos that got removed for violating YouTube community guidelines, was higher than the average engagement for YouTube videos without an information panel and that were still available.

[16] See https://help.twitter.com/en/rules-and-policies/notices-on-twitter.

[17] See https://github.com/medialab/minet.

[18] The placement of an interstitial can depend on accounts settings and we noticed that the field "possibly_sensitive" returned "True" for Tweets for which we could no longer see an interstitial when inspecting them manually directly on Twitter. Hence we suspect that the interstitial might be placed temporarily. Moreover, we suspect that these interstitials might result from users reporting, see https://help.twitter.com/en/safety-and-security/sensitive-media.

[19] Among the eleven unique Twitter users having created a Tweet to which a notice was attributed, four Twitter users had between 4 and 300 followers, two users had between 1k and 6k followers and finally four users had between 64k to 400k followers.

Panel a Panel b Panel c

Figure 6: "Stay Informed" notice and "possibly sensitive" interstitial on Twitter
(Screenshots taken on May 11, 2022)

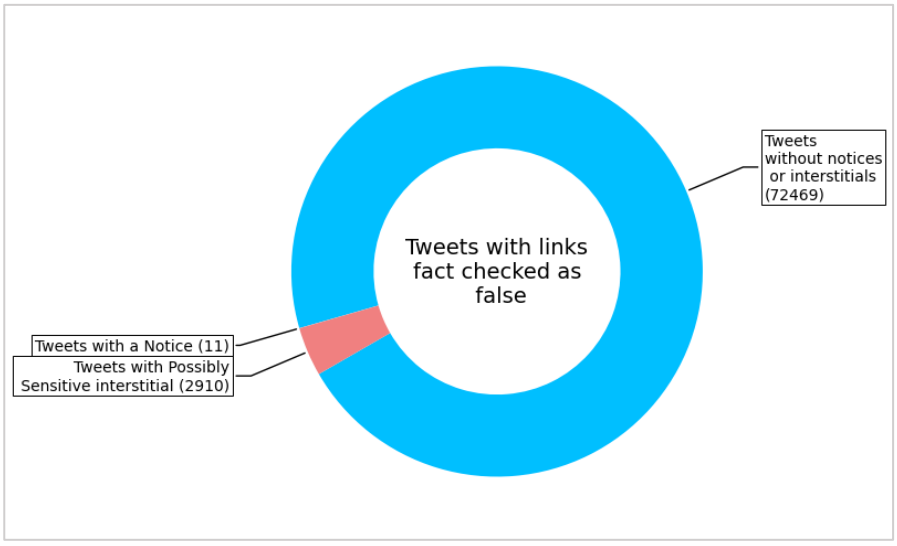


Figure 7: Tweets containing at least one link fact-checked as false

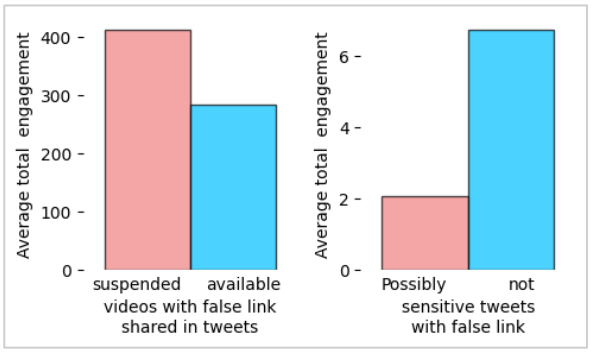


Figure 8: Total engagement for (left) suspended YouTube videos shared within Tweets and (right) Tweets marked as possibly sensitive

4. Recommendations: data & information needed to monitor and audit digital platforms' misinformation related interventions

The previous section illustrates and discusses that currently available data does not allow us to monitor most of the misinformation-related actions platforms take. In this section we list our recommendations regarding which data and information should be made available to effectively monitor, audit, and investigate misinformation related platforms' interventions.

Many of the following data and information requests can be achieved via differential access that accounts for the sensitivity of the data and information:

(1) For highly sensitive data, official APIs can be utilized by creating an extended specific data access to vetted members of the academic community or via academic partnerships similar to Social Science One. Such access would occur within a legal framework (e.g. Article 31 of the DSA) and according to a rigorous code of conduct to preserve privacy and address ethical concerns (e.g. Article 40 of the GDPR).

(2) For data that do not raise ethical concerns and with very minimal risk to privacy, current official APIs can have enriched fields, allowing a wider audience to conduct research and investigate misinformation related issues.

1. Data needed to measure the amount of misleading content in circulation and its visibility

Access to pertinent data needs to be provided in order to measure the scale of the issue and the impact of platforms' actions on the total amount of attention misinformation gets. Data needs to be available timely, without a delay of several months and it needs to be complete. For example, the "Privacy-Protected Full URLs Data Set" only includes URLs with more than 100 "public" shares, as part of their privacy-protective procedures (see box 2); so one cannot study websites with limited audiences. Moreover, there is a delay of 6 to 18 months so researchers cannot use it to study on-going events.

Data needed

- The "reach" of a piece of content, which is its actual number of views. *Collectible via a new field in existing APIs*
- list of all content (including internal posts and external fact-checked URLs) that a platform has identified as sharing misleading information, including recent content and low virality content.

2. Data needed to investigate platforms' labeling actions

Labels added by platforms to provide context on problematic content can only be obtained by scraping or visual inspection so far and a programmatic access is needed.

Data needed

- Indication of the presence of a banner, information panel or notice.
  Collectible via a new field in existing APIs. To the best of our knowledge, these are not available on the official APIs, with the exception of the 'withheld' and 'possibly sensitive' content interstitials in the Twitter API v2, which are not directly related to misinformation.
- If a banner exists, indicate whether the labeling process was algorithmic (e.g., signal based on a list of keywords) or the result of a human decision (e.g. by a moderator or users' reporting) or both. *Collectible via a new field in existing APIs.*

3. Data and information needed to investigate platforms' downranking actions

To quantify the impact of interventions consisting in reducing the visibility of a piece of content or all content produced by a given source (e.g. downranking), researchers need to be able to know when the intervention (e.g. strike) has occurred.

Data needed

- Indication, for a website or an account (Facebook group or page, Twitter accounts, Youtube channels etc.), of the number of "strikes" it received for a given misinformation-related policy violation and the specific policy violated. *Collectible via a new field in existing APIs.*
- list of all accounts/websites that have been down ranked by the platform and the periods for which the intervention was in effect.

Information needed

- the definition of the strike system of platforms when it exists and the policy violations that lead to a strike as well as the consequences of having strikes for an account.
  To the best of our knowledge, this information exists for Twitter (as part of their "Civic Integrity Policy" and "COVID 19 misleading information policy") and for Youtube for any community guidelines violation.
- details about further actions which are used to reduce the visibility of content, such as 1) excluding a piece of content from appearing via a direct search on a platform (e.g. search box of Twitter), 2) prohibiting users from sharing specific domain names in a post.

4. Data needed to monitor platforms' suspension/deletion of content/accounts and to study misinformation over time

Digital platforms can either directly remove a piece of content that violates their rules or invite users to delete it themselves to regain access to their account for example. Lost data linked to deleted content can bias the study of long-term trends of users' behavior and/or the impact of platforms' moderation policies. It can disrupt ongoing research projects, because when a page or account is suspended by a platform, researchers must adapt their protocols to deal with the missing data.

Data needed

- Indication of the specific policy violation that led to the suspension of an account and the content that violated the stated policy.
  Collectible via a new field in existing APIs.
- list of content/accounts deleted by a platform in relation to misinformation or hate speech policies.
- count of views/engagement related to removed content following a policy violation,
  - including views/engagement for all the content produced by suspended accounts, and not just the problematic content.
  - including data that was deleted by users themselves following a notification from the platform.

5. Data needed to study algorithmic recommendations of misleading content

Researchers need to be able to understand whether engagement (e.g., following a page or an account) with accounts identified as repeatedly sharing misinformation or misleading content, comes from direct search by users or from algorithmic recommendations.

Data needed

- Proportion of views of identified misleading content that results from algorithmic recommendation: *This can take the form of a field indicating the proportion of views resulting from a direct search for the content or an external link versus the proportion of views resulting from an algorithmic recommendation within the platform.*
- Proportion of the total number of views on all content shared by known misinformation sources that comes from the platform's recommendation algorithm. For example, the proportion of

views on all Youtube videos published by a channel that regularly publishes misinformation that come from the Youtube "watch next" recommendation algorithm.
- Proportion of followers gained by a given account (notably known misinformation sources) that come from the platform's recommendation algorithm: *This can take the form of a field indicating the share of followers gained via direct search and the share of followers gained via an algorithmic recommendation to follow an account or page.*

Information needed
- whether and how a platform's recommendation algorithm excludes or downranks i) content that was found to be misleading information or ii) accounts that have repeatedly shared misinformation.

6. Tracking demonetization of content/accounts spreading misinformation

As the EU Code of Practice on Disinformation asks that platforms "reduce revenues of the purveyors of disinformation", independent researchers would need access to data to verify this intervention.

Data needed
- indicating whether an account (e.g., a YouTube channel or Facebook page) was demonetized for policy violations and indication of which policies were violated. *Collectible via a new field in existing APIs.*
- list of content demonetized by a platform due to its identification as misinformation;
- list of accounts demonetized by a platform because they were identified as repeatedly spreading misleading content;
- list of content/accounts that have been considered for demonetization by the platform (e.g., following reports by the platform's users) but have not been demonetized.

Summary table

See Annex 1

5. Further readings

Misinformation: governance

- Gorwa, R. (2019). The platform governance triangle: conceptualising the informal regulation of online content. Internet Policy Review, 8(2). https://doi.org/10.14763/2019.2.1407
- Ó Fathaigh, R. & Helberger, N. & Appelman, N. (2021). The perils of legally defining disinformation. Internet Policy Review, 10(4). https://doi.org/10.14763/2021.4.1584
- Persily, N. (2021), Opening a Window into Tech: The Challenge and Opportunity for Data Transparency. In-report, Cyber Policy Recommendations for the New Administration, January 27th, 2021 Stanford Cyber Policy Center. Also see, https://law.stanford.edu/press/the-platform-transparency-and-accountability-act-new-legislation-addresses-platform-data-secrecy/
- Code of Practice on Disinformation, European Commission: https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation and the Annex Current practices from the Signatories of the Code: https://ec.europa.eu/information_society/newsroom/image/document/2018-29/annex_to_msf_cop_on_disinformation_13_07_99F63CFE-A8CE-39BF-687C68BFC0668569_53544.pdf
- Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access (2022), https://edmo.eu/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf
- Publicité Numérique Responsable, Good in Tech (2021), see discussion about BrandGuard developed by

NewsGuard: https://www.goodintech.org/uploads/files/23d72079-c50f-461b-8aa6-1bdb55b7b871.pdf

Misinformation: interventions

- Allen, J., Mobius, M., Rothschild D. M., & Watts, D. J. (2021). Research note: Examining potential bias in large-scale censored data. *Harvard Kennedy School (HKS) Misinformation Review*. https://doi.org/10.37016/mr-2020-74
- Saltz, E., Barari, S., Leibowicz, C. R., & Wardle, C. (2021). Misinformation interventions are common, divisive, and poorly understood. Harvard Kennedy School (HKS) Misinformation Review, 2(5). https://doi.org/10.37016/mr-2020-81
- Shabayek, Théro, Almanla, Vincent (2022). Monitoring misinformation related interventions by Facebook, Twitter and YouTube: methods and illustration. Available at HAL Open Archives: https://hal.archives-ouvertes.fr/hal-03662191
- Yesilada, M. & Lewandowsky, S. (2022). Systematic review: YouTube recommendations and problematic content. Internet Policy Review, 11(1). https://doi.org/10.14763/2022.1.1652
- Quick access to community guidelines & platform misinformation policies, for Facebook, Twitter and YouTube: see Annex 2.

Misinformation: fact-checking

- Annany, M. (2018). The Partnership Press: Lessons for Platform-Publisher Collaborations as Facebook and News Outlets Team to Fight Misinformation, https://doi.org/10.7916/D85B1JG9 and Checking in with the Facebook fact-checking partnership.
- Rich, T. S., Milden, I., & Wagner, M. T. (2020). Research note: Does the public support fact-checking social media? It depends on whom and how you ask. *Harvard Kennedy School (HKS) Misinformation Review*. https://doi.org/10.37016/mr-2020-46

Misinformation: general insights

- Knuutila, A., Neudert, L.-M., Howard, P. N. (2022). Who is afraid of fake news? Modeling risk perceptions of misinformation in 142 countries. Harvard Kennedy School (HKS) Misinformation Review.
- Pasquetto, I., Swire-Thompson, B., Amazeen, M. A., Benevenuto, F., Brashier, N. M., Bond, R. M., Bozarth, L. C., Budak, C., Ecker, U. K. H., Fazio, L. K., Ferrara, E., Flanagin, A. J., Flammini, A., Freelon, D., Grinberg, N., Hertwig, R., Jamieson, K. H., Joseph, K., Jones, J. J. Yang, K. C. (2020). Tackling misinformation: What researchers could do with social media data. Harvard Kennedy School (HKS) Misinformation Review.
- Rogers, R. (2020). Research note: The scale of Facebook's problem depends upon how 'fake news' is classified. Harvard Kennedy School (HKS) Misinformation Review.

Ethics: the data collection and processing complied with the EU General Data Protection Regulation (GDPR). The case studies use the Facebook Privacy-Protected Full URLS Data set and have been submitted for pre-publication review as indicated in the updated (April 2022) Research Data Agreement. The review concluded that no Personal Information or Confidential Information was found.

Co-funded by the European Commission under the Connecting Europe Facility (CEF) – Telecommunications Sector (Agreement No. INEA / CEF / ICT / A2020 / 2394372).

Annex 1: Summary table - data & information needed to monitor and audit digital platforms' misinformation related interventions

Policy area or type of content	New Field collectible via an API	Dataset to provide to vetted researchers (time period to be determined)	Information to provide	Research projects/questions
1. Content in circulation identified as misleading	**Reach metric**, i.e. actual number of views of a piece of content	List of content (internal posts & external fact-checked URLs) that a platform has identified as misleading + recent content & low virality content.		This data is needed to study the effectiveness of fact checking by investigating the circulation of links identified as misleading (within posts marked as such, e.g. "false" rating). How do users engage with labelled content (e.g. flagged as "false")? How effective are users' reports (crowd-sourcing approach) in identifying misleading or problematic content, when compared to fact-checking? More broadly, what is the reach of misinformation ? How do users interpret most common misinformation related interventions ?
2. Labeling actions	Presence of a banner or information panel If banner, indicate whether the labeling process is **algorithmic** (e.g. signal, keywords) or the result of a **human decision** (e.g. fact-checking partner, moderator or users' reports) or both			
3. Downranking actions (in relation to misinfo)	If strike system, **number of "strikes"** + specific policy violated	List of all accounts/websites that have been downranked by the platform and the periods for which the intervention was in effect	definition of the strike system of platforms when it exists + policy violations that lead to a strike + consequences of having strikes for an account.	The data is needed to study the effectiveness of platforms' interventions to try and limit the spread of misinformation.
4. Suspension of accounts & deletion of content (in relation to misinfo)	**Specific policy violated**, when a message appears indicating an account was suspended/content was deleted	List of content/accounts deleted by a platform linked to misinformation policies or by users themselves following a notification from the platform + pre-existing data before the suspension of an account (content, engagement, etc.) or deletion of content (engagement).		This data is needed to investigate the (direct and indirect) effects of suspending accounts and deleting content in relation to misinformation. Do the targeted pages, accounts, groups or channel migrate to other platforms? How much enagegement did the removed content in relation to a platforms' misinformation intervention receive?
5. Algorithmic recommendations (misleading content)	**Proportion of views & followers**from **direct search** for the content and proportion of views resulting an **algorithmic recommendation**		whether and how recommendation algorithm excludes or downranks: content that was found to be misleading or accounts that have repeatedly shared misinformation. + How algorithms account for the multiple appearances of the same piece of content (e.g. same video within a post with different urls)	This data is needed to understand whether platforms are improving their recommendation algorithms to recommend more credible content. Does downranking (misleading) content result in decreased audiences?
6. Demonetization of content/accounts spreading misinformation	Whether an account was demonetized for policy violations + specific policy violated	List of content/accounts demonetized by a platform, because identified as misleading/spreading misleading content List of content/accounts that have been considered for demonetization by the platform but have not been demonetized (borderline content)		This data is needed to understand whether demonetizing content identified as misleading, leads to a change in posting behavior of the targeted pages, accounts or channels. Does demonetizing (misleading) content thin audiences ?

Annex 2: Quick access to community guidelines & platform misinformation policies, for Facebook, Twitter and YouTube

Ressource	Platform	Link
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/
		https://law.yale.edu/yls-today/news/facebook-data-transparency-advisory-group-releases-final-report
		transparency.twitter.com/en/reports.html
		transparencyreport.google.com/?hl=en
Policy regarding Covid-19		https://www.facebook.com/help/230764881494641/
		help.twitter.com/en/rules-and-policies/medical-misinformation-policy
		blog.twitter.com/en_us/topics/company/2021/updates-to-our-work-on-covid-19-vaccine-misinformation
		blog.twitter.com/en_us/topics/company/2020/covid-19
		support.google.com/youtube/answer/9891785
Fact-checking policy		facebook.com/journalismproject/programs/third-party-fact-checking/how-it-works
		support.google.com/youtube/answer/9229632
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/#policies
		blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/
Strike System		See in the following link the section What is the number of strikes a person or Page has to get to before you ban them?
		about.fb.com/news/2018/08/enforcing-our-community-standards/
		https://transparency.fb.com/en-gb/enforcement/taking-action/counting-strikes/ (updated July 29,2021)
		See in the following links the section: Account locks and permanent suspension
		help.twitter.com/en/rules-and-policies/election-integrity-policy
		help.twitter.com/en/rules-and-policies/medical-misinformation-policy
		https://support.google.com/youtube/answer/2802032?hl=en
Account suspension		
		https://help.twitter.com/en/managing-your-account/suspended-twitter-accounts
Flags, Notices and Information Panels		https://www.facebook.com/business/help/341102040382165
		https://help.twitter.com/en/rules-and-policies/notices-on-twitter
		https://support.google.com/youtube/answer/9004474?hl=en
Summary of ressources, last accessed on July 5, 2021		