Mündges, Stephan; Park, Kirsty

Article

But did they really? Platforms' compliance with the code of practice on disinformation in review

Internet Policy Review

Provided in Cooperation with:
Alexander von Humboldt Institute for Internet and Society (HIIG), Berlin

Suggested Citation: Mündges, Stephan; Park, Kirsty (2024) : But did they really? Platforms' compliance with the code of practice on disinformation in review, Internet Policy Review, ISSN 2197-6775, Alexander von Humboldt Institute for Internet and Society, Berlin, Vol. 13, Iss. 3, pp. 1-21,
https://doi.org/10.14763/2024.3.1786

This Version is available at:
https://hdl.handle.net/10419/300750

INTERNET POLICY REVIEW
Journal on internet regulation

Volume 13 **Issue** 3

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

But did they really? Platforms' compliance with the Code of Practice on Disinformation in review

Stephan Mündges *TU Dortmund University*
Kirsty Park *European Digital Media Observatory*

DOI: https://doi.org/10.14763/2024.3.1786

Published: 25 July 2024
Received: 28 December 2023 **Accepted:** 27 March 2024

Funding: Kirsty Park received funding for this research from the European Union under action number 2020-EU-IA-0282 and agreement number INEA/CEF/ICT/A2020/2381686, Stephan Mündges received funding for this research from the European Union, project number 101083573.
Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: Mündges, S. & Park, K. (2024). But did they really? Platforms' compliance with the Code of Practice on Disinformation in review. *Internet Policy Review*, *13*(3). https://doi.org/10.14763/2024.3.1786

Keywords: Disinformation, Platform regulation, Code of Practice on Disinformation, EU internet policy, VLOP

Abstract: A key pillar in the EU's approach to regulating disinformation is the Strengthened Code of Practice on Disinformation. This self-regulatory Code proposes a broad range of measures for different stakeholders. It has been signed by platform companies that thereby agreed to report on compliance with the Code. This study investigates Google's, Meta's, Microsoft's, TikTok's and Twitter's (now X) compliance with their reporting obligations for the first time. Analysing the platform baseline reports published in early 2023, we find that, overall, platforms are only partly compliant with the Code. Qualitative information provided by platforms often lack detail and/or relevance. Reported quantitative data is, in several cases, missing, incomplete, or not robust. We point out claims by platforms that are doubtful or have been proven wrong in the past, and highlight avenues for future research and investigations. Additionally, we reflect on the framework in place for monitoring the Code of Practice and ways to improve it. This study is particularly relevant as the EU is transitioning from a self-regulatory to a co-regulatory model when regulating disinformation. The Code of Practice may soon become a code of conduct under the Digital Services Act (DSA) making non-compliance with it sanctionable and increasing the need for systematic monitoring.

Introduction

Dis- and misinformation is considered by many scholars, journalists, and policy-makers to be a substantial threat to democratic systems and public discourse (Lewandowsky et al., 2017; van der Linden, 2023). There is debate within academia regarding the scope of this challenge, with some scholars noting that empirical evidence for the prevalence of mis- and disinformation and how it influences public attitudes, behaviours, and participation in political processes is contentious (Altay et al., 2023; Altay et al., 2022; Jungherr & Schroeder, 2021). However, other findings indicate that false claims and misinformation are indeed quite prevalent, particularly, on digital platforms (Yang et al., 2023; Ecker et al., 2024). Independent of this academic debate, several political actors have taken action to counter mis- and disinformation, hinder its spread in digital information spaces and – in some instances – penalise actors publishing disinformation (Cipers et al., 2023). Among democratic political systems, the European Union is arguably the most prominent and most active in this regard. Apart from non-legislative actions (e.g. more strategic communication, more funding for media literacy programmes), it has initiated the Strengthened Code of Practice on Disinformation (hereinafter referred to as 'the Code' or 'the CoP'), a self-regulatory framework signed by the major technology companies (Google, Meta, Microsoft, TikTok), among others (CoP, 2022). By becoming signatories, platforms commit to take numerous very specific actions to counter mis- and disinformation on their services.

The Code is detailed and wide-ranging, covering fields such as political advertising, recommendation systems, and support for researchers and fact-checkers. Thus, the reporting that the signatories of the Code publish biannually is extensive. Even though only a voluntary commitment by the platforms, the Code is a regulatory instrument which may help to provide transparency about disinformation on large online platforms, but the extent of this transparency requires systematic monitoring. This study summarises data and findings of such a monitoring, that was carried out as a pan-European collaboration by independent academic researchers and has been reported to stakeholders (Park & Mündges, 2023). Assessments of signatories' compliance with certain parts of the Code have been conducted by other stakeholders (EFCSN, 2023; Kiely et al., 2023), but, to our knowledge, this is the first comprehensive monitoring of the Code's platform signatories. For this, we developed a monitoring framework which we applied to the first set of signatory reports which were published in early 2023. Our analysis provides insights into platforms' compliance with the Code but also offers valuable lessons for future monitoring of similar policy instruments and may be instructive for ongoing formal

proceedings against VLOPs. This is particularly relevant as the Code of Practice on Disinformation might become a Code of Conduct, and hence turn into a legal imperative, under the Digital Services Act (DSA).

The Code of Practice on Disinformation as a cornerstone of the EU's response to disinformation

Disinformation is a complex issue with varying definitions and the need for balancing different fundamental rights, making it challenging to regulate (Peukert, 2023). Stasi and Parcu (2021) offer three conceptual models that can be applied when regulating disinformation:

TABLE 1: Regulatory models, adapted from Stasi & Parcu, 2021

MODEL	DESCRIPTION	STATE INFLUENCE LEVEL	WEAKNESSES
Statutory model	Rules passed by lawmakers or regulatory bodies, enforced by the government, government agencies, or regulatory authorities	strong	Unlikely to keep up with fast changing technology landscape; could be used to justify censorship
Co-regulatory model	Rules which are negotiated by those subject to them and the regulator	medium	Risk of regulatory capture
Self-regulatory model	Rules relying entirely on voluntary compliance, where legislators and regulators mainly observe private efforts and essentially play no active role	weak	Shift of power from public to private sector; lack of accountability; lack of protection of fundamental rights

These models are, of course, not fully distinct from one another but should be thought of as a "spectrum of possibility" (Stasi & Parcu, 2021, p. 422). In the case of regulating disinformation, the EU started out with the self-regulatory model, adopting a multi-stakeholder approach, but is shifting its approach over recent years to the co-regulatory and in part statutory model.

The Code was initiated by the European Commission and published in 2018, marking the first time that platforms agreed on a set of self-regulatory standards (European Commission, 2018). The text of the Code was informed by the report of the High-Level Group on fake news and online disinformation and signed by several stakeholders including advertising associations and platform companies. After its release, the Code underwent monitoring by the European Commission itself, as well as by various stakeholders, all of which concluded similar points of critique: the Code was vague concerning certain commitments and insufficiently comprehensive in others. Moreover, it was noted that the Code lacked measurable objectives, key performance indicators, and a lack of independent verification of plat-

form claims (Culloty et al., 2021; European Regulators Group for Audiovisual Media Services, 2020; Stasi & Parcu, 2021). As a result, it was revised and re-published as "Strengthened Code of Practice on Disinformation" in 2022 (CoP, 2022). This new iteration incorporates numerous suggestions previously put forth by stakeholders and monitoring bodies, is better structured, more detailed, and clearer in what information must be provided. Additionally, signatories are obliged to report more granular and country-specific data.

The Strengthened CoP defines disinformation broadly to encompass "misinformation, disinformation, information influence operations, and foreign interference in the information space" (CoP, 2022, p. 1). It consists of nine sections, which can be divided into three categories: reaction, empowerment, and monitoring.

- Reaction: addresses monetisation of disinformation and manipulative techniques.
    - Sections: scrutiny of ad placements, political advertising, integrity of Services.
- Empowerment: mandates cooperation with stakeholders.
    - Sections: empowering users, research community, fact-checking community.
- Monitoring: details how the Code is overseen and its outcomes disseminated.
    - Sections: transparency centre, permanent task-force, monitoring of the Code.

Each section has three layers:

1. Commitments: overall objectives and actions
2. Measures: specific steps for Signatories
3. Reporting elements: either "Qualitative Reporting Elements" (QRE) or quantifiable "Service Level Indicators" (SLI); they detail what signatories have to report in order to prove compliance with measures

For example, commitment 27 obliges platforms "to provide vetted researchers with access to data necessary to undertake research on Disinformation by developing, funding, and cooperating with an independent, third-party body that can vet researchers and research proposals" which is operationalized in four measures: to contribute to the development of the aforementioned independent third-party body (measure 27.1); commit to co-fund this body (27.2): work with this third-party body "to enable sharing of personal data necessary to undertake research on Disinformation with vetted researchers" (27.3); engage in pilot programs for granting data access to researchers without the third-party body being set up (27.4). For

each measure signatories have to report qualitative information, specified in the corresponding QREs (e.g. QRE 27.2.1: disclose funding for the development of the third-party body), and for measure 27.3 they also have to report quantitative data about the number of research projects which have been granted data access.

Currently, the Code includes 44 commitments which are operationalized in 128 measures. It has been signed by 44 stakeholders, including big tech companies Google, Meta, Microsoft, and TikTok. Twitter (now called X; as the platform committed to the Code as Twitter this name is used throughout this paper) had been a signatory until May 2023 but the company decided to leave the Code under its new owner Elon Musk. Given its design, it is evident the CoP primarily aims to regulate Big Tech's role in disinformation dissemination. By signing the Code, signatories commit to report regularly, every six months in the case of signatories that qualify as Very Large Online Platforms (VLOPs) or Search Engines (VLOSEs) under the DSA, on their compliance with the Code. They do so using a predefined reporting template which enables both quantitative as well as qualitative comparisons between signatories. The template was developed by representatives of the European Regulators Group for Audiovisual Media Services (ERGA) that are part of the CoP Permanent Task-force in consultation with the signatories.

Participation in the Code is voluntary. However, the DSA will introduce obligations for VLOPs and VLOSEs defined as platforms with more than 45 million users in the EU. The DSA's Article 45 commends the formulation of voluntary codes of conduct, referencing the CoP as a potential example (Regulation 2022/2065, 106). Concretely, article 45(1) tasks the Commission and the European Board for Digital Services with promoting and facilitating the creation of codes of conduct to ensure proper application of the DSA, particularly addressing the challenges of systemic risks. Article 45(2) states that if significant systemic risks involving several VLOPs and/or VLOSEs arise, the Commission may invite stakeholders to develop codes with specific risk mitigation measures and a reporting framework.

Hence, codes of conduct are particularly relevant for VLOPs and VLOSEs in fulfilling two cornerstone articles of the DSA, namely articles 34 “Risk assessment” and 35 “Mitigation of risks”. Risk assessments must encompass various issues. Although the article does not explicitly mention disinformation, it implies comprehensive coverage of the phenomenon by including “any actual or foreseeable negative effects on civic discourse and electoral processes, and public security” (Art. 34(1)(c)). Mitigation measures must be taken by VLOPs and VLOSEs if systemic risks are identified in the assessments conducted as part of article 34. These measures may include modifying service designs and features, updating terms and conditions, im-

proving content moderation processes to quickly address illegal content, or adjusting "internal processes, resources, testing, documentation, or supervision of any of their activities in particular as regards detection of systemic risk" (Art. 35(1)(f)).

Codes of conduct offer the opportunity to specify these broad obligations and operationalise them into specific commitments. Additionally, they may provide detailed standards to facilitate internal compliance and external assessments, give a diverse set of stakeholders the chance to shape DSA implementation and allow for future action and adaptation to technological progress by addressing areas or gaps not covered in the DSA.

While these codes are soft law instruments and will continue to be voluntary, the DSA does specify that non-compliance with codes may attract penalties, which is a much bigger incentive for major platforms to cooperate (Griffin & Vander Maelen, 2023) and signals the EU's shift from a self- to a co-regulatory model when regulating disinformation. As a probable inaugural code under the DSA, it is crucial for the CoP to be unambiguous, strictly adhered to, and consistently monitored.

Considering the broad scope of the CoP, its focus on very large digital platforms and the relevance of monitoring compliance systematically, we ask two research questions:

RQ1: To what extent do platform signatories of the Code comply with their reporting obligations?

RQ2: How can compliance with the CoP be monitored systematically, regularly and comprehensively?

Method

This analysis centres on platform signatories classified as VLOPs/VLOSEs under the DSA, which are Google, Meta, Microsoft, TikTok, and Twitter. These platforms are often exploited by malicious entities and unsuspecting users to disseminate disinformation and misinformation. Given their vast user bases, they warrant significant scrutiny. To answer RQ1 we analysed the first set of CoP reports published by the platforms in early 2023. The reports are extensive (859 pages in total) but the structure of the predetermined reporting template enabled robust comparisons of signatory reports. We opted to analyse at the measure level, viewing commitments as too overarching and reporting elements as overly granular.

In our analysis we drew from content analysis methodology (Puppis, 2019). We for-

mulated a customised coding scheme, incorporating three quantitative and two qualitative metrics, detailed in the subsequent table.

TABLE 2: Coding scheme

VARIABLE	EXPRESSION	EXPLANATION
MISSING QUALITATIVE RESPONSES	yes	Information that is requested in a Measure or QRE is not provided comprehensively, and/or the information provided by the Signatory does not meet the obligations set out in the Measure or QRE. As some QREs cover several aspects of an issue, this variable should also be coded as "yes" when the Signatory fails to provide information of one or more relevant aspects.
	no	The qualitative responses provided by the Signatory are comprehensive, concise and to the point.
	irrelevant	If a Measure does not contain QREs, this variable might be coded irrelevant.
MISSING QUANTITATIVE DATA	yes	SLIs are ignored or relevant data that is requested as part of a Measure or SLI is not provided by the Signatory. This might also be the case if Member State specific data is requested but not provided. The variable is also to be coded "yes" if data provided is unsuitable or the methodology is ill-suited to provide reliable data.
	no	Quantitative data provided by the Signatory is comprehensive and the methodology to collect or calculate the data is appropriate.
	irrelevant	Measure does not contain SLI or there is reasonable ground why an SLI cannot be provided at this point in time.
OVERALL SCORE	1	Poor: the response significantly falls short of meeting the requirements of the Measure. This is the case for responses that lack major details, are incomplete or irrelevant, or fail to address the specific information requests outlined in the measure.
	2	Adequate: the response shows effort towards meeting the requirements of the Measure but there are notable issues or areas that require improvement. This is the case for responses that partially address the question, but may lack important details, evidence, or context.
	3	Good: the response fully meets the requirements of the Measure. This is the case for responses that are complete, relevant, and provide clear and comprehensive information that directly addresses the specific information requests outlined in the measure.
	n/a	Not Applicable: if a Signatory claims a Measure they subscribed to is not relevant to their services and the Assessor believes this claim to be correct e.g. the Measure relates to displaying information alongside political advertising and the Signatory's product does not allow political advertising.
SIGNATORY LEVEL COMMENTS	free text	Qualitative comments explaining the Coder's results for the previous three variables, e.g. "lacked detail on x". Coders should include any observations or feedback on the response.
CODE LEVEL COMMENTS	free text	Qualitative comments regarding the CoP; this might include comments on wording, structure, needed clarification, etc.

For coding, we adopted a three-step process. Initially, two coders independently analysed specific sections, with the option to flag uncertain codings as "needs a second opinion". After these initial rounds, a third round sought to finalise the results. The lead authors reviewed all discrepancies between the initial coders, refer-

encing comments and reports content to determine the final outcome. Additionally, they reviewed all codings for inconsistencies.

We present numerical outcomes for the first six sections. For the final three, focused on the Code's monitoring methods, we offer general remarks without numerical data, as many measures lacked distinct reporting elements or were absent entirely. From our qualitative analysis of the last three sections as well as from insights gained through the coding process for the first six sections we draw conclusions to offer potential answers to RQ2.

Results for RQ1: Assessment of platforms' compliance with the Code of Practice on Disinformation

TABLE 3: Overall scores per signatory (scores correspond to grades of 1 = Poor, 2 = Adequate and 3 = Good)

SIGNATORIES	% OF MEASURES MISSING QUALITATIVE INFORMATION	% OF MEASURES MISSING QUANTITATIVE DATA	AVERAGE OVERALL SCORE
Google	44%	53%	2.1
Meta	52%	65%	2.0
Microsoft	54%	59%	1.9
TikTok	43%	50%	2.0
TOTAL w/o Twitter	**49%**	**58%**	**2.0**
Twitter	100%	100%	1.0
TOTAL	**55%**	**64%**	**1.9**

The average score of signatories is a subpar 1.9, with only Google exceeding an overall grade of 2 (Adequate) at 2.1. Measures containing QREs revealed that 55% (with Twitter excluded 49%) were either incomplete or lacked qualitative information, while 64% (with Twitter excluded 58%) of Measures with SLIs were lacking at least some data.

Twitter left the Code in May 2023, so several months after the baseline reports had been published (Hendrix, 2023). Evidence suggests its commitment was waning well before its exit, averaging the lowest possible score 1 (Poor) and exhibiting a 100% deficiency in QREs/SLIs across all measures. The report contained numerous empty sections, limited general comments on some commitments, and information that often seemed directly copied from their website. This suggests Twitter had little intention of adhering to the Code when reporting.

Relevance and comprehensiveness of qualitative responses

The relevance of provided information is problematic, with many responses failing to be comprehensive, i.e. leaving out important aspects or answering reporting obligations by providing only superficial information. Instances where responses highlighted policy details instead of the requested implementation processes were common, and redirection to external links or other QREs often resulted in only partially relevant information.

To provide examples, in the "Empowering Users" section, QRE 17.1.1 asks Signatories to describe tools they develop or maintain to help users improve their media literacy and critical thinking skills. Instagram described an on-platform campaign for youth to raise awareness about the online safety features contained within Instagram such as reviewing privacy settings or managing screen time. This information is irrelevant to the aims of the commitment.

In response to measure 21.2, which in essence demands the application of labelling and warning systems for fact-checked content, Meta states "we know this program is working". But the signatory provides little evidence to back up this claim. Similarly, on measure 21.3 which obliges signatories to design labelling and warning systems "in accordance with up-to-date scientific evidence", Meta provides a vague description of working "in close consultation with fact-checkers and misinformation experts" (Meta, 2023, p. 94). The response needs a much clearer description of how they incorporate scientific evidence and user needs in their implementation of these programmes.

In the section "Empowering the Research Community" Google only lists Google Trends as a public, real-time data source for YouTube (Google, 2023, pp. 137–138). Google Trends, while useful for tracking general search interests over time, does not offer the most relevant data for studying disinformation. Arguably, the search function is not the most relevant feature of YouTube. More relevant data would include details on content that has been flagged, removals due to policy violations or the types of videos suggested to users.

Several responses in the "Empowering the Fact-Checking Community" section are also either lacking relevance, detail or both. This is particularly apparent for commitment 30, which obliges signatories to cooperate with EU fact-checkers under "transparent, structured, open, financially sustainable, and non-discriminatory" (p. 31) frameworks.

For example, Google stresses in their report their financial commitments to the In-

ternational Fact Checking Network (IFCN) as well as their financial contributions through the European Media and Information Fund (EMIF) (Google, 2023, pp. 153–155). However, Google does not state the number of agreements with fact-checking organisations through EMIF and IFCN per member state and language. It is problematic that Google does not disclose the fact that EMIF's support for fact-checking activities is only one of four funding lines within the fund's portfolio. Additionally, Google conflates contributions made to the international fact-checking community with support for EU fact-checkers, making it impossible to determine their true contribution in the scope of this commitment.

Microsoft's reporting for LinkedIn also lacks detailed information regarding commitment 30. For instance, there are undefined "pilot fact-checking arrangements" with news agencies (Microsoft, 2023, p. 147). There is also no clear indication whether the two mentioned agencies focus on the EU or globally. Additionally, LinkedIn does not offer any explanation on how they will ensure fact-checking coverage in all EU member states (a core provision in this section) or how they plan to aid fact-checkers at the national level. Furthermore, no information about the financial aspect of supporting fact-checking organisations is provided.

Overall, these examples illustrate the ways in which signatories fail to provide comprehensive and relevant qualitative information.

Lack of quantitative data

The absence of quantitative data is substantial. The methodology was occasionally dubious, and the data, often imprecise or completely absent. Comprehensive and reliable data submission is crucial for effective Code monitoring and evaluation of signatories' compliance. Otherwise researchers and regulators can neither conduct comparative analyses across platforms nor will reliable, longitudinal data become available in the long term.

For example, the lack of quantitative data was apparent for Measure 14.2. It obliges signatories to maintain and update a list of public policies of prohibited Tactics, Techniques, and Procedures (TTP) and report on enforcement of them. SLIs 14.2.1-4 ask for specific data on actions taken against TTPs at member state level. The reporting template lists 12 TTPs for which signatories could provide data, e.g. the creation of inauthentic accounts or botnets, use of fake engagement and fake followers, the creation of inauthentic pages, groups etc., inauthentic coordination of content creation or amplification, and non-transparent compensated messages by influencers (for full list see reporting template, e.g. Meta, 2023, p. 51).

Quantitative data on these TTPs is often missing or incomplete. For example, Microsoft only reports data about four TTPs (creation of inauthentic accounts or botnets; use of fake / inauthentic reactions; use of fake followers or subscribers; creation of inauthentic pages, groups, chat groups, fora, or domains). Microsoft argues that data for the remaining eight TTPs could not be computed. However, it is possible to apply most of these TTPs on LinkedIn. Furthermore, other signatories did provide data for other TTPs as well (Microsoft, 2023, pp. 53–64).

Additionally, how the data is computed is methodologically not robust: the figures reported for three TTPs are derived from a subset of inauthentic accounts reported for TTP1, likely resulting in an understatement of the actual presence of the respective TTPs. For example, for "use of fake followers or subscribers" Microsoft reports "a subset of the fake accounts reported in TTP 1 SLI 14.2.1 that followed a LinkedIn profile or page" (Microsoft, 2023, p. 60). This overlooks the possibility of a single fake account following multiple profiles or pages. Consequently, if reported accurately, the numbers for this TTP would likely be substantially higher.

In another example, TikTok reported that the number of fake accounts as a percentage of monthly active users is 0.0067 % – an implausibly low number (TikTok, 2023, p. 33). In contrast, Meta approximates the figure on their platforms to be around 5%. TikTok's claim necessitates a detailed examination, which is not within the scope of this study. Furthermore, the supplied quantitative data from TikTok is lacking: there's an absence of member state level data for SLI 14.2.4, and for some TTPs, nothing is reported.

Results per section

Signatories' adherence to the commitments varies within the first six subsections. All sections lack qualitative information, but the degree diverges, with "Empowering Users" missing at least some information in 37% of measures, in stark contrast to a missing 83% in the "Empowering Researchers" section, the latter scoring the lowest at 1.6. Responses in this section generally did not fulfil the information requirements, and SLIs were incomplete for all but Google. Excluding "Political Advertising", all sections scored below 2.

TABLE 4: Overall scores per section (scores correspond to grades of 1 = Poor, 2 = Adequate and 3 = Good)

SECTION	% OF MEASURES MISSING QUALITATIVE INFORMATION	% OF MEASURES MISSING QUANTITATIVE DATA	AVERAGE OVERALL SCORE
Scrutiny of ads placements	58%	64%	1.9

SECTION	% OF MEASURES MISSING QUALITATIVE INFORMATION	% OF MEASURES MISSING QUANTITATIVE DATA	AVERAGE OVERALL SCORE
Political Advertising	56%	0%	2.2
Integrity of services	56%	94%	1.9
Empowering Users	37%	52%	1.9
Empowering Researchers	83%	89%	1.6
Empowering the Fact-Checking Community	63%	72%	1.7

The “Empowering the Fact-checking Community” section was the next lowest at 1.7, marked by substantial deficiencies in both qualitative and quantitative data. The “Integrity of Services” section notably lacked 94% of quantitative data in relevant measures, due to absent member state level data, flawed methodologies, or complete data absence.

Conversely, the “Political Advertising” section saw the best performance, scoring 2.2, with signatories meeting quantitative data requirements. However, three out of five Signatories opted out of most measures in this section. Microsoft, TikTok, and Twitter cited their advertising policies which include a ban of political and issue ads, deeming most of this section irrelevant to their services. The scope of this study did not allow for actually testing whether these policies were enforced. It must be noted that previous research has found the enforcement on TikTok's service in this regard to be insufficient (Mozilla Foundation, 2021; Visser et al., 2023).

Results for RQ2: The need for systematic monitoring

In our analysis four aspects emerged as most relevant when considering how to monitor compliance with the CoP systematically, regularly, and comprehensively.

First, Section I(p) of the Code states that signatories should collaborate with the European Digital Media Observatory (EDMO) and ERGA, alongside the European Commission, for monitoring compliance with the Code. However, competencies and responsibilities for such monitoring remain undefined. If the Code is integrated as a code of conduct under the DSA, the enforcement and monitoring tasks might also be overseen by the Commission’s DSA enforcement team and the European Centre for Algorithmic Transparency (ECAT). Given the extensive scope of the involved tasks and the requisite for planning and allocation of resources, a detailed plan to guarantee appropriate monitoring of the Code should be devised.

Second, the Commission aspires to have the Code monitored at a member state

level, emphasising aspects of QREs or SLIs offering insights about a member state, possibly via ERGA representatives or EDMO hubs.[1] Presently, the scarcity of member state level data in reports raises concerns about the feasibility and value of such an approach. Any redundant efforts should be avoided.

Third, the Code contains several commitments and measures that are impossible to monitor externally as their assessment is delegated to the Permanent Task-force consisting of representatives from signatories, the European Commission, EDMO and ERGA. Subsequently, signatories state in their reports in some cases that they engage with the Permanent Task-force and its subgroups. The substance and sincerity of such engagement cannot be assessed externally without access to meeting minutes and/or interviews with participants.

Fourth, it is crucial to differentiate between monitoring the reports' compliance with the actual implementation of measures, ensuring the reported information is accurate. Previous monitoring, such as the CovidCheck report (Culloty et al., 2021), has highlighted discrepancies between platforms' claimed actions and reality. Verifying claims necessitates domain expertise, research capacity, and data access. For instance, commitment 14 of the Code requires information and data on the enforcement of outlined policies against specified TTPs, including the use of fake followers and engagement. A series of reports by NATO's Strategic Communications Centre of Excellence illustrates platforms' struggle to curb such manipulative techniques by commercial entities (Fredheim et al., 2023), emphasising the need for similar comprehensive evaluations across the Code.

One important step towards a more in-depth, systematic monitoring is actually codified in the CoP itself. As per commitment 41, the development and implementation of so-called Structural Indicators is foreseen in the Code. The exact scope of such indicators is not defined in the Code, but they are supposed to "assess the effectiveness of the Code in reducing the spread of online disinformation for each of the relevant Signatories, and for the entire online ecosystem in the EU and at Member State level" (CoP, 2022, p. 40).

The release of the initial set of Structural Indicators was scheduled within nine months after the Code's signing, subsequent to the publication of the baseline reports. However, this has not yet happened, with no workable proposal for Structur-

1. The European Digital Media Observatory (EDMO) and its 14 regional hubs form a network of researchers, fact-checkers, journalists and media literacy practitioners working on countering disinformation. Their goal is to provide insights to counter disinformation and serve as a platform for collaboration within the anti-disinformation ecosystem. Legally, EDMO and its hubs are 15 EU-funded projects without a legal entity.

al Indicators tabled by the responsible Working Group within the set timeframe, as outlined in measure 41.3. A pilot study proposing and implementing a limited set of Structural Indicators has been conducted by the company TrustLab (TrustLab, 2023). Additionally, a more comprehensive framework was proposed by EDMO (Nenadic et al., 2023) which encompasses indicators gauging the prevalence, sources, audience, and demonetisation of disinformation, collaboration, and investments in fact-checking, as well as investments in the overall implementation of the Code. While the pilot study by TrustLab is too limited in scope and scale, the EDMO proposal is too vague and in part redundant with reporting obligations that are already part of the Code itself. To assess the prevalence of disinformation on different platforms, the pilot study by TrustLab uses the concept of discoverability, defined as the "percentage of content returned from searching disinformation keywords which is mis/disinformation" (p. 13). This concept allows for cross-platform comparisons but does not reflect common user practices, i.e. accessing content through personalised feeds. In addition, the TrustLab study is limited geographically, only studying disinformation in three countries. The EDMO working paper proposes several indicators to gauge the different dimensions of disinformation (e.g. prevalence, audience, sources), but some of these indicators are only very vaguely described (in particular on demonetisation of disinformation (p. 17). Thus, the commitment by signatories to develop and implement Structural Indicators remains unfulfilled. This oversight is significant given the crucial role of Structural Indicators in assessing the essence of the signatories' reporting.

Conclusion

This study presents the first comprehensive monitoring of the CoP platform signatories baseline reports. The analysis is particularly relevant as the European Commission shifts from a self-regulatory framework to a co-regulatory model. With the likelihood of the Code being recognised as a code of conduct under the DSA, compliance could become effectively mandatory.

Our analysis involved a newly developed monitoring framework applied to the baseline reports from Google, Meta, Microsoft, TikTok, and Twitter. The findings indicate that these platforms fall short in fulfilling their reporting obligations. Notably, Twitter's complete failure, evidenced by their inadequate report even before withdrawing from the Code, suggests a lack of intent to comply. Our findings align with prior research that illustrates Twitter's transformation into a largely unmoderated platform, disregarding legal obligations to oversee and regulate speech on the platform. As a result, the EU Commission opened formal proceedings against

Twitter under the DSA citing shortcomings of “measures taken to combat information manipulation on the platform” (European Commission, 2023).

Regarding the remaining platforms, the information and responses provided often fail to precisely address the Code's specific requirements, with many instances of incomplete information. The reporting of quantitative data requires significant improvement in both scope and detail. The methodologies for calculating quantitative data need to be more accurate, with an emphasis on providing member state-level data.

It is crucial to emphasise that our analysis must be regularly replicated and refined to maximise its utility for policy evaluation and enforcement. We encourage researchers to build upon and apply our methodological framework to other platform reports on compliance with the CoP. We believe it is easily replicable and platform-agnostic, i.e. if more VLOPs or VLOSEs join the Code their reports can be analysed applying the same methodological framework.

Signatories and regulators should pay special attention to ensuring comparable methodologies when reporting SLIs. Although challenging, due to the platforms' differing focuses, metrics, and specifications that complicate comparisons, there are still data points that allow for comparisons (e.g. related to TTPs). Additionally, the data currently hidden in platform reports could be integrated into the DSA Transparency Database[2] and made available in a machine-readable format.

Repeated research over time will yield invaluable insights into how platforms evolve in countering disinformation, how the implementation of the DSA influences their policies and enforcement actions, and their overall behaviour. This iterative process will refine the methodology, identify gaps in platform reporting, and highlight statements that warrant further investigation. This will be particularly significant when the Code of Practice transitions into a code of conduct under the DSA. As it is likely to be the first one, the way this transition is managed, its impact, and its implementation by regulators and policymakers will largely determine the DSA's effectiveness in regulating harmful but mostly lawful speech.

Given recent developments and investigations initiated by the European Commission into the behaviour and policies of very large online platforms, the importance of this work cannot be overstated. In its press release announcing the opening of formal proceedings against Meta, the European Commission mentions specifically

2. See European Commission (n.d.).

that it "suspects that Meta does not comply with DSA obligations related to addressing the dissemination of deceptive advertisements, disinformation campaigns and coordinated inauthentic behaviour in the EU" (European Commission, 2024). Monitoring compliance with the Code of Practice may offer valuable insights for such investigations and indicate areas where platforms fall short in their responsibility to counter disinformation. Even though our analysis was focussed on monitoring compliance with reporting obligations, we found several instances in which platforms' claims are at least doubtful if not even contradicted by already published empirical evidence. Areas of concern are the enforcement of bans on political advertisements and actions against fake engagement and inauthentic behaviour. Therefore, future monitoring should be accompanied by in-depth investigations into specific focus areas to further enhance our understanding and improve regulatory practices. The European Commission should ensure this monitoring is adequately funded and resourced long-term, conducted by independent researchers with the necessary background knowledge. Only after these steps towards a systematic monitoring have been taken, will we be able to answer with certainty the question: but did platforms really do what they claim to have done?

Limitations

While this study provides valuable insights into platforms' compliance with the CoP, it is essential to acknowledge certain limitations. Firstly, our analysis is based on the initial set of reports published by platform signatories in early 2023. Since then, these signatories have published a second round of reports in September 2023. Our results do not reflect any changes to reporting compliance through these reports. Secondly, the monitoring framework developed for this analysis, while comprehensive, may have inherent subjectivity in its assessment criteria. The interpretation of policy compliance and reporting quality can vary. Through application of a three-step coding process we try to minimise this subjectivity but may not eliminate it entirely. Thirdly, the analysis is focused on monitoring compliance with reporting obligations. It does not comprehensively assess the accuracy of information and data reported by signatories. As noted, we identified several instances in which accuracy is at least doubtful. This calls for more investigatory approaches to monitor CoP compliance. Lastly, the analysis focuses on platform signatories, and the findings may not generalise to platforms that are not (yet) signatories of the Code.

ACKNOWLEDGEMENTS

The authors thank all colleagues who contributed to the analysis of platform reports and gave feedback during the drafting of the article:

Mato Brautović, University of Dubrovnik

Eileen Culloty, Dublin City University

Dawn Holford, University of Bristol

Anastasia Kozyreva, Max Planck Institute for Human Development

Stephan Lewandowsky, University of Bristol, University of Potsdam, and University of Western Australia

Trisha Meyer, Vrije Universiteit Brussel

Susanne Wegner, TU Dortmund University

Kirsty Park and Eillen Culloty received co-funding for this research from the European Union under action number 2020-EU-IA-0282 and agreement number INEA/CEF/ICT/A2020/2381686.

Susanne Wegner and Stephan Mündges received co-funding for this research from the European Union, project number 101083573.

Mato Brautović received co-funding for this research from the European Union, project number 101083909.

Trisha Meyer received co-funding for this research from the European Union, project number INEA/CEF/ICT/A2020/2394296.

Stephan Lewandowsky has received funding from Jigsaw (a technology incubator created by Google) for several empirical projects unrelated to the present paper. He also acknowledges financial support from the European Research Council (ERC Advanced Grant 101020961 PRODEMINFO) and the European Commission (Horizon 2020 grants 964728 JITSUVAX and 101094752 SoMe4Dem). He also receives support from UK Research and Innovation (through EU Horizon replacement funding grant number 10049415).

Anastasia Kozyreva and Stephan Lewandowsky acknowledge funding from the Volkswagen Foundation (grant “Reclaiming individual autonomy and democratic discourse online: How to rebalance human and algorithmic decision making”).

References

Altay, S., Berriche, M., & Acerbi, A. (2023). Misinformation on misinformation: Conceptual and methodological challenges. *Social Media + Society*, *9*(1). https://doi.org/10.1177/20563051221150412

Altay, S., Kleis Nielsen, R., & Fletcher, R. (2022). Quantifying the "infodemic": People turned to trustworthy news outlets during the 2020 coronavirus pandemic. *Journal of Quantitative Description: Digital Media*, *2*. https://doi.org/10.51685/jqd.2022.020

Cipers, S., Meyer, T., & Lefevere, J. (2023). Government responses to online disinformation unpacked. *Internet Policy Review*, *12*(4). https://doi.org/10.14763/2023.4.1736

Culloty, E., Park, K., Feenane, T., Papaevangelou, C., Conroy, A., & Suiter, J. (2021). *Covidcheck: Assessing the implementation of EU code of practice on disinformation in relation to Covid-19* [Project report]. DCU Institute for Future Media, Democracy and Society. https://doras.dcu.ie/26472/

Ecker, U., Roozenbeek, J., van der Linden, S., Tay, L. Q., Cook, J., Oreskes, N., & Lewandowsky, S. (2024). Misinformation poses a bigger threat to democracy than you might think. *Nature*, *630*(8015), 29–32. https://doi.org/10.1038/d41586-024-01587-3

European Commission. (2018). *Tackling online disinformation: A European approach* (Communication COM(2018) 236 final). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52018DC0236

European Commission. (2023). *Commission opens formal proceedings against X under the Digital Services Act* [Press release]. https://ec.europa.eu/commission/presscorner/detail/en/IP_23_6709

European Commission. (2024). *Commission opens formal proceedings under DSA* [Press release]. https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2373

European Commission. (n.d.). *DSA transparency database*. https://transparency.dsa.ec.europa.eu/

European Fact-Checking Standards Network. (2023). *Fact-checkers' feedback on the*

baseline reports of the VLOP signatories of the Code of Practice on Disinformation [Report].

European Regulators Group for Audiovisual Media Services. (2020). *ERGA Report on disinformation: Assessment of the implementation of the Code of Practice* [Report]. https://erga-online.eu/wp-content/uploads/2020/05/ERGA-2019-report-published-2020-LQ.pdf

Fredheim, R., Bay, S., Dek, A., Stolze, M., & Haiduchyk, T. (2023). *Social media manipulation 2022/2023: Assessing the ability of social media companies to combat platform manipulation* [Report]. NATO Strategic Communications Centre of Excellence. https://stratcomcoe.org/publications/social-media-manipulation-20222023-assessing-the-ability-of-social-media-companies-to-combat-platform-manipulation/272

Google. (2023). *Code of Practice on Disinformation – Report of Google for the period 1 July 2022—30 September 2022* [Dataset]. Transparency Centre. https://disinfocode.eu/reports-archive/?years=2023

Griffin, R., & Vander Maelen, C. (2023). *Codes of conduct in the Digital Services Act: Exploring the opportunities and challenges*. SSRN. https://doi.org/10.2139/ssrn.4463874

Hendrix, J. (2023, May 26). Musk's Twitter ditches EU Code of Practice on Disinformation. *Tech Policy Press*. https://techpolicy.press/musks-twitter-ditches-eu-code-of-practice-on-disinformation/

High Level Group on Fake News and Online Disinformation. (2018). *A multi-dimensional approach to disinformation* [Report]. European Commision. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50271

Jungherr, A., & Schroeder, R. (2021). Disinformation and the structural transformations of the public arena: Addressing the actual challenges to democracy. *Social Media + Society*, *7*(1). https://doi.org/10.1177/2056305121988928

Kiely, K. P., Margova, R., Gargova, S., Dobreva, M., Gandova, T., & Stefanova, T. (2023). *Evaluating VLOP and VLOSE implementation of the Strengthened EU Code of Practice on Disinformation in Bulgaria* [White paper]. GATE Institute Big Data for Smart Society. https://brodhub.eu/en/research/evaluating-vlop-and-vlose-implementation-of-the-strengthened-eu-code-of-practice-on-disinformation-in-bulgaria/

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008

Meta. (2023). *Code of Practice on Disinformation – Meta baseline report* [Dataset]. Transparency Centre. https://disinfocode.eu/reports-archive/?years=2023

Microsoft. (2023). *Code of Practice on Disinformation Baseline Report – January 2023 Microsoft* [Dataset]. Transparency Centre. https://disinfocode.eu/reports-archive/?years=2023

Mozilla Foundation. (2021). *Th€se are not po£itical ad$: How partisan influencers are evading TikTok's weak political ad policies* [Report]. https://foundation.mozilla.org/documents/178/TikTok-Advertising-Report_e5GrWx5.pdf

Nenadic, I., Brogi, E., & Bleyer-Simon, K. (2023). *Structural indicators to assess effectiveness of the EU's Code of Practice on Disinformation* (Working Paper 2023/34). European University Institute. https://hdl.handle.net/1814/75558

Park, K., & Mündges, S. (2023). *CoP monitor: Baseline reports. Assessment of VLOP and VLOSE Signatory reports for the Srengthened Code of Practice on Disinformation* [Report]. EDMO Ireland & German-Austrian Digital Media Observatory. https://gadmo.eu/wp-content/uploads/2023/09/CoP-Monitor-Report.pdf

Peukert, A. (2023). Desinformationsregulierung in der EU – Überblick und offene Fragen [Disinformation regulation in the EU - overview and open questions]. *JuristenZeitung*, *78*(7), 278–296. https://doi.org/10.1628/jz-2023-0095

Puppis, M. (2019). Analyzing talk and text I: Qualitative content analysis. In H. Van den Bulck, M. Puppis, K. Donders, & L. Van Audenhove (Eds.), *The Palgrave handbook of methods for media policy research* (pp. 367–384). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-16065-4_21

Regulation 2022/2065. (2022). *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)*. European Parliament and Council. https://eur-lex.europa.eu/eli/reg/2022/2065

Stasi, M. L., & Parcu, P. L. (2021). Disinformation and misinformation: The EU response. In P. L. Parcu & E. Brogi (Eds.), *Research handbook on EU media law and policy*. Edward Elgar Publishing. https://doi.org/10.4337/9781786439338.00030

The Strengthened Code of Practice on Disinformation. (2022). https://disinfocode.eu/wp-content/uploads/2023/01/The-Strengthened-Code-of-Practice-on-Disinformation-2022.pdf

TikTok. (2023). *Code of Practice on Disinformation – Report of TikTok for the period 16 June – 16 December 2022* [Dataset]. Transparency Centre. https://disinfocode.eu/reports-archive/?years=2023

TrustLab. (2023). *A comparative analysis of the prevalence and sources of disinformation across major social media platforms in Poland, Slovakia, and Spain* [Report]. https://my.visme.co/view/vdpvxy4j-code-of-practice-on-misinformation-september-2023#s1

van der Linden, S. (2023). *Foolproof: Why misinformation infects our minds and how to build immunity*. W.W. Norton & Company.

Visser, F., Smirnova, J., & Martiny, C. (2023). *Cashing in on conflict: TikTok profits from pro-Kremlin disinformation ads* (Digital Dispatches). Institute for Strategic Dialogue. https://www.isdglobal.org/digital_dispatches/cashing-in-on-conflict-tiktok-profits-from-pro-kremlin-disinformation-ads/

Yang, Y., Davis, T., & Hindman, M. (2023). Visual misinformation on Facebook. *Journal of Communication*, *73*(4), 316–328. https://doi.org/10.1093/joc/jqac051

Published by

Alexander von Humboldt Institute for Internet and Society

in cooperation with

CREATe

cnrs centre internet et societe

R&I IN3 Internet interdisciplinary Institute Universitat Oberta de Catalunya

University of Tartu Johan Skytte Institute of Political Studies