COVIDCHECK

ASSESSING THE IMPLEMENTATION OF EU CODE OF PRACTICE ON DISINFORMATION IN RELATION TO COVID-19

Institiúid DCU um Meáin Todhchaí, Daonlathas agus Sochaí
DCU Institute of Future Media, Democracy and Society

ÚDARÁS CRAOLACHÁIN NA hÉIREANN
BROADCASTING AUTHORITY OF IRELAND

Eileen Culloty
Kirsty Park
Trudy Feenane
Charis Papaevangelou
Alex Conroy
Jane Suiter

Acknowledgements

The authors wish to acknowledge Ciarán O'Connor and the Institute for Strategic Dialogue for their contribution to the research; Kevin Doyle for his assistance with the report; and all those who provided expert insights.

DCU Institute for Future Media, Democracy and Society

Directed by Prof. Jane Suiter, the Institute for Future Media, Democracy and Society (FuJo) is a research centre located in Dublin City University's School of Communications. FuJo's multidisciplinary research investigates how to counter digital pathologies including disinformation and digital hate; how to enhance public participation through democratic innovations; and how to secure the sustainability of high-quality journalism.

www.fujomedia.eu

Acknowlegements	02
Contents	03
Forewords	04
Executive summary	06
Introduction and Background	10
Methodology	12
Analysis of reported actions	16
Automated analysis of reporting	24
Case study: Facebook	28
Case study: TikTok	36
Case study: Automation and AI	40
Conclusion and recommendations	42

Forewords

BAI Foreword

The Code of Practice on Disinformation ("The Code") was adopted by the major online platforms operating in the EU at the end of 2018. In the intervening period the world has witnessed some historic events including the COVID-19 pandemic, the Black Lives Matter Movement and significant electoral contests in Ireland, UK, Europe, and the USA. The importance of social media in our information ecosystem has also continued to grow as evidenced in the 2021 Reuters Digital News Report Ireland. The need for increased oversight of, and accountability by, social media platforms is now accepted and legislative underpinning for this is progressing at national and European levels. The BAI is playing a leading role in this process and is committed to supporting the establishment of an effective regulatory regime in Ireland that will serve the needs of Irish and European citizens. The implementation of the Code is part of this process, and this experience will provide a useful reference point for future regulatory engagement with the signatories.

CovidCheck is the third monitoring report that has been commissioned by the BAI and prepared by the DCU Institute for Future Media, Democracy and Society (FuJo), on the implementation of the Code in Ireland. Each of these reports has been part of a larger monitoring process undertaken by European Regulators Group for Audiovisual Media Services (ERGA), at the request of the EU Commission. As with the first two reports, the authors of CovidCheck conclude that, while the Code is a significant first step in fighting disinformation, significant weaknesses in terms of structure, content and enforcement remain to be addressed. This conclusion also underpins the guidance issued by the EU Commission in May 2021 on how the Code should be strengthened by the signatories to become a more effective tool in fighting disinformation.

Work on the revised Code is currently underway and the BAI believes that the analysis and recommendations in CovidCheck can make a valuable contribution to this process. As with the previous reports, the authors highlight concerns over the quality and transparency of the information presented by the signatories in their implementation reports. The ongoing lack of any meaningful country-specific data, and the slow pace at which previous monitoring recommendations are being implemented, are particularly noteworthy findings. While the scale of the problem of disinformation, particularly during the COVID-19 crisis, is generally acknowledged, the case studies in this report highlight specific issues relating to an inconsistent approach to content labelling and comment moderation. In addition, the lack of metrics on the promotion of authoritative content and the unchecked propagation of groups (particularly on Facebook) involved in circulating disinformation are important findings. The recommendation for enhanced transparency in relation to the operation of AI merit attention by all stakeholders. The revised Code will need to reflect significant enhancements in terms of its structure and content, and on the accountability of platforms arising from the Code, if it is to address the concerns highlighted by the monitoring reports to date.

I endorse the recommendation by FuJo in relation to the resources required for effective monitoring of the Code. The BAI has commissioned this work to date in the context of its commitments to ERGA and its statutory requirement to undertake research that promotes plurality. However, more systematic and detailed monitoring will require significantly increased resources.

Engaging in strategic partnerships has always underpinned the BAI's approach to its activities and our relationship with FuJo is a good example of this philosophy in practice. I would like to thank the team for their work on this report and welcome the decision by the European Commission to make this institution one of the European Digital Media Observatory (EDMO) hubs.
Finally, I would like to acknowledge the work of the BAI team that has been working on this project and on related activities in the fight against disinformation. The BAI has consistently highlighted the need for co-ordinated action nationally and internationally in this area and has ongoing engagement with stakeholders in Ireland, the EU and internationally to further its work on disinformation. This global challenge, we believe, requires a coordinated international response, and the BAI remains committed to playing its part in this regard.

Celene Craig
Deputy CEO
Broadcasting Authority of Ireland

September 2020

DCU FuJo Foreword

The COVID-19 crisis underscored the harmful impacts of online disinformation and the need for effective responses. As we move on from that crisis, we must remain mindful of the threat posed by disinformation and devise appropriate mechanisms to counteract it. Since 2018, the EU's self-regulatory Code of Practice on Disinformation has offered a means for signatories to provide transparency about their efforts to combat disinformation. To date, DCU FuJo has published three reports assessing the implementation of the Code. Unfortunately, many critical shortcomings have not been addressed since the Code was adopted. We hope the findings and recommendations of the current report contribute to the strengthening of the Code including the development of robust procedures for reporting and monitoring.

Ireland has a particular obligation to exercise oversight in this area as many technology companies maintain their European headquarters here. Earlier this year, DCU FuJo and DCU ABC made a joint submission to Ireland's Joint Committee on Media, Tourism, Arts, Culture, Sport and the Gaeltacht regarding the General Scheme of the Online Safety and Media Regulation Bill. Our submission highlighted the need to ensure there is a specific responsibility to tackle harmful disinformation and the need for greater transparency and accountability surrounding automated decision-making, among other areas.

DCU FuJo is committed to addressing the problem of disinformation and the need for democratic resilience. It coordinates the Ireland Hub for the European Digital Media Observatory and H2020 Provenance on content verification. As a partner on H2020 EUComMeet, we are investigating whether deliberative innovations can improve democratic legitimacy and foster greater perspective-taking and empathy. Ultimately, we believe a whole-of-society approach is needed to address the difficult problem of disinformation and a strengthened Code of Practice has an important role to play in that.

We are grateful to the BAI for their support and to all those who contributed to this research.

Prof. Jane Suiter,
Director
Institute for Future Media, Democracy and Society
Dublin City University

September 2020

Executive Summary

Commissioned by the Broadcasting Authority of Ireland, this report presents a systematic analysis of the transparency reports submitted by Facebook, Google, Microsoft, Mozilla, TikTok, and Twitter in response to the European Commission's June 2020 Communication on tackling COVID-19 disinformation. The researchers analysed 47 reports that were submitted between August 2020 and April 2021. The analysis involved manual coding to identify individual actions and an automated textual analysis to identify themes and patterns. In addition, the report presents findings from case studies investigating the implementation of COVID-19 policies on Facebook and TikTok and the signatories' transparency regarding the use of AI and automation.

Analysis of reported actions

Across the 47 reports, 1114 individual actions were identified. Google reported the highest number of actions (387). Among the signatories that operate a social media platform (i.e. excluding Mozilla), TikTok reported the lowest number of actions (87). However, direct comparisons are complicated by the different size and nature of the signatories' operations.

Only 32 percent of the 1114 actions were new initiatives, which means the same actions were reported multiple times across the reports, sometimes with new updates or information. The majority (58%) of Facebook's reported actions concerned new initiatives whereas only 12 percent of Google's reported actions were new.

A quarter of all actions concerned the promotion of authoritative content such as links to information by the World Health Organisation (WHO) or national health authorities. The next most common action areas were advertising responses (17%) and blocking, removing or demoting content (13%).

Just over a quarter of all actions were undertaken in collaboration with third-party organisations such as the WHO. Among the signatories, TikTok reported the highest level of collaborative actions at 46 percent followed by Twitter at 38 percent.

Although signatories were asked to report on policies and actions that addressed COVID-19 disinformation, the reported actions were sometimes unrelated to the topic. Twitter and Facebook, in particular, often reported actions with only a tenuous link to COVID-19.

The Commission requested data relating to the EU and at a Member State level. However, the regional application was unclear or unstated for 40 percent of actions. Regarding the EU, 34 percent of the actions covered all of the EU while 12 percent covered some but not all EU Member States. Google and Microsoft reported the lowest number of actions where the regional application was unstated or unclear and the highest number of actions that were stated to be applicable to all EU Member States.

Overall, 32 percent of actions reported outcomes, metrics, or results. However, this figure does not take into account the reporting of new actions where outcomes were not yet available. More than two-thirds of Mozilla's actions and more than half of TikTok's actions were reported with outcomes. Less than a third of actions reported by Google, Microsoft, and Twitter and just over a third of Facebook actions included information about outcomes.

Most actions with reported outcomes did not include EU-specific data. When EU-specific data was reported, it variously referred to an EU aggregate (e.g. 24 million views across the EU), a partial

EU breakdown (e.g. two million in Germany and three million in France); and, less commonly, a full breakdown by EU Member States.

Automated analysis of reporting

An automated textual analysis was applied to gain a broader understanding of the themes, relevance, and level of repetition in the reports. Thematically, a substantial amount of TikTok's reporting addressed metrics including click-through rates (CTR), impressions, and views. This is consistent with the manual coding, which found that more than half of TikTok's actions were reported with outcomes. Reporting by Google and Facebook prioritised work on account removals and combatting coordinated networks and campaigns. Twitter put emphasis on its efforts to serve the research community while Microsoft emphasised its work on advertising.

In terms of relevance, COVID-19 related keywords were most prominent in Microsoft's reports while vaccine-related keywords were most prominent in TikTok's reports. A Labbé distance analysis found that the reports submitted by Google, Facebook, Microsoft and Twitter had high levels of repetition. Although this could indicate a coherent and consistent style of reporting, it is clear from a close reading of the reports that there was a considerable amount of unnecessary repetition in some reports.

Case study: Facebook

In the absence of country-level data to verify reported actions, DCU FuJo cooperated with the Institute for Strategic Dialogue (ISD) to undertake a case study of Irish Facebook Groups and Pages known to propagate COVID-19 vaccine misinformation. The analysis sought to verify the implementation of Facebook's reported actions and policies.

The number of posts about COVID-19 vaccines declined sharply in the Groups after January, which may be indicative of strengthened enforcement of content policies. The analysis identified 35 instances of false claims that had been debunked by Facebook's factchecking partners, but the factchecks were not applied. From a user perspective, the application of content policies gives rise to inconsistencies as the same claims are factchecked in some circumstances, but not in others.

Similar inconsistencies were evident in the application of content labels. Although the vast majority of posts about COVID-19 vaccines carried an appropriate label, there were numerous instances of unlabelled posts about COVID-19 vaccines. Moreover, users can choose not to view labels.

There was a lack of clarity about whether content removal policies applied to comments as well as posts. Content that merited removal was more likely to feature in comments than in posts. Moreover, Groups were more likely than Pages to host posts that merited removal.

Of the 22 Groups analysed, eighteen carried a notification to warn new members they were joining a Group that had Community Standards violations. Only three carried the educational pop-up intended to direct users to credible information from health organisations.

Group admins were allocated greater responsibility for moderating content. However, the admins were themselves organisers of anti-lockdown and anti-vaccine Groups and proponents of disinformation. There was evidence of admins operating multiple Facebook accounts to circumvent sanctions including account blocking and suspensions.

The Groups and Pages were interlinked with other platforms that are not signatories to the Code. Positioned within a network of platforms, Facebook was discussed by users as a strategic means to gain followers who were then directed to platforms with minimal content moderation.

Case study: TikTok
A limited analysis of TikTok verified claims about the promotion of authoritative content and the application of content labels. TiKToK stated that when "users search for coronavirus-related topics" they will "find videos from verified accounts that are providing trusted information from credible sources." The analysis found that search terms related to COVID-19 (e.g. covid) generally returned videos from authoritative sources. However, vaccine search terms (e.g. vaccination) did not return videos from recognised authoritative sources. It appears that the promotion of authoritative content only applied to vaccine search results when a COVID-19 specific term was also included. Moreover, anti-vaccine disinformation featured prominently in search results. TikTok also reported that labels would be applied to COVID-19 content. However, among the top 20 videos for #covid, #vaccine, and #vaxx, only some of the videos about COVID-19 were appropriately labelled. The unlabelled content included anti-vaccine disinformation and conspiracy theories. Notably, the user comments accompanying videos were a source of disinformation.

Case study: AI and Automation
Although signatories were not asked to report on automation, the Commissions' May 2021 Guidance on Strengthening the Code of Practice on Disinformation urges platforms to harmonise their content moderation practices with the relevant provisions of the EU's proposed AI Act. With the exception of Mozilla, each signatory reported actions in areas where automation plays a key role such as the automatic detection of content for labelling. However, signatories did not provide a consistent account regarding their use of AI, automation, or machine learning. Notably, Twitter did provide data for actions arising from automated solutions, but only in relation to advertising and, more specifically, concerning content that contained words related to COVID-19. More generally, signatories referenced the use of AI in relation to dealing with manipulative behaviour, fraudulent commercial content, and deepfakes. However, the percentage of actions (e.g. content removal) that were taken as the result of automated versus human moderation remains unclear. Moreover, as the dominant language of the AI systems used by the signatories is English and as Ireland is the only English-speaking country in the post-Brexit EU, further information is required regarding the adequacy of signatories' systems for detecting and addressing disinformation in languages beyond English.

Recommendations
Recommendation 1: We recommend that reporting be standardised, as far as possible, to ensure necessary and relevant information is provided and in a manner that facilitates monitoring. For each action, we recommend that signatories clearly state: the specific policy, if any, associated with the action; the relevance of the action to the Code or the specific information requested; whether the reported action is a new initiative or part of an ongoing initiative; the regional application of the action and, in particular, the application across EU Member States; and whether metrics or other outcome data are available at the level of EU Member States.

Recommendation 2: We recommend that signatories provide clear definitions of relevant policies to combat disinformation, clear definitions of common terms, and how those terms are operationalised on their services. This information should be available as part of a reference resource.

Recommendation 3: We recommend that relevant stakeholders introduce a framework to address disinformation in comments that is consistent with Article 10 of the European Convention on Human Rights and the principle of freedom of opinion.

Recommendation 4: Expanding on Article 23(2) of the Digital Services Act, which requires more detailed data regarding platforms' active users, we recommend that clear parameters be defined for the reporting of granular data about specific action areas and in relation to EU Member States. A more comprehensive picture of the signatories' actions on specific types of content related to disinformation is necessary to evaluate the effectiveness of the Code and signatories' actions.

Recommendation 5: We recommend that meaningful KPIs be defined for the reporting of results and outcomes in relation to key areas including: content labels, content and account removals, factchecking, and media literacy campaigns. We also recommend that signatories report on their own efforts to measure the efficacy of these actions and provide data to independent researchers to verify that efficacy.

Recommendation 6: We recommend that the original commitment to establish an independent auditor be implemented under the revised Code. Further, we recommend that signatories provide adequate funding and resources to support this position, which will contribute to the monitoring work of ERGA and EDMO.

Recommendation 7: We recommend that standardised procedures to verify the implementation of actions be agreed for future monitoring. This will ensure consistency in monitoring and provide an important counterpoint to the signatories' reported metrics.

Recommendation 8: We recommend that signatories report on their use of automated systems to combat disinformation including an explanation of what systems are used, what languages are covered, what kinds of disinformation they are trained to detect, and what risk assessments have been conducted on the AI systems used to tackle disinformation. Additionally we propose that the European Commission specifically articulates the need for risk assessments related to disinformation in the strengthened Code.

Recommendation 9: We recommend that signatories embrace the need for transparency and data-sharing with researchers, as well as expand and improve services that allow researchers to access data. Moreover, we suggest that the Commission create a clear regulatory framework for accessing data for research on disinformation and further expand the scope of its current proposal to include more stakeholders, including members of civil society organisations, rather than just university-affiliated researchers.

Introduction & Background

Throughout the COVID-19 pandemic, false claims about the virus, public health measures, and vaccines have circulated widely online. The prevalence of false claims has contributed to public confusion and potentially undermined efforts to stop the spread of the virus. In response to these concerns, the European Commission issued a Joint Communication in June 2020: ***Tackling COVID-19 disinformation – Getting the facts right***[1]. It established a COVID-19 monitoring programme for the six platform signatories of the EU Code of Practice on Disinformation[2]: Facebook, Google, Microsoft, Mozilla, TikTok, and Twitter. These signatories were asked to report on their policies and actions to address COVID-19 disinformation with a particular emphasis on the following areas:

Initiatives to promote authoritative content at the EU and Member State level: including data on the actions taken to promote information from national and international health agencies, national and EU authorities, as well as professional media.

Initiatives and tools to improve users' awareness: including data about the implementation of policies to inform users when they interact with disinformation.

Manipulative behaviour: including reporting on all instances of social media manipulation, malign influence operations or coordinated inauthentic behaviour detected on their services. Signatories were also asked to cooperate with EU Member States and institutions in order to facilitate the assessment and attribution of disinformation campaigns and influence operations.

Data on flows of advertising linked to COVID-19 disinformation: including the provision of data, broken down by Member State where possible, on policies undertaken to limit advertising placements related to COVID-19 disinformation on their services. As applicable, signatories were asked to provide data on policies to limit such advertising placements on third-party websites.

In addition, signatories were asked to "broaden and intensify their cooperation with factcheckers and offer access to their factchecking programmes to organisations in all EU Member States – as well as in its neighbourhood – for all languages". In anticipation of the roll-out of vaccination programmes, signatories were subsequently asked to provide information on actions taken to combat false information about vaccines.

Reporting and monitoring

Signatories were asked to provide a baseline report detailing relevant actions implemented prior to 31 July 2020. Thereafter, signatories were asked to submit monthly transparency reports. The monthly reporting period, September to December 2020, was extended to June 2021 due to the continuation of the pandemic. The reports were published online by the European Commission[3].

The European Regulators Group for Audiovisual Media Services (ERGA) was tasked with evaluating the implementation of the programme. In May 2021, ERGA published an interim report[4], which found that signatories had intensified their efforts to counter COVID-19 disinformation and had provided a useful overview of their actions in the transparency reports. However, ERGA observed that the reported actions could not be verified at the country level and that the absence of country-specific data impeded efforts to monitor and assess the effectiveness of the reported actions. In 2021, ERGA asked the signatories to provide country-specific and disaggregated data. Although the signatories provided

1. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020JC0008
2. https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation
3. https://digital-strategy.ec.europa.eu/en/library/latest-set-reports-and-way-forward-fighting-covid-19-disinformation-monitoring-programme

additional context for the transparency reports, the information provided fell short of what was requested.

The present report

Ireland is one of the ten countries participating in ERGA's evaluation. As part of its role within ERGA, the Broadcasting Authority of Ireland (BAI) commissioned the Institute for Future Media, Democracy and Society at Dublin City University (DCU FuJo) to undertake research on the signatories' transparency reports. This followed two previous reports on the Code of Practice published by the BAI and DCU FuJo: ElectCheck 2019[5] and CodeCheck 2020[6].

This report presents findings from a systematic analysis of the 47 reports submitted by the signatories between August 2020 and April 2021. The analysis involved manual coding of the transparency reports to identify and parse the individual actions reported by the signatories and an automated textual analysis to identify themes and patterns across the sets of reports. In addition, the reports were reviewed to assess transparency regarding the use of AI and automation during COVID-19.

In the absence of country-level data to verify reported actions, DCU FuJo cooperated with the Institute for Strategic Dialogue (ISD) to undertake a case study of Irish Facebook Pages and Groups known to propagate COVID-19 vaccine disinformation. The analysis examined the application of factchecks and content labels, content removals, and the implementation of Facebook policy changes between January and April 2021. Notably, the capacity to verify Facebook's reported measures is largely due to Facebook's provision of the research tool CrowdTangle and the work undertaken by ISD. It was not possible to perform a comparable analysis of the other platform signatories. Nevertheless, a small and limited case study of TikTok was undertaken.

The research methodology and findings are presented in the following sections. The report concludes with recommendations for the reporting and monitoring of the Code of Practice on Disinformation.

4. ERGA (2021). Interim Report on Monitoring of the COVID19 Disinformation

5. www.bai.ie/en/new-report-on-political-social-media-ads-identifies-inconsistencies-in-datasets-and-definitions/

6. www.bai.ie/en/new-report-highlights-inconsistencies-across-digital-platforms-in-tackling-disinformation/

Methodology

Manual coding of reported actions

The reports varied considerably in terms of structure, presentation, and information provided. To compensate for the lack of a standardised reporting format and to facilitate analysis, each report was manually coded to identify and parse individual actions and any associated information. This manual coding covered the 47 reports submitted between August 2020 and April 2021.

Using individual actions as the unit of analysis, a codebook was developed to focus on three areas: the nature of the reported action; the region of application; and the reporting of results (see Text Box 1). Each action was categorised into one of the following action types:

- ***Advertising responses***: e.g. actions addressing problematic advertising, profiteering, or advertising grants.

- ***Blocking, removing or demoting content***: e.g. actions addressing the visibility of disinformation.

- ***Combatting organised manipulation***: e.g. actions addressing disinformation campaigns, malign influence operations, or coordinated inauthentic behaviour.

- ***Factchecking and labelling content***: e.g. actions addressing the veracity of content claims, supporting users' evaluation of content, or support for factcheckers.

- ***Promoting authoritative content***: e.g. actions addressing the promotion of public health messaging and support for public health authorities and other reliable sources about COVID-19 or COVID-19 vaccines such as the World Health Organisation.

- ***Public health or media literacy***: e.g. actions promoting media literacy or public health issues such as mental health during the pandemic.

- ***Research responses***: e.g. actions empowering the research community such as funding, API access, or provision of data.

- ***Other***: any action outside the above categories.

The coding was undertaken in two phases. In phase one, each report was coded leading to a review of the codebook and a discussion of any borderline cases. In phase two, each report was re-coded using the finalised codebook. Each action was reviewed by at least two coders.

Text box 1: Codebook for manual coding of reported actions.

For each action, the following information was recorded:

1. Text extract describing the action
2. Platform division (if relevant)
3. Action type (select from list):
    - advertising responses
    - blocking, removing or demoting content
    - combating organised manipulation
    - factchecking and labelling content
    - promoting authoritative content
    - public health or media literacy initiatives
    - research responses
    - other
4. A new action: yes/no
5. Involvement of partner organisations: yes/no
6. Stated region of application (select from list):
    - Global
    - Full EU
    - Partial EU
    - Non-EU
    - Unstated/ unclear
7. Named EU countries of application: list as applicable
8. Results reported: yes/no
9. Results reported at EU aggregate level: yes/no
10. Results reported at country level: yes/no
11. Countries for which results are provided: list as applicable

Automated analysis of reporting

As a supplement to the manual coding, an automated analysis of the reports was conducted using IRaMuTeQ[7], an open-source software developed by LERASS[8], a laboratory at the University of Toulouse 3 – Paul Sabatier, for the statistical analysis of text. Two types of analysis were performed: Specific Words Analysis and Labbé Intertextual Distance Index. Specific Words Analysis examines the frequency and uniqueness of specific words across a textual corpus. It was applied to identify the words, and subsequently themes, that the signatories prioritised in their transparency reports. As such, it provides an indication of the thematic commonalities and differences between each signatory's set of reports. Labbé Intertextual Distance Index estimates the extent to which a set of texts is similar or different from a lexicological perspective. It was applied to examine the textual variance across each signatory's set of reports. As such, it provides an indication of the level of repetition in the reporting. In terms of comparing the signatories' reports it should be noted that there were only two Mozilla reports and the TikTok reports were significantly shorter than other reports. These factors should be taken into consideration when assessing the findings of the automated analysis.

Case study: Facebook

In the absence of country-level data to verify reported actions, DCU FuJo consulted with European factcheckers and cooperated with the Institute for Strategic Dialogue (ISD) to undertake a case study of Irish Facebook Groups and Pages known to propagate COVID-19 vaccine misinformation. ISD provided analysis that was conducted using Facebook's CrowdTangle tool. In summary, a comprehensive list of keywords relating to COVID-19 vaccines was used to search CrowdTangle for Irish Pages and Groups that posted, shared, or hosted content containing false information about COVID-19 vaccines. The available data included: time and date of publication; the Group or Page sharing the posts; the text of posts; any associated links or media; and engagement metrics. Importantly, the data included factual information about vaccines, such as links to reliable sources, as well as vaccine misinformation and disinformation. The data covers a four month period: January to April 2021.

To gain a detailed understanding of the data, one week from each month was sampled for in-depth analysis regarding the application of factchecks, content labels and the removal of content (see Table 1). Claims that were judged to merit removal were reviewed by at least three researchers.

Application of factchecks: The sampled posts were examined to identify: (i) posts that had been factchecked by a Facebook factchecking partner and (ii) inconsistencies in the application of factchecks.

Application of content labels: In March 2021, Facebook announced[9] that it would roll "out labels on all posts generally about COVID-19 vaccines". Sampled posts from March and April were examined to (i) verify the application of labels to relevant content and (ii) identify any issues or inconsistencies.

Content removals: In December 2020, Facebook announced[10] that it would start removing false claims about COVID-19 vaccines that have been debunked by leading global health organisations. The sampled posts were examined to assess whether or not they merited removal on the grounds outlined by Facebook's COVID-19 and vaccine policy updates and protections.[11] As this policy refers broadly to "content" and "claims", the analysis examined the sampled posts as well as the comments accompanying those posts. Any posts and comments deemed to be in violation of Facebook's policy were recorded for further review.

Verification of other actions: Additional analysis was conducted on a sample of Groups and Pages to assess the extent to which Facebook policy changes were implemented. Specifically, this included the application of educational pop-ups[12] at the top of COVID-19 related Groups (announced April 2020); the provision of notifications to users[13] before they join a Group that has accrued Community Standards violations and actions to reduce the reach and visibility of those Groups[14] (both actions announced March 2021).

7. http://www.iramuteq.org/
8. https://www.lerass.com/
9. https://about.fb.com/news/2021/03/mark-zuckerberg-announces-facebooks-plans-to-help-get-people-vaccinated-against-covid-19/
10. https://about.fb.com/news/2020/12/coronavirus/#latest
11. https://www.facebook.com/help/230764881494641/
12. https://about.fb.com/news/2020/12/coronavirus/

Case study: TikTok

A more limited case study of TikTok was undertaken to verify claims about the promotion of authoritative sources and the application of content labels. Eight sets of searches for Covid-19 and vaccine related terms were conducted in July 2021. The following keywords were used: COVID, COVID-19, coronavirus, vaccine, COVID vaccine, and vaccination. To account for variances in user history, the searches were performed across six TikTok accounts maintained by the researchers. Each researcher recorded whether: (i) the top results were from authoritative sources such as the WHO and (ii) whether a clickable link to access further information was visible on screen without scrolling. In addition, the top 20 TikTok posts available under the hashtags #covid, #vaccine and #vaxx were examined to verify whether they were labelled appropriately. At the time of analysis, #covid had a total of 26.3 billion views, #vaccine had a total of 4.2 billion views, and #vaxx had a total of 8.3 million views on TikTok.

Case study: AI and Automation

Neither the Code of Practice on Disinformation nor the European Commissions' June 2020 Communication on COVID-19 disinformation address the use of Artificial Intelligence (AI) in content moderation. However, the Commissions' May 2021 ***Guidance on Strengthening the Code of Practice on Disinformation***[15] urges platforms to harmonise their content moderation practices with the relevant provisions of the EU's proposed AI Act[16]. In this context, the transparency reports submitted by the signatories were reviewed to identify any references to the use of AI and automation in their efforts to tackle COVID-19 disinformation.

Table 1: Overview of the sampled data

Sampled Week	Group Posts	Pages Posts
1-8 January 2021	138	33
22-28 February 2021	28	28
22-28 March 2021	27	29
15-21 April 2021	26	37

13. https://about.fb.com/news/2021/03/changes-to-keep-facebook-groups-safe/
14. https://about.fb.com/news/2021/03/changes-to-keep-facebook-groups-safe/
15. https://digital-strategy.ec.europa.eu/en/library/guidance-strengthening-code-practice-disinformation
16. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

Analysis of Reported Actions

Reported actions

Between August 2020 and April 2021, the signatories collectively submitted 47 reports. Each individual action referenced in the reports was coded as a discrete item, resulting in the identification of 1114 individual actions (see Table 2). The number of actions reported by the signatories varied greatly. However, direct comparisons are complicated by the different size and nature of the signatories' operations. For example, Mozillia is the only signatory that does not operate a social media platform and confined its updates to just two reports. In contrast, Google operates a wide range of platforms and applications including: Ads, Images, Maps, News, Play, Search, and YouTube[17]. Similarly, Facebook operates three large-scale platforms - Facebook, Instagram, and WhatsApp - as well as a large advertising business.

Google reported the highest number of actions (387). TikTok reported the lowest number of actions (87) among the signatories that operate a social media platform. Unsurprisingly, the highest number of actions were reported in August 2020 when signatories were asked to provide baseline reports outlining all actions undertaken to counteract COVID-19 disinformation between the start of the pandemic and the end of July 2020.

New and continuing actions

There was considerable repetition in the reporting of actions across the time period. In some instances, previous actions were referenced to provide new updates or to report an extended application. In other instances, the same action was presented in successive reports without new information. To gain a better understanding of how many new actions were undertaken in response to the pandemic, each action was coded as a new action or a continuing action, beginning with the baseline reports (see Table 3). If an action was in place prior to the COVID-19 pandemic, such as a media literacy initiative that began in 2019, then it was classified as a continuing action when it was reported in the baseline report of August 2020. Of the 1114 reported actions, only 351 (32%) were new while 763 (68%) were repetitions of previously reported actions. Notably, the majority (58%) of actions reported by Facebook concerned new actions whereas only 12 percent of Google's reported actions concerned new initiatives.

Broken down by month (see Table 4), the vast majority of actions presented in the baseline reports of August 2020 were new initiatives. In subsequent reports, Facebook, Twitter and, to a lesser extent, TikTok regularly presented new actions while Google and Microsoft largely presented continuing actions.

17. Google did not report actions relating to its browser Chrome.

Table 2: Individual actions reported by the signatories

	Facebook	Google	Microsoft	Mozilla	TikTok	Twitter	**Total**
Aug 20	51	34	23	12	10	30	**160**
Sep 20	24	38	24	-	3	17	**106**
Oct 20	30	39	18	-	6	22	**115**
Nov 20	30	40	21	-	7	29	**127**
Dec 20	17	44	20	-	7	28	**116**
Jan 21	11	46	19	-	13	16	**105**
Feb 21	9	47	19	4	12	25	**116**
Mar 21	25	49	19	-	15	22	**130**
Apr 21	40	50	19	-	14	16	**139**
Total	**237**	**387**	**182**	**16**	**87**	**205**	**1114**

Table 3: New actions reported by each signatory

	Facebook	Google	Microsoft	Mozilla	TikTok	Twitter	**Total**
All Actions	237	387	182	16	87	205	**1114**
# New Actions	137	47	25	12	28	102	**351**
% New Actions	**58%**	**12%**	**14%**	**75%**	**32%**	**50%**	**32%**

Table 4: New actions reported by month

		Facebook	Google	Microsoft	Mozilla	TikTok	Twitter
Aug 20	#	50	31	21	12	10	30
	%	98%	91%	91%	100%	100%	100%
Sep 20	#	13	3	0	-	0	9
	%	54%	8%	0%	-	0%	53%
Oct 20	#	14	1	0	-	2	11
	%	47%	3%	0%	-	33%	50%
Nov 20	#	21	1	3	-	2	14
	%	70%	3%	14%	-	29%	48%
Dec 20	#	5	3	0	-	2	15
	%	29%	7%	0%	-	29%	54%
Jan 21	#	4	2	0	-	7	5
	%	36%	4%	0%	-	54%	31%
Feb 21	#	2	1	1	-	3	7
	%	22%	2%	5%	0%	25%	28%
Mar 21	#	11	2	0	-	1	8
	%	44%	4%	0%	-	7%	36%
Apr 21	#	17	3	0	-	1	3
	%	43%	6%	0%	-	7%	19%

Action types

While signatories were asked to report on all policies and actions to address COVID-19 disinformation, the following areas were highlighted by the Commission: initiatives to promote authoritative content at the EU and Member State level; initiatives and tools to improve users' awareness; manipulative behaviour on their services; data on flows of advertising linked to COVID-19 disinformation; and support for factchecking. A quarter of all actions concerned the promotion of authoritative content such as links to information by the WHO or national health authorities (see Table 5). The next most common action areas were advertising responses (17%) and blocking, removing or demoting content (13%).

The emphasis placed on these action areas varied across the platforms (see Table 6). Promoting authoritative content was the most common type of action reported by each signatory, with the exception of Facebook. For Facebook, the most common action type related to blocking, removing, or demoting content, which accounted for more than a fifth of all Facebook actions. Almost a quarter of all actions reported by TikTok concerned factchecking and content labelling and one fifth related to blocking, removing, or demoting content. Mozilla did not have actions in a number of categories, reflecting the fact that it does not provide a social media type service.

Apart from Mozilla, all signatories operate significant advertising businesses. They were asked to provide information on advertising linked to COVID-19 disinformation including policies to limit such adverts on their own services and the placement of such adverts on third-party websites. Google reported the largest percentage of actions relating to advertising, which accounted for 23 percent of all its actions.

Collaborations

Just over a quarter of all actions were undertaken in collaboration with third-party organisations such as the WHO (see Table 7). Among the signatories, TikTok reported the highest level of collaborative actions at 46 percent followed by Twitter at 38 percent.

Unsurprisingly, certain types of action attracted a higher level of collaboration including factchecking, the promotion of authoritative content, public health and media literacy, and research (see Table 8). These actions often involved working with factchecking partners, health agencies, and non-profit organisations. Almost half of all research actions were reported as collaborations although the overall number of research actions is relatively low. Notably, only four percent of actions relating to both content moderation (i.e. blocking, removing, demoting content) and organised manipulation were reported as collaborative endeavours.

Table 5: Action types (n=1114)[18]

Action type	# Actions	% of Total
Advertising responses	187	17%
Blocking, removing, or demoting	150	13%
Combating organised manipulation	91	8%
Factchecking and labelling	101	9%
Promoting authoritative content	279	25%
Public health or media literacy	113	10%
Research responses	94	8%
Other	99	9%

Table 6: Types of actions reported by each signatory as a percentage[19]

	Facebook	Google	Microsoft	Mozilla	TikTok	Twitter
	n = 237	n =387	n =182	n =16	n =87	n =205
Advertising responses	14%	23%	18%	-	9%	12%
Blocking, removing, or demoting	21%	13%	5%	-	20%	11%
Combating organised manipulation	8%	8%	14%	-	-	7%
Factchecking and Labelling	11%	2%	15%	-	24%	8%
Promoting authoritative content	15%	25%	36%	44%	43%	20%
Public health or media literacy	9%	12%	5%	6%	3%	15%
Research responses	7%	12%	1%	13%	1%	13%
Other	14%	5%	7%	38%	-	14%

Table 7: Percentage of actions involving collaborations

	Facebook	Google	Microsoft	Mozilla	TikTok	Twitter	**Total**
	n = 237	n = 387	n = 182	n = 16	n = 87	n = 205	**n = 1114**
No collaboration	77%	82%	73%	81%	54%	62%	**74%**
Collaboration	23%	18%	27%	19%	46%	38%	**26%**

Table 8: Collaborative actions by action type

	# Actions	% Collaborative
Advertising responses	187	15%
Blocking, removing, or demoting	150	4%
Combating organised manipulation	91	4%
Factchecking and Labelling	101	39%
Promoting authoritative content	279	39%
Public health or media literacy	113	40%
Research responses	94	48%
Other	99	18%

18. Due to rounding, the total adds to 99%.

19. Due to rounding, the Facebook total adds to 99%.

Relevance to COVID-19

Although signatories were asked to report on policies and actions that addressed COVID-19 disinformation, the reported actions by Facebook, Google, and Twitter were sometimes unrelated to the topic (see Table 9). For example, they reported on marketing workshops that provided support for start-ups and the launch of media literacy campaigns ahead of elections. Twitter and Facebook, in particular, dedicated large segments of their reports to describing charity work and general public-health initiatives. The links to COVID-19 were often tenuous. For example, Twitter reported the launch of a new campaign and emoji to commemorate International Holocaust Memorial Day, which was linked to COVID-19 by referencing the rise in hateful and racist rhetoric during the pandemic.

It should be noted, however, that signatories were asked to report "all instances of social media manipulation, malign influence operations or coordinated inauthentic behaviour detected on their services". This may be interpreted as an invitation to report instances of organised manipulation even when those instances were unrelated to COVID-19. When those cases were removed, Google's percentage of irrelevant actions fell to nine percent, but the Facebook and Twitter percentages remained high at 28 percent each (see Table 10).

Regional application

The Commission requested data relating to the EU and at a Member State level. The actions described in the report concerned a mix of regions (see Table 11). In many instances, it was difficult to discern which regions were covered by the reported actions as no specific region was mentioned or geographic reach was vaguely defined as "available in 32 countries". In total, 13 percent of actions were referenced as having a global reach. Regarding the EU, 34 percent of actions covered all of the EU while 12 percent covered some but not all EU Member States. The regional application was unclear or unstated for 40 percent of actions. It is possible that these cases were intended to be read as global actions. However, as new initiatives are frequently rolled out in a phased manner across markets, these actions were designated unstated or unclear in the absence of clarity in the reports. One percent of actions did not apply to any EU Member State. These mainly concerned the US market and were irrelevant to the scope of Code.

Across the signatories, Google and Microsoft reported the lowest number of actions where the regional application was unstated or unclear and the highest number of actions that were stated to be applicable to all EU Member States (see Table 12). Both Google and Microsoft offered clarity about the regional application of their reported actions. At the beginning of each report, Microsoft stated: "we generally track and report these efforts on a global or EU-wide basis. In those instances where our efforts are limited to a certain Member State, we have stated that below". Similarly, Google frequently clarified that: "unless specified otherwise, the content of this [report] section applies equally to all EU Member States." Consequently, it was generally possible to determine the regional application of actions for these signatories. In contrast, the regional application could not be determined for more than two-thirds of the actions reported by Twitter and Facebook and half of the actions reported by TikTok.

Actions that applied only partially to the EU ranged from single-country actions, such as a media literacy program in France, to actions that applied to most, but not all, Member States. Most of the non-EU actions reported by Facebook and Twitter applied to the United States. Mozilla reported one action applying to Africa; this accounted for six percent of the total due to Mozilla's small number of overall actions.

Table 9: Percentage of actions unrelated to COVID-19

	Facebook n = 237	Google n = 387	Microsoft n = 182	Mozilla n = 16	TikTok n = 87	Twitter n = 205
Unrelated Actions	32%	18%	0%	0%	0%	30%

Table 10: Percentage of actions unrelated to COVID-19 and organised manipulation

	Facebook n = 237	Google n = 387	Microsoft n = 182	Mozilla n = 16	TikTok n = 87	Twitter n = 205
Unrelated Actions	28%	9%	0%	0%	0%	28%

Table 11: Actions reported by region (N=1114)

	% of total actions
Global	13%
Full EU	34%
Partial EU	12%
Non-EU	1%
Unstated/unclear	40%

Table 12: Actions reported by region across the signatories[20]

	Facebook n = 237	Google n = 387	Microsoft n = 182	Mozilla n = 16	TikTok n = 87	Twitter n = 205
Global	22%	13%	12%	38%	5%	5%
Full EU	5%	52%	75%	19%	25%	3%
Partial EU	3%	15%	5%	6%	20%	20%
Non-EU	2%	0%	0%	6%	0%	2%
Unstated/Unclear	67%	20%	7%	31%	51%	70%

20. Due to rounding, some totals do not add to 100.

Reported outcomes

For each action, we recorded whether any outcomes, metrics, or results were reported such as the provision of information about the volume of content that was removed or the number of people who visited an information centre. Overall, 32 percent of actions reported outcomes (see Table 13). However, this figure does not take into account the reporting of new actions - such as a newly launched funding scheme or policy change - where outcomes were not yet available. Nevertheless, there were many instances in which outcomes could have been reported, but no information was provided. In total, more than two-thirds of Mozilla's actions and more than half of TikTok's actions were reported with accompanying outcomes, although the overall number of actions by these signatories was comparatively low. Less than a third of actions reported by Google, Microsoft, and Twitter and just over a third of Facebook actions included information about outcomes.

There are variances across the signatories' reporting of actions within the same category (see Table 14). For example, all signatories provided outcomes for actions relating to the promotion of authoritative content, but the percentage of actions with reported outcomes was low for Google and Twitter, at nine percent and 15 percent respectively. TikTok provided outcomes for all actions relating to blocking, removing or denoting content whereas Facebook provided outcomes for about half of its actions in this area.

As noted, signatories were specifically asked to provide data relating to the EU and at a Member State level (see Table 15). When outcomes were reported about the EU, the data variously referred to an EU aggregate (e.g. 24 million views across the EU), a partial EU breakdown (e.g. two million in Germany and three million in France); and, less commonly, a full breakdown by EU Member States.

Most actions did not include any of these additional levels of results. TikTok's reporting covered four main markets leading to 47% of actions including partial EU results, but along with Facebook and Twitter, none of their actions included a full Member State breakdown. Additionally, some actions reported results which were not directly relevant to COVID-19 or which combined COVID-19 related results with non-COVID-19 related actions such as hate speech.

Table 13: Actions with reported outcomes

	Facebook n = 237	Google n = 387	Microsoft n = 182	Mozilla n = 16	TikTok n = 87	Twitter n = 205	**Total n = 1114**
# reporting outcomes	83	104	55	11	45	64	**362**
% reporting outcomes	35%	27%	30%	69%	52%	31%	**32%**

Table 14: Actions with reported outcomes by action type

	Facebook	Google	Microsoft	Mozilla	TikTok	Twitter
Advertising responses	26%	40%	56%	-	13%	76%
Blocking, removing, or demoting	52%	0%	0%	-	100%	70%
Combating organised manipulation	70%	100%	0%	-	-	71%
Factchecking and Labelling	33%	100%	0%	-	52%	6%
Promoting authoritative content	40%	9%	55%	100%	41%	15%
Public health or media literacy	5%	0%	0%	100%	0%	10%
Research responses	41%	22%	0%	50%	33%	22%
Other	13%	43%	0%	33%	0%	11%

Table 15: . Number of actions reporting data on outcomes relating to the EU

	Facebook n = 83	Google n = 104	Microsoft n = 55	Mozilla n =11	TikTok n = 45	Twitter n = 64
No EU data	66	66	26	4	14	57
Aggregate EU data	13	15	17	1	10	0
Some Member State data	4	16	6	3	21	7
Full Member State data	0	7	6	3	0	0

Automated Analysis of Reporting

The manual coding identified the individual actions covered by each report. However, the space afforded to describing an individual action ranged from one sentence to multiple paragraphs. An automated textual analysis was applied to gain a broader understanding of the themes covered by the reports and the commonalities and differences between the reports submitted by each signatory.

Themes and priorities

An analysis of the most frequently cited words and the typical text segments associated with those words provides an indication of what topics or themes were prioritised by each signatory (see Table 16). It should be mentioned that the software used for analysis 'lemmatises' the words; put simply, it groups together the inflected forms of a word for analysis as a single item. For example, 'person' and 'people' are grouped as a single item.

As noted above, more than half of TikTok's actions were reported with outcomes and TikTok frequently provided data for specific EU Member States. The automated analysis indicates that a substantial amount of TikTok's reporting addressed metrics including click-through rates (CTR), impressions, and views. Figure 1 presents an example of a report segment featuring TiKTok's actions on labeling COVID-19 related videos in some EU Member States. Notably, TikTok's reports were considerably shorter than those submitted by other signatories, which indicates that TikTok emphasised the reporting of metrics and outcomes.

Facebook prioritised its work on removing groups and networks. However, as illustrated by the typical text segments in Figure 2, this reporting tended to offer global figures and often concerned countries beyond the EU. Similarly, Google's reporting emphasised the termination of accounts and efforts to combat coordinated campaigns. Again, reporting provided high-level data that was not tailored for the EU. For example, each report offered broad figures for the number of accounts that were terminated, but with little granular data to assess the relevance for the Code or the EU.

Among the other signatories, advertising was one of the themes prioritised in Microsoft's reporting. Interestingly, this theme did not feature as prominently in reporting by Google and Facebook even though these two signatories are much larger players in the online advertising market. Microsoft was the only signatory for which disinformation featured among the most frequently cited specific words. Mozilla, unsurprisingly, prioritised actions relating to its web browser. Twitter put significant emphasis on its efforts to serve the research community with API access to study data on public conversations (see Figure 3). Twitter also launched a 'product track' for academic research[21], which may partially address the Commission's request for more research data.

21. https://developer.twitter.com/en/products/twitter-api/academic-research

Table 16: Most frequently cited specific words across each signatory's set of reports

Facebook	Google	Microsoft	Mozilla	TikTok	Twitter
Instagram	Google	Linkedin	Mozilla	TikTok	Twitter
person	Youtube	Microsoft	Firefox	video	tweet
Facebook	terminate	Bing	Pocket	view	conversation
remove	channel	company	consumer	CTR	prompt
group	influence	disinformation	pandemic	tag	API
network	finding	advertise	privacy	click	public
standard	consistent	Newsguard	web	impression	organisation
piece	operation	advertisement	browser	notice	good
audience	coordinate	effort	Foundation	month	datum
community	campaign	service	open	user	label

Number of COVID-19 Notice Tag Videos

Month	Italy	Spain	France	Germany	All EU
July 2020	8493	14518	7000	14195	N/a
August 2020	6899	15285	8752	10960	N/a

Figure 1: Example of a data table taken from a TikTok transparency report

May 2020 Coordinated Inauthentic Behavior Report

We removed two networks of accounts, Pages and Groups. One of them - from Tunisia - focused on countries across Francophone Africa, and the other one targeted domestic audiences in the Kurdistan region of Iraq. As noted above, we have not found evidence of COVID-19 focused influence operations. The networks reported below were removed for behaviours that violated our Inauthentic Behaviour Policy.

Figure 2: Excerpt taken from Facebook's July 2020 report

Academic Research product track on the new Twitter API

Since the Twitter API was introduced, academic researchers have used data from the public conversation to study topics as diverse as the conversation on Twitter itself - from state-backed efforts to disrupt the public conversation to floods and climate change, from attitudes and perceptions about COVID-19 to efforts to promote healthy conversation online. Today, academic researchers are one of the largest groups of people using the Twitter API.

Figure 3: Excerpt taken from Twitter's March 2021 report

Relevance and repetition

As noted, many reports offered information that was not directly relevant to COVID-19 disinformation. Each signatory's set of reports were analysed to identify the frequency of keywords relating to the pandemic (e.g. COVID, coronavirus, pandemic) and vaccines (e.g. vaccine, vaccination). COVID-related keywords were most prominent in Microsoft's reports while vaccine-related keywords were most prominent in TikTok's reports (see Figure 4). This is consistent with the impression derived from the manual coding, which found that these two signatories tended to report on relevant actions. It should be noted that vaccines did not become a prominent part of the reports until later in the reporting period, when vaccines were slowly becoming available.

The manual coding identified a considerable amount of repetition in the reporting of actions. A Labbé distance analysis was applied to estimate the level of similarity across each signatory's set of reports. Labbé allocates a value between 0 and 1 where 0 (colored white) indicates complete similarity, 1 (coloured blue) indicates complete dissimilarity and median values indicate modest similarity or dissimilarity (coloured green along with its tonal variations). As shown in Figure 5, the values for Google and Microsoft are close to 0, which indicates considerable repetition across the text of their reports. The values for Facebook and Twitter are all less than 0.5, which indicates a high level of repetition. In contrast, the values for TikTok fluctuate, but its reports are generally less repetitive than those provided by the other signatories. Mozilla's reports are modestly similar, although only two reports were provided.

Similarity across the reports is not necessarily negative. It could, for example, indicate a coherent and consistent style of reporting. However, it is clear from our close reading of the reports that there was a considerable amount of unnecessary repetition that resulted in longer reports and sometimes obscured the identification of new actions. To overcome this issue, a standardised reporting format would be highly desirable.

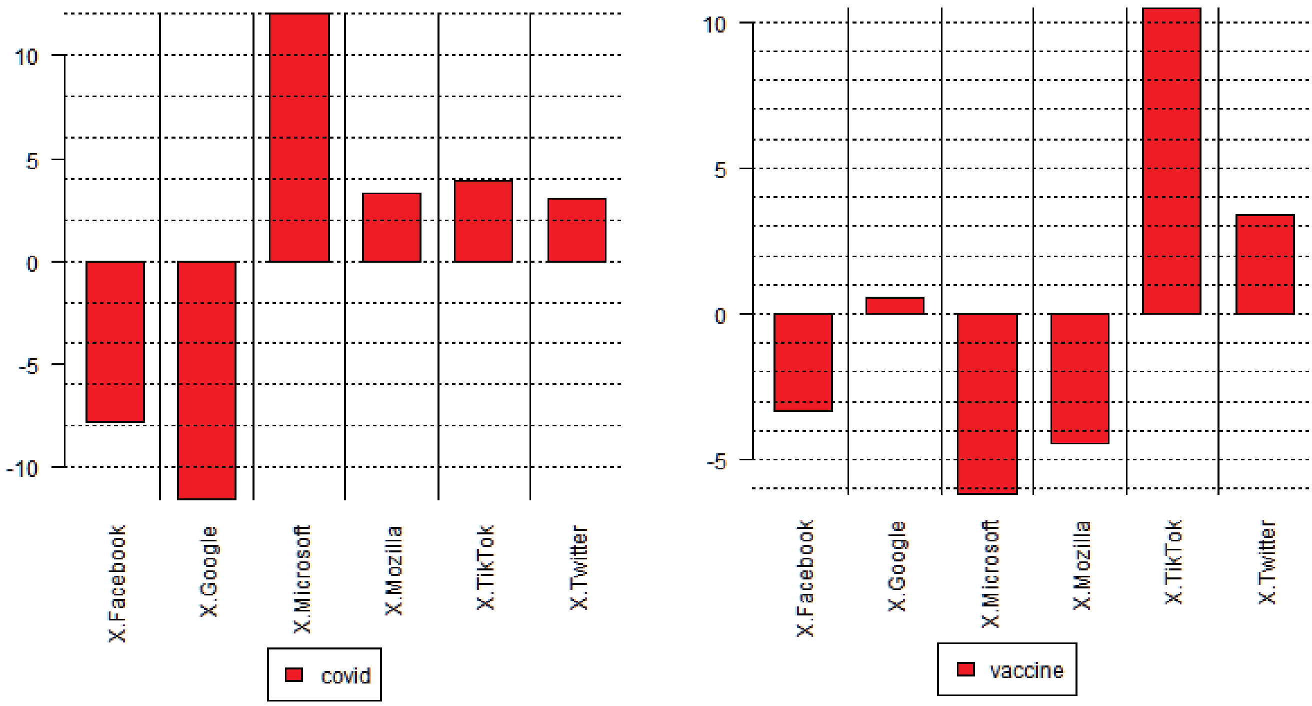

Figure 4: Frequency comparison of word clusters

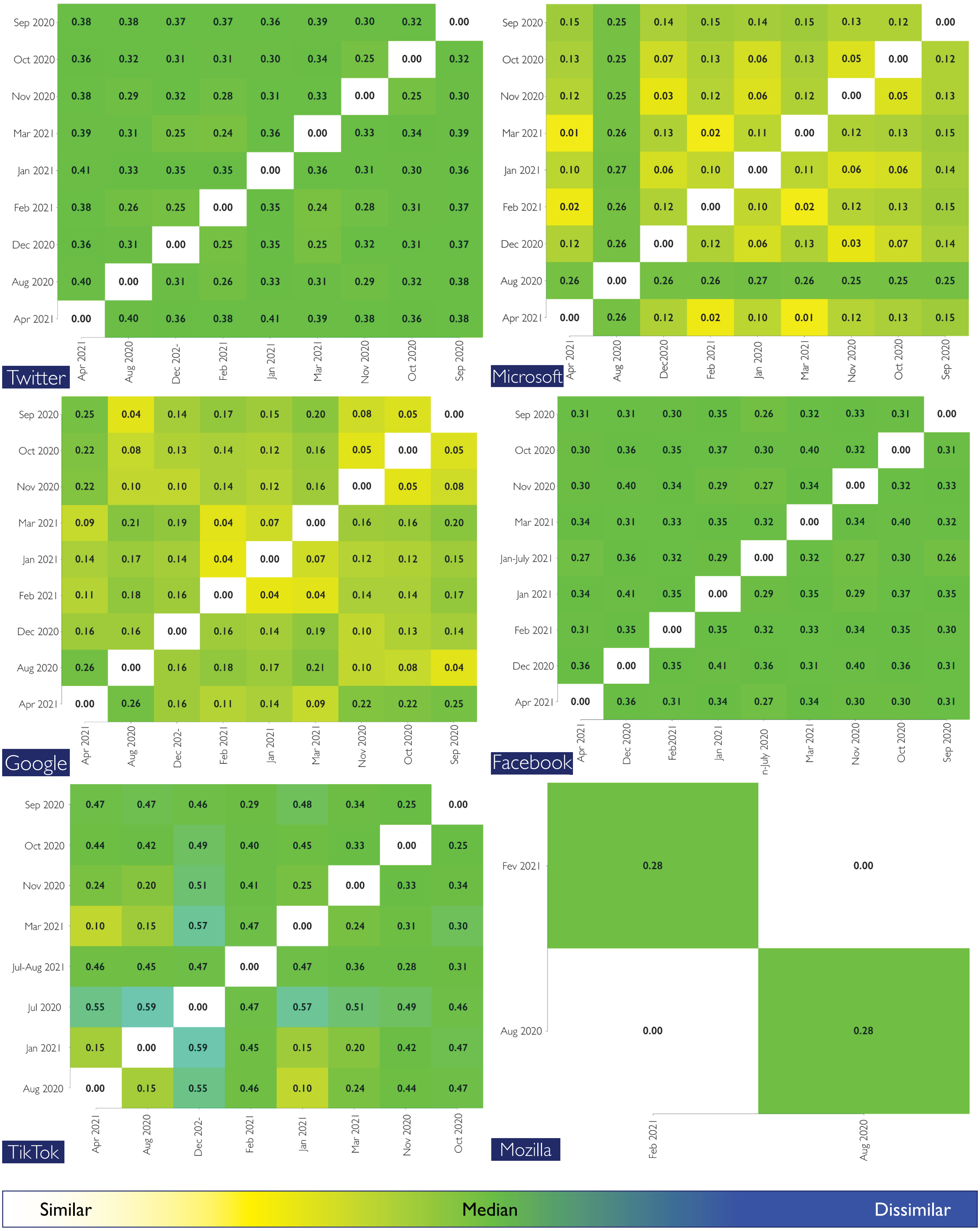

Figure 5: Labbé distance analysis to estimate similarity[22]

22. Note: the order of the reports within the figure are random rather than chronological.

Case Study: Facebook

In the absence of country-level data to verify reported actions, DCU FuJo cooperated with the Institute for Strategic Dialogue (ISD) to undertake a case study of Irish Facebook Groups and Pages known to propagate COVID-19 vaccine misinformation. Importantly, this analysis was only possible because Facebook provides access to data and analytics through its CrowdTangle tool. Facebook Groups and Pages differ from personal profiles. Pages act as public profiles connecting individuals, organisations, or brands with followers. Groups allow users sharing the same interests to communicate. Groups, which have their own administrators and moderators, can customise privacy settings to determine who can join a group and who can see what is posted by the Group[23]. Users become 'members' of a Group whereas they 'like' or 'follow' a Page.

The original analysis related to 49 Irish Facebook Pages and 37 Irish Facebook Groups. The number of Groups and Pages varied somewhat over the four-month analysis period based on the results generated from CrowdTangle. In addition, as Facebook implemented its policies, the visibility of some Pages, Groups, and posts changed over time. Table 17 provides an overview of the number of posts that were reviewed within the Groups and Pages during each sampled week.

Fact Checking

During COVID-19, Facebook expanded its factchecking programme and offered grants and training to support factcheckers. European factcheckers consulted for this research reported that, in their opinion, Facebook had become more proactive in its approach to factchecking. Grants, training, and regular meetings made the platform more accessible to factcheckers and factcheckers gained more insights regarding the utility and efficacy of their work. However, the factcheckers argued that detailed data, rather than rough percentages, about users' interaction with factchecks is necessary to evaluate and improve the scope of their work.

The implementation of Facebook's factchecking policies was examined in reference to the sampled weeks. Of the 219 posts about COVID-19 vaccines in the analysed Groups, eight were factchecked (see Table 18). However, this does not mean the remaining posts merited a factcheck. Following a review of these posts, we identified 30 additional posts with misleading claims for which there were factchecks available from Facebook's factchecking partners but they were not applied. Of the 127 posts about COVID-19 vaccines posted to the analsed Pages, we identified five misleading posts for which factchecks were available but not applied (see Table 19). Overall, the Groups posted more content that required a factcheck than the Pages. This may be partly due to the nature of Groups, which allow all members, and sometimes non-members, to post content. In contrast to Pages, there is less oversight regarding the content posted, especially if Group admins do not invoke post-approval tools[24]. However, it is also notable that the number of posts about COVID-19 vaccines declined sharply in the Groups after January, which may be indicative of strengthened enforcement of content policies.

23. https://www.facebook.com/help/337881706729661
24. https://www.facebook.com/help/1686671141596230

Table 17: Overview of the sampled data

Sampled Week	Group Posts	Pages Posts
1-8 January 2021	138	33
22-28 February 2021	28	28
22-28 March 2021	27	29
15-21 April 2021	26	37

Table 18: Factchecking of posts in the Groups

	# Posts about COVID-19 vaccines	# Factchecked posts	# Posts for which factchecks were available
1-8 January	138	8	19
22-28 February	28	0	5
22-28 March	27	0	4
15-21 April	26	0	2
Total	**219**	**8**	**30**

Table 19: Factchecking of posts in the Pages

	# Posts about COVID-19 vaccines	# Factchecked posts	# Posts for which factchecks were available
1-8 January	33	0	3
22-28 February	28	0	1
22-28 March	29	0	0
15-21 April	37	0	1
Total	**127**	**0**	**5**

Facebook asserts that it detects duplicates of debunked claims through similarity detection methods. However, similarity detection appears to be limited in its capacity to address duplicate claims in different formats. For example, a video created by a high-profile figure within the anti-lockdown movement presented false claims about people dying as a result of the COVID-19 vaccines; these claims were factchecked by Facebook factchecking partners at the time of circulation. The video was posted multiple times across various Groups and Pages in different formats. As shown in Figure 6, a factcheck warning was applied when a user shared the video by posting a URL from the video-sharing platform BitChute. However, no factcheck was applied when a user posted the false claim as text and shared a link to a webpage hosting the video. Examples like this highlight the difficulty of applying factchecks across formats and the inconsistencies that exist from the perspective of users and those seeking to spread disinformation.

Additionally, there were numerous instances whereby disinformation content was labelled with a factcheck when it was shared as a post, but not when it was shared in the comments section. In Figure 7, for example, a disinformation video urging people to "not get the COVID vaccine" was debunked as false information when it was shared as a post, but the same disinformation video was shared by a user in the comment section and did not carry any factcheck warning. Indeed, some of the most egregious examples of COVID-19 and vaccine disinformation appeared in the comments section accompanying posts.

Another inconsistency concerned the factchecking of political content. Under Facebook's Programme Policy, content created by politicians or their campaigns is not eligible to be reviewed and factchecked. During COVID-19 some political figures were leading proponents of disinformation. In Ireland, the prominent anti-lockdown campaigner depicted in Figure 6 ran for election and thereby became ineligible for factchecking during the campaign period. The candidate's campaign posts on Facebook used anti-vaccine disinformation including the false claim that: "over 1.7 million people have suffered adverse events and over 16,500 are reported dead in the EU". Factchecks were applied retrospectively after the candidate failed to win the election. However, as shown in Figure 8, these fackchecks were not applied in all circumstances; the false claims appeared in video format without a factcheck.

Figure 6: Links to a video with false claims about vaccines: a factchecked link shared via BitChute (left), and a weblink to the video without a factcheck (right).

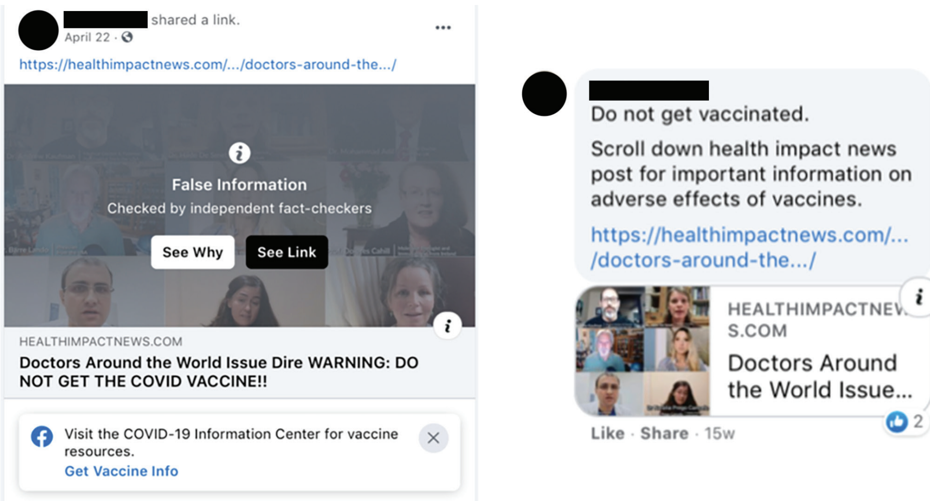

Figure 7: A disinformation video posted by a Group admin carries a factcheck label (left) and the same video posted in the comments section does not carry a factcheck (right).

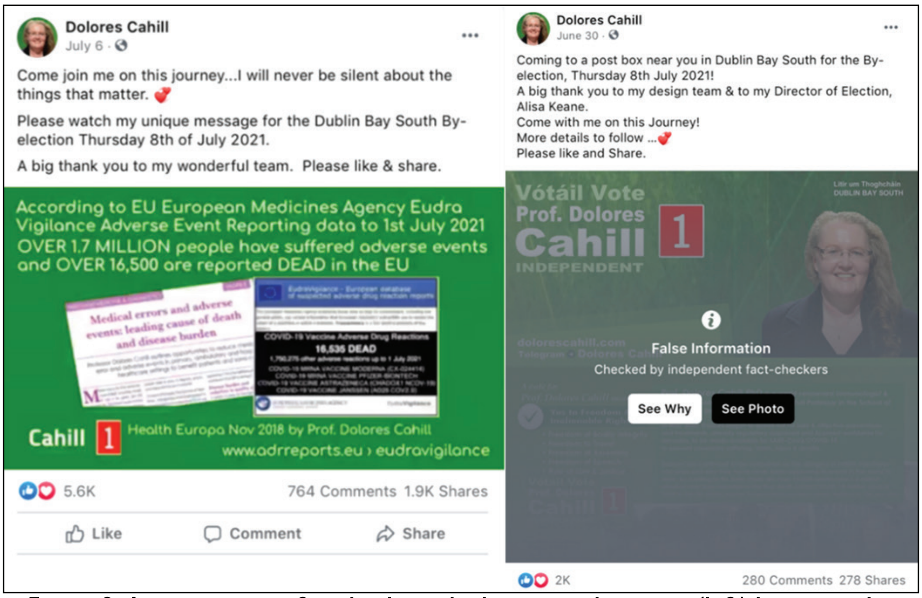

Figure 8: A retrospective factcheck applied to textual content (left) but not video content depicting similar false claims (right).

Labelling

Similar posts about COVID-19 vaccines carried an information label in some instances, but not in others. Most of these cases concerned links to external sources. For example, Figure 9 shows two versions of a COVID-19 story from the same source. The stories were posted to the same page one day apart, but only one carried a label pointing users to the COVID-19 Information Centre.

In March 2021, Facebook announced it would label posts about COVID-19 vaccines to "point people to the COVID-19 Information Centre" and "show additional information from the World Health Organisation". An analysis of March and April content indicates that this was applied inconsistently. Of the 53 posts about COVID-19 vaccines in the Groups only 39 were labelled while 56 of the 66 posts about COVID-19 vaccines in the Pages were labelled (see Table 20). It is also worth noting that users can choose not to view these labels by clicking the 'x' function in the top right-hand corner. After a page has been reloaded, the posts will appear without the labels.

Informing Users

In March 2021, Facebook announced that it would "start to let people know when they're about to join a Group that has Community Standards violations", reduce notifications for these Groups "so people are less likely to join", and reduce the distribution of the Group's content. In a sample of 22 Groups, eighteen carried the notification warning for new members. In April 2020, Facebook announced that it would apply an "educational pop-up" to direct "members of COVID-19 related Groups" to credible information from health organisations. Of the 22 Groups analysed, only three carried these educational pop-up labels.

Although the policy announcements explicitly mentioned Groups, they appeared to be applied to Pages in some instances. In a sample of 18 Pages known to disseminate COVID-19 vaccine disinformation, six notified users that they were about to like a Page that had previously violated Facebook's Community Standards. Just five of these Pages carried the educational pop-up directing users to credible information.

Facebook also announced that it would "limit invite notifications" for Groups that have Community Standards violations. An analysis conducted in May 2021 found that some Groups welcomed no new members over the course of the month. However, one Group welcomed approximately 1,723 new members. This Group did not display an educational pop-up nor did it notify users about the Group's violation of Community Standards. The Group posted frequently about COVID-19 and COVID-19 vaccines and hosted content that was rated false by factcheckers. Moreover the Group's admins promoted information about anti-lockdown protests.

Content Removals

In December 2020, Facebook announced that it would start removing false claims about Covid-19 vaccines that had been debunked. This included false claims about the safety, efficacy, ingredients, or development of vaccines and vaccine conspiracy theories. An expanded list of claims was included in February 2021 after consultations with leading global health organisations. COVID-19 content that increases "the likelihood of exposure to or transmission of the virus" or contributes to "adverse effects on the public health system" was also eligible for removal. Some posts that were judged to merit removal on these grounds were labelled rather than removed (see Figure 10).

Table 20: Appropriate labelling of posts about COVID-19 vaccines in the Groups and Pages[25]

	Groups		Pages	
	# Posts requiring a label	# Posts that were labelled	# Posts requiring a label	# Posts that were labelled
22-28 March	27	23	29	20
15-21 April	26	16	37	36
Total	**53**	**39**	**66**	**56**

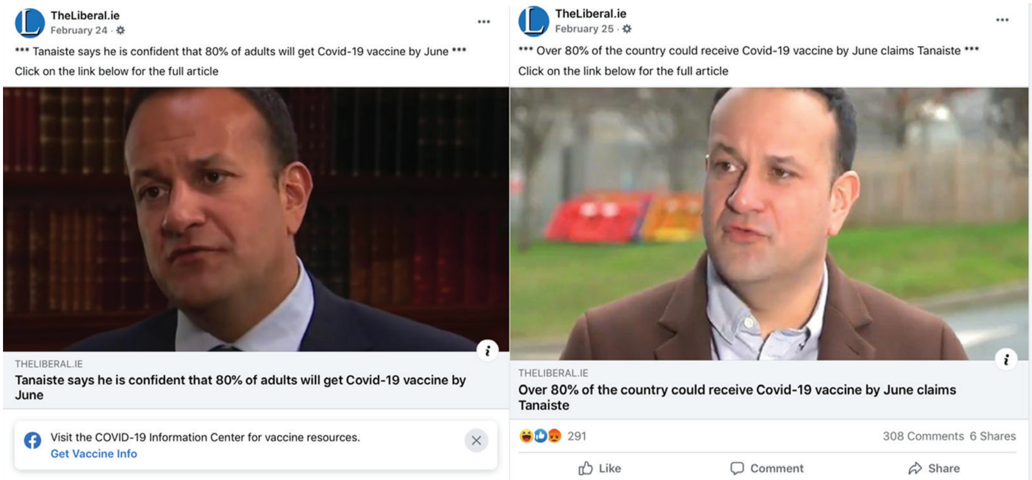

Figure 9: Two versions of the same story with only one labelled

Figure 10: Posts presenting conspiracy theory claims about COVID-19 vaccines (left) and claims that COVID-19 vaccines kill or seriously harm people (right)

25. As the availability of posts was subject to change, these figures represent the live posts that were observable at the time of analysis.

As Facebook does not provide specific definitions for 'content' and 'claims', there is a lack of clarity about whether content removal policies apply to comments. A recent research report[26] indicated that violating comments should be removed under Facebook's Community Standards. This analysis evaluated both posts and comments. Of the 219 posts about COVID-19 vaccines in the Groups, 34 were deemed to merit removal (see Table 21). An additional 41 posts hosted violating comments. Of the 127 posts about COVID-19 vaccines in the Pages, six were deemed to merit removal and an additional 52 posts hosted violating comments (see Table 22). For clarity, when the content in a post and the accompanying comments were both found to be in violation, this is reported as one instance. Similarly, when there are multiple violating comments accompanying a post, these are grouped as one instance as they appeared under a single post. For example, in the sampled January week, there were 17 posts in groups hosting violating content in the comments and this is in addition to the 24 posts that were deemed to merit removal.

Notably, Groups were more likely to host posts that merited removal than Pages and violating content appeared far more regularly in the comments than in posts. See Figure 11 for an example of violating comments. Importantly, violating comments often appeared under posts that did not violate content policies such as factual news articles. A recent analysis by First Draft highlighted this issue in relation to comments accompanying posts on the Pages of two Australian news organisations[27]. These findings indicate the pressing need for a framework to address disinformation in comments.

Admins and moderation

In March 2021, Facebook allocated a greater moderation role to the admins and moderators of Groups whose members had violated content policies. In these cases, admins were asked to temporarily approve all posts. However, if admins are the organisers of anti-lockdown and anti-vaccine Groups, they appear unlikely to enforce these measures against disinformation.

Among the analysed Groups known to propagate COVID-19 and vaccine disinformation, we found that admins were often members of multiple Groups. One admin was a member of at least thirteen Groups. We also found evidence of admins operating multiple Facebook accounts to circumvent sanctions including account blocking and suspensions.

Subsequent to the analysis period, Facebook announced in June that it was "testing and rolling out" new tools for admins. Notably, this includes comments moderation; admins can now proactively prevent certain keywords from appearing in the comments section; they can reduce promotional content by declining posts or comments with specific weblinks and they can limit the number of comments allowed under posts. Additionally, a tool known as 'conflict areas' is being trialled whereby Facebook will notify Group admins of "contentious or unhealthy conversations taking place in their Group so they can take action as needed"[28]. While these measures are positive developments, the core question about the intentions of admins who have organised anti-lockdown and anti-vaccine Groups remains.

Disinformation networks

While this analysis concentrated on Facebook it is important to note that the content on the Facebook Groups and Pages was interlinked with other platforms that are not signatories to the Code. Links to third-party websites regularly redirected users to platforms, such as BitChute, which hosted vast amounts of COVID-19 and vaccine conspiracies. Relatedly, Facebook users encouraged Group members and Page followers to migrate to alternative platforms - including Telegram, Gab and MeWe - that apply minimal content moderation. Positioned within a network of platforms, Facebook was discussed as a strategic means to gain followers. As one user stated: "Facebook should be used for gaining numbers but conversations should be elsewhere." Finding ways to address the role of Facebook, and other Code signatories, within the wider network of platforms would seem to be a challenge for the future of the Code.

26/27. https://firstdraftnews.org/articles/vaccine-misinformation-in-facebook-comment-sections-a-case-study/

28. https://www.facebook.com/community/whats-new/new-tools-features-nurture-community/

Table 21: Posts and comments judged to merit removal in the Groups.

	# Posts about COVID-19 vaccines	# Posts deemed to warrant removal	# Comments deemed to warrant removal
1-8 January	138	24	17
22-28 February	28	1	11
22-28 March	27	5	3
15-21 April	26	4	10
Total	**219**	**34**	**41**

Table 22: Posts and comments judged to merit removal in the Pages

	# Posts about COVID-19 vaccines	# Posts deemed to warrant removal	# Comments deemed to warrant removal
1-8 January	33	1	12
22-28 February	28	0	16
22-28 March	29	2	9
15-21 April	37	3	15
Total	**127**	**6**	**52**

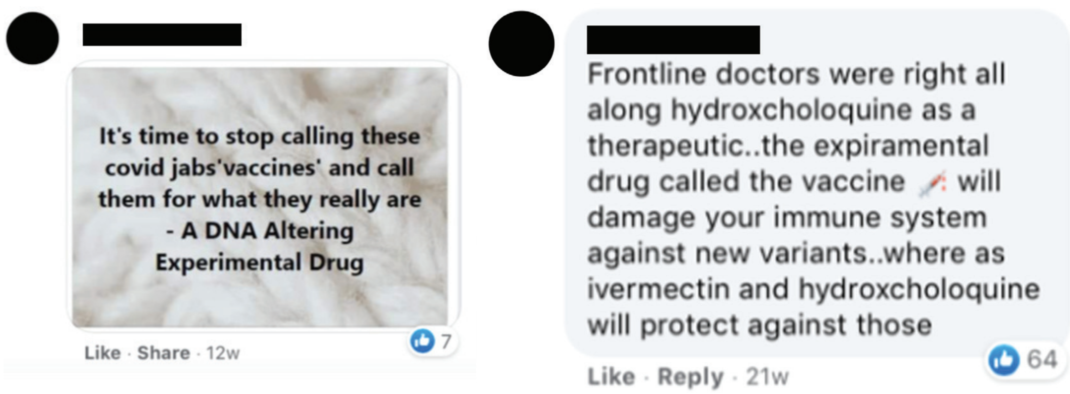

Figure 11: Comments presenting false claims about the safety of COVID-19 vaccines

Case Study: TikTok

A micro-analysis of TikTok verified claims about the promotion of authoritative content and the application of content labels.

Promoting authoritative content

A quarter of all the signatories' actions and 43 percent of TikTok's actions were about the promotion of authoritative content. In its baseline report from August 2020, TiKToK stated that when "users search for coronavirus-related topics" they will "find videos from verified accounts that are providing trusted information from credible sources." For example, as illustrated in Figure 12, the first results shown to a user in response to a COVID-19-related search should present reliable sources as the WHO and British Red Cross.

However, a series of searches conducted across six accounts in July 2021 found that the promotion of authoritative content was inconsistent. Search terms related to COVID-19 (e.g. covid) generally, but not always, return videos from authoritative sources such as the WHO. However, the two vaccine search terms - vaccination and vaccine - did not return any videos from recognised authoritative sources. In the vast majority of cases, these videos were about COVID-19 vaccines. In one instance, one of the top results (i.e. visible without scrolling) returned a video, which claimed a child was harmed by "the experimental Moderna jab". It encouraged people to share the video to raise awareness of the damage "caused to human beings" by vaccines (see Figure 13). Although the video did carry a label to "learn more about COVID-19 vaccines", its presence at the top of the search results is contrary to the promotion of authoritative content.

It appears that the promotion of authoritative content only applied to vaccine search results when a COVID-19 specific term was also included. For example, all searches for "Covid vaccine" returned a prominent banner about COVID-19 vaccines (see Figure 14), but none of the searches for "vaccine" or "vaccination" resulted in the same banner, despite the fact that the search results were clearly related to COVID-19. For example, the top search results for "vaccine" in Figure 15 did not display a banner and returned one video applauding healthcare workers for refusing to take a COVID-19 vaccine.

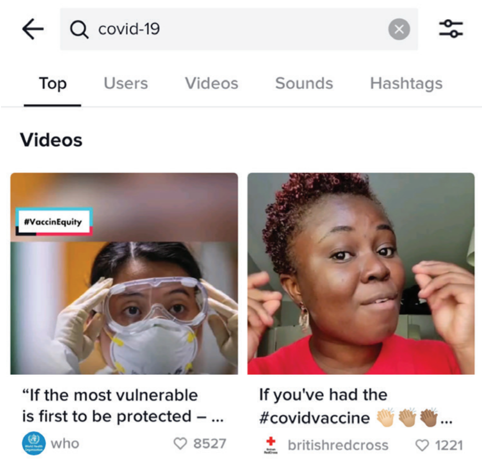

Figure 12: Promotion of authoritative content

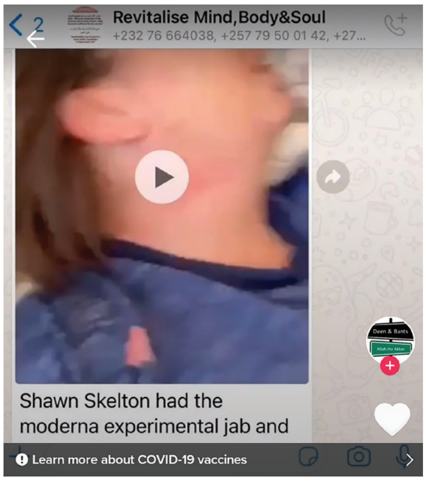

Figure 13: Top search result for "vaccine"

Figure 14: Search term "Covid Vaccine"

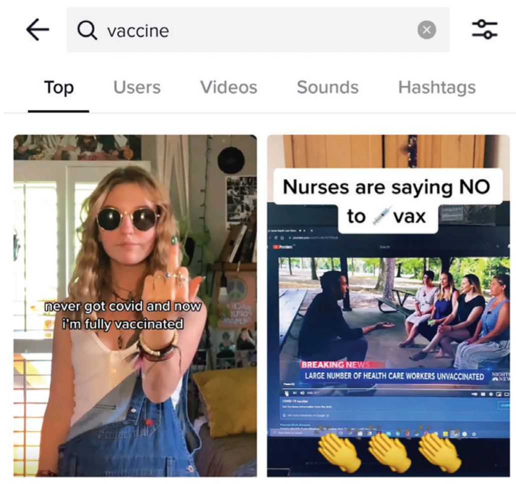

Figure 15: Search term "vaccine"

Content labelling

TikTok also reported that labels would be applied to COVID-19 content: "Across all markets, we detect and tag all videos which have words, hashtags or music related to COVID-19 and we attach a 'sticker' to those videos with a message 'Learn the facts about COVID-19". TikTok later reported that a similar label would apply to "all videos with words and hashtags related to the COVID-19 vaccine" beginning in December 2020. However, the labels were applied inconsistently. In addition to using search terms, we also viewed the top 20 TikTok posts available under the hashtags #covid, #vaccine and #vaxx. These posts represent the 'most engaged with' content for those hashtags and, as such, give an indication as to whether COVID-19 and COVID-19 vaccination content was being accurately detected and tagged.

As shown in Table 23, only some of the top 20 videos about COVID-19 were appropriately labelled. Among the top 20 videos for #covid, only two were tagged with a label. Sixteen of the top 20 videos for #vaccine were about COVID-19; only nine of these were labelled. Seventeen of the top 20 videos for #vaxx were about COVID-19; only two of these were labelled.

While the #vaxx hashtag has the smallest number of overall views on TikTok - eight million in contrast to the 23 billion for #covid - it is a keyword associated with the anti-vaccine movement and anti-vaccine disinformation. Thirteen of the top 20 results for #vaxx were clearly speaking against COVID-19 vaccines, yet only four of these were labelled. One of these results is about accessing forms "to legally refuse this experimental vaccine" (see Figure 16). It has been shared 11,900 times.

Anti-vaccine and COVID-19 disinformation is easily accessible and searchable on TikTok and user comments reinforce conspiracy theories (see Figure 17). Generic searches return conspiracy theories such as those illustrated in Figure 10 and user comments make it easy to find others who share and create similar content. The promotion of authoritative content and the application of labels is inconsistent. If AI systems are being used to detect content in English, then it appears that those systems fail to capture relevant content. While content labels were present on some instances of disinformation, the generic nature of the label does little to challenge the presence of disinformation. This underscores the need for meaningful metrics to assess the effectiveness of signatories' actions to counter disinformation.

Table 23. Results of hashtag searches

	Total number of hashtag views	# COVID-19 related videos in top 20	# Videos tagged with label in top 20
#covid	23.6b	20	2
#vaccine	4.2b	16	9
#vaxx	8.3m	17	7

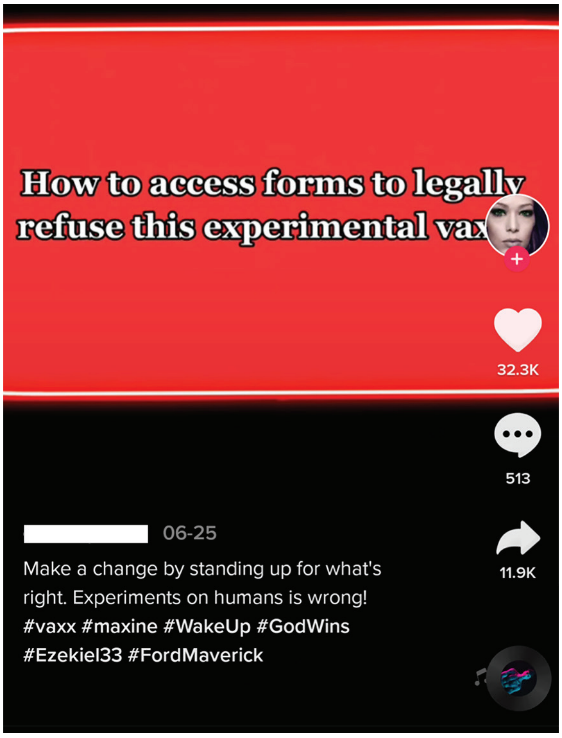

Figure 16. A top 20 #vaxx result

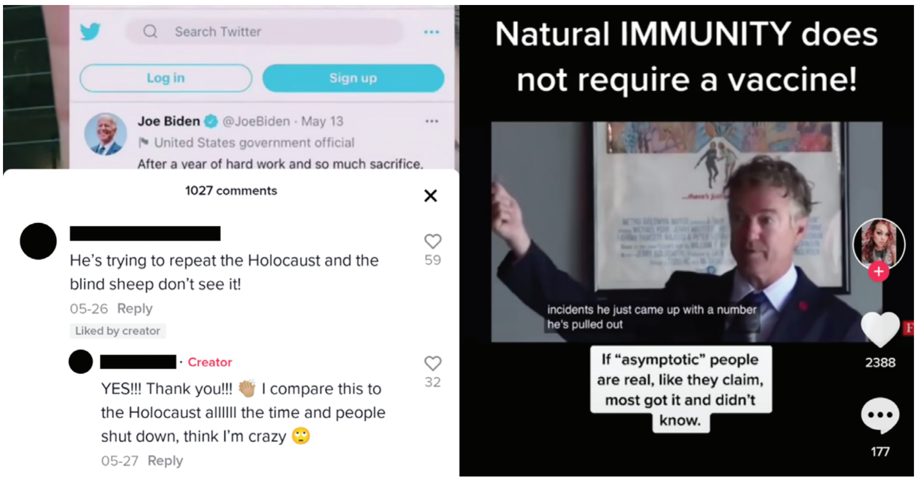

Figure 17. Anti-vaccine claims and conspiracy theories in posts and comments

Case study: Automation and AI

Neither the Code of Practice of Disinformation nor the European Commissions' June 2020 communication on COVID-19 disinformation address the use of Artificial Intelligence (AI). However, the Commissions' May 2021 ***Guidance on Strengthening the Code of Practice on Disinformation***[29] urges platforms to harmonise their content moderation practices with the relevant provisions of the EU's proposed AI Act[30]. The manual coding found that 25 percent of all actions involved the promotion of authoritative content; 13 percent related to blocking, removing, or demoting content; and nine percent related to factchecking and labelling content. Automation plays a key role in these three areas as AI systems are used to automatically detect content for labelling and search queries that merit authoritative results. With the exception of Mozilla, each signatory reported actions in the three areas mentioned. In this context, the signatories' reports were reviewed to identify any references about the use of AI or automation in their efforts to tackle COVID-19 disinformation.

Signatories did not provide a consistent account regarding the use of AI, automation, or machine learning. The percentage of actions (e.g. content removal) that were taken as the result of automated versus human moderation remains unclear. Notably, Twitter did provide data for actions arising from automated solutions, but only in relation to advertising and, more specifically, concerning content that contained words related to COVID-19. The lack of data regarding the use of AI in content moderation is important to note given the implications of automated decision-making for freedom of expression. The signatories, with very few exceptions, did not provide adequate explanations of how their AI systems function to identify problematic content or bad actors. Instead, they simply refer to such systems in a rather vague way.

29. https://digital-strategy.ec.europa.eu/en/library/guidance-strengthening-code-practice-disinformation

30. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

31. https://www.buzzfeednews.com/article/ryanmac/facebook-employees-bias-arabs-muslims-palestine.

32. https://santaclaraprinciples.org/

The signatories who mentioned the use of AI in their content moderation systems primarily referenced efforts to identify and address manipulative behaviour such as Coordinated Inauthentic Behaviour (CIB) networks, fraudulent commercial content (e.g. Facebook's Marketplace), and synthetic images/videos or deepfakes. This is rather positive given the fact that the strengthened Code refers to commitments against such content. What remains somewhat vague is the grey area of content which is characterised as harmful but lawful. It seems that signatories primarily reported actions on content that explicitly violated their policies; yet, there is empirical evidence showing repeated false positives and biased enforcement against marginalised groups[31].

Finally, we know that the dominant language of the AI systems used by the signatories is English. As Ireland is the only English-speaking country in the post-Brexit EU, further information is required regarding the adequacy of these systems for detecting and addressing disinformation in languages beyond English or, even, other global languages such as Spanish and Portuguese.

Many platforms have signed up to the Santa Clara Principles on Transparency and Accountability in Content Moderation[32]. These principles ask platforms to reveal the numbers of actions taken on content according to the type of detection (e.g. government-mandated or automatically detected) and to explain how their automated systems work. Some of this information is provided by the signatories in their own transparency reports, but it is not provided in the transparency reports submitted as part of the EU Code of Practice.

Conclusions and Recommendations

While the Code has proven a useful instrument insofar as it has prompted signatories to respond to concerns about disinformation, there are evident shortcomings in its implementation and scope. It remains difficult to assess the timeliness, completeness and impact of the actions undertaken by signatories. Here we offer our recommendations for more effective reporting and monitoring.

Recommendation 1: standardise the reporting format
In the absence of a standardised reporting format, each signatory reports in an idiosyncratic manner. Although some variance is to be expected given the differences in the services offered by the signatories, the current reporting mechanism is not conducive to an assessment of signatories' actions. Considerable effort is required to extract a clear picture of what actions were undertaken, how those actions relate to the signatory's policies, whether those actions were new, and whether they were relevant to the Code and EU Member States. In particular, the free-text nature of the reports affords signatories the opportunity to produce repetitive and irrelevant information. In some cases, the reports were excessively long due to considerable amounts of repetition and lengthy descriptions of irrelevant actions. Notably, some signatories copied company press releases and self-promotional announcements without any effort to tailor the information for the Code or the EU.

Regarding the transparency reporting envisaged in the Digital Services Act (DSA), Dot Europe, the lobby group representing tech companies, cautions that: "there is no 'magic button' which will enable a service provider of any size to automatically pull and deliver the data in the right structure and format of a transparency report"[33]. Nevertheless, a certain level of standardisation is necessary for effective reporting and monitoring.

We recommend that reporting be standardised, as far as possible, to ensure necessary and relevant information is provided and in a manner that facilitates monitoring. For each action, we recommend that signatories clearly state:

- the specific policy, if any, associated with the action
- the relevance of the action to the Code or the specific information requested
- whether the reported action is a new initiative or part of an ongoing initiative
- the regional application of the action and, in particular, the application across EU Member States
- whether metrics or other outcome data are available at the level of EU Member States.

Recommendation 2: create a reference resource to clarify policies and definitions
It is often difficult to discern how the reported actions relate to a signatory's policies, whether those policies relate to disinformation per se, and whether those policies apply across all EU Member-states. In addition, generic terms, such as "content' and "label", are often used by the signatories without specific details. It would be helpful to have a clear definition of what is meant by these terms and what is included and excluded. For example, many signatories apply content labels, but there are important differences between the application of generic labels (e.g. read the facts about COVID-19 vaccines) and the application of specific labels to pieces of content (e.g. this claim has been rated false). Furthermore, more information is required about the application of labels such as their placement and whether they can be turned off by users. Similarly, when a signatory's service facilitates multiple content types (e.g. posts, comments), the term 'content' needs to be clearly defined to indicate what is included and what is excluded from a policy or action.

33. https://doteurope.eu/library/dot-europe-questions-and-recommendations-on-dsa/

We recommend that signatories provide clear definitions of relevant policies to combat disinformation, clear definitions of common terms, and how those terms are operationalised on their services. This information should be available as part of a reference resource.

Recommendation 3: introduce a framework to address disinformation in comments
User comments have been overlooked as a source of disinformation. The Code asked signatories to address content "in search, feeds, or other automatically ranked distribution channels." We note that signatories' policies typically prioritise posts rather than the comments that accompany them. Moreover, factckechers concentrate their limited resources on posts as the potential reach of comments is much lower than the potential reach of posts. However, as our and related case studies[34] indicate, user comments facilitate the fermentation of harmful disinformation. The inattention to comments creates notable contradictions whereby disinformation claims that appear in posts are labelled or removed, but the same claims appearing in comments are not. Relatedly, there is an obvious contradiction in the expectation that posters who are already known to propagate disinformation will take responsibility for the veracity of the comments accompanying their posts.

We recommend that signatories and relevant stakeholders introduce a framework to address disinformation in comments that is consistent with Article 10 of the European Convention on Human Rights and the principle of freedom of opinion. Provided that comments are indeed automatically ranked, it would be reasonable for the Commission to include them in the revised Code as points where disinformation can be spread. Although comments are already being filtered using automated solutions, albeit with questionable results, we recommend taking a more nuanced approach, one that would not endanger freedom of expression: empowering, on the one hand, platforms' communities to better moderate comments and, on the other hand, implementing mechanisms that could also have educational value. A mechanism to address hateful and inappropriate comments has been introduced by some signatories[35/36] whereby users are prompted to reevaluate their comment based on its potentially harmful nature or, alternatively, such comments are hidden or filtered. In instances where a claim in a post has been reviewed and labelled by a factchecking partner, we also recommend that the same claims be actioned when they appear as user comments.

Recommendation 4: define parameters for granular data
In many instances, the reported data was aggregated at the global level. When EU data was provided, it was frequently presented in aggregate, which is of little use when assessing implementation across EU Member States. Moreover, the current reporting format allows signatories to report broad data that may or may not be relevant. For example, some signatories report metrics about hate speech violations without clarifying how many, if any, of those violations were related to COVID-19 or COVID-19 disinformation. Relatedly, broad action areas such as factchecking require more granular data including how many Member States have factchecking partners, and how many COVID-19 posts have been factchecked, removed, or otherwise actioned at the Member State level.

Expanding on Article 23(2) of the Digital Services Act, which requires more detailed data regarding platforms' active users, we recommend that clear parameters be defined for the reporting of granular data about specific action areas and in relation to EU Member States. A more comprehensive picture of the signatories' actions on specific types of content related to disinformation is necessary to evaluate the effectiveness of the Code and signatories' actions.

34. https://firstdraftnews.org/articles/vaccine-misinformation-in-facebook-comment-sections-a-case-study/
35. https://blog.twitter.com/en_us/topics/product/2021/tweeting-with-consideration
36. https://support.tiktok.com/en/using-tiktok/messaging-and-notifications/comments

Recommendation 5: define meaningful KPIs and demonstrate effectiveness
When signatories do provide data it tends to be in the form of engagement metrics. However, these metrics rarely provide much insight into the effectiveness of an action. There are two dimensions to this problem: First, the reported engagement metrics are often too broad. For example, when reporting on a media literacy campaign, signatories typically report the number of people who clicked on the campaign link. As it seems likely that many people will click on a link by mistake and decline to engage with the content, more meaningful metrics would provide information about how many people remained on the campaign page and the time-spent on the page relative to the volume of content. Second, engagement metrics reveal little about the effectiveness of the action in terms of combating disinformation. For example, to understand the effectiveness of actions such as media literacy campaigns it would be instructive to know whether those who engaged with a campaign were less likely to engage with disinformation as a result.

We recommend that meaningful KPIs be defined for the reporting of results and outcomes in relation to key areas including: content labels, content and account removals, factchecking, and media literacy campaigns. Moreover, we recommend that signatories report on their own efforts to measure the efficacy of these actions and provide data to independent researchers to verify that efficacy.

Recommendation 6: fund and appoint an independent auditor
In the 2018 Code, signatories committed "to select an objective 3rd party organisation" to review the self-assessment reports and to evaluate progress, "which would include accounting for commitments" signatories have agreed to under the Code. However, this commitment was not implemented. The lack of a well-resourced and independent auditor to review the quality of what has been reported is a significant weakness. Current and proposed oversight structures - ERGA and the European Digital Media Observatory (EDMO) - appear inadequate to address this gap and the needs of ongoing monitoring. For example, we note that this report required considerable resources in terms of funding, personnel, and time and it relied on the research infrastructure of an academic institute.

We recommend that the original commitment to an independent auditor be implemented under the revised Code. Further, we recommend that signatories provide adequate funding and resources to support this position, which will contribute to the monitoring work of ERGA and EDMO.

Recommendation 7: define procedures to independently verify actions
As our case studies indicated, the reported actions were not applied consistently. The reporting of engagement metrics - when they are made available - does not address this issue. For example, providing country-level metrics about content removals or engagement with COVID-19 information centres does not provide any indication about whether those actions were applied consistently and appropriately. Consequently, there is a major gap in the monitoring of the Code as it assesses what actions the signatories have reported without the certainty that those actions have actually been implemented across EU Member States and are working as stated.

We recommend that standardised procedures to verify the implementation of actions be agreed for future monitoring. This will ensure consistency in monitoring and provide an important counterpoint to the signatories' reported metrics.

Recommendation 8: require the provision of data on automated systems
The reports provided only limited insight into the use of AI and automated systems to combat disinformation. Yet, many of the reported actions, such as the application of generic content labels, presumably rely on automated systems. In some instances, signatories may be using AI to identify harmful content such as harassment or hate speech. In other instances, they may be using automated tools to

match policy-violating content against a blacklist of known instances. Consequently, it is important to discern how signatories frame AI and automated tools and how they are deployed against disinformation. Relatedly, it is important to understand how datasets are used to train content moderation models. As AI solutions are typically not equally advanced in all the languages of the EU, there may be regional gaps in the application of actions. Moreover, given major concerns about the transparency of automated content moderation and the potential implications for freedom of expression, further details are desirable.

We recommend that signatories report on their use of automated systems to combat disinformation including an explanation of what systems are used, what languages are covered, what kinds of disinformation they are trained to detect, and what risk assessments have been conducted on the AI systems used to tackle disinformation. Additionally, we propose that the European Commission specifically articulates the need for risk assessments related to disinformation in the strengthened Code.

Recommendation 9: render access to data for research on disinformation binding

The European Commission has acknowledged the need for greater access to data that will allow independent researchers to better understand the role platforms play in various domains including disinformation. The Digital Services Act introduces obligations for platforms to make data available to "vetted researchers" (Art. 31) and this provision is echoed in the Guidance on Strengthening the Code (8.1). Some social media platforms have made access to their data available either through APIs (e.g. Twitter) or through curated services (e.g. Facebook's CrowdTangle). While the former certainly affords greater opportunities for analysis, the latter is also welcome for its ease of use and accessibility. However, platforms should not be left to decide the criteria for accessing data nor their format, as evidenced by Facebook's recent decision to block a team of New York University researchers studying disinformation[37] from accessing its services.

We recommend that signatories embrace the need for transparency and data-sharing with researchers, as well as expand and improve services that allow researchers to access data. Moreover, we suggest that the Commission create a clear regulatory framework for accessing data for research on disinformation and further expand the scope of its current proposal to include more stakeholders, including members of civil society organisations, rather than just university-affiliated researchers.

37. https://www.npr.org/2021/08/04/1024791053/facebook-boots-nyu-disinformation-researchers-off-its-platform-and-critics-cry-f?t=1628525534571