Michalon, Barthélémy

Article

The role of civil society organisations in co-regulating online hate speech in the EU: A bounded empowerment

Internet Policy Review

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

Suggested Citation: Michalon, Barthélémy (2025) : The role of civil society organisations in co-regulating online hate speech in the EU: A bounded empowerment, Internet Policy Review, ISSN 2197-6775, Alexander von Humboldt Institute for Internet and Society, Berlin, Vol. 14, Iss. 1, pp. 1-29,
https://doi.org/10.14763/2025.1.1826

This Version is available at:
https://hdl.handle.net/10419/315583

INTERNET POLICY REVIEW

Volume 14 **Issue** 1

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

The role of civil society organisations in co-regulating online hate speech in the EU: A bounded empowerment

Barthélémy Michalon *Tecnologico de Monterrey* bmichalon@tec.mx

DOI: https://doi.org/10.14763/2025.1.1826

Published: 31 March 2025
Received: 19 March 2024 **Accepted:** 11 October 2024

Funding: The author did not receive any funding for this research.
Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: Michalon, B. (2025). The role of civil society organisations in co-regulating online hate speech in the EU: A bounded empowerment . *Internet Policy Review*, *14*(1). https://doi.org/10.14763/2025.1.1826

Keywords: Civil society organisations, Hate speech, Co-regulation, Social media platforms, Code of conduct

Abstract: Civil society organisations (CSOs) have been gradually expanding their role in monitoring the Code of Conduct jointly developed in 2016 by the European Commission and four large social media platforms to counter online hate speech. While their function was initially limited to collecting data and transmitting it to the EU executive, over the years CSOs have not only expressed critical views on the "monitoring exercise" designed to assess the agreement's effectiveness but also devised ways to support their claims and elevated their involvement in the functioning of the mechanism. Drawing on data from two surveys carried out in 2019 and 2022, over twenty interviews conducted in the same timeframe, as well as reports published by these organisations, this paper provides insights into the evolution of CSOs' perceptions on the matter and examines how this category of players became closely embedded in the operational side of the Code of Conduct. The evidence presented suggests that civil society has expanded its role beyond the specific task initially entrusted to it by the Commission. Yet, despite this shift, CSOs' role remains limited to that of a third-party participant, an expression that reflects both the reality and the limitations of their role within the co-regulatory scheme.

This paper is part of **Content moderation on digital platforms: beyond states and firms**, a special issue of *Internet Policy Review* guest-edited by Romain Badouard and Anne Bellon.

Introduction

In May 2016, Facebook, Twitter,[1] Microsoft and YouTube undersigned the "Code of Conduct to counter illegal hate speech online", promoted by the European Commission. In this voluntary, three-page document, the companies agreed to 12 commitments, chief among them the review within 24 hours of the majority of content reported as hate speech by their users, and its removal if it indeed fell within this category. Since then, eight other firms have successively joined the initiative.[2]

With the aim of creating an incentive for the platform companies to implement their commitments, the Commission attached a "monitoring exercise" (ME) to the Code in the months following its adoption. This exercise has taken place for about six weeks on a yearly basis. It involved civil society organisations (CSOs) reporting instances of hate speech to the firms. Importantly, when flagging this content, they first do it through the platform's reporting tools under the guise of regular users. If the company ignores their initial notice, they start the process over, this time through specific means reserved to "trusted flaggers".[3] CSOs register the companies' responses – if any – and then transmit their data to the Commission, which incorporates it in a publicly available annual report.

Over time the number of involved organisations grew from 12 in the first round, carried out at the end of 2016, to 36 from 21 different states in the seventh, at the beginning of 2022. At the time of writing, no ME rounds have taken place since then, as negotiations for a renewed version of the Code started in March 2023 to align it with the formal requirements established by the Digital Services Act for codes of conduct (European Commission, 2023, p. 8). Though the new Code has been approved by mid-2024, it has not been released yet.

1. Throughout this paper, I use the names that these two companies (now Meta and X, respectively) had at the time when the document was negotiated, to maintain consistency with their designation in the primary sources I use.
2. Instagram, Snapchat and Dailymotion joined the Code in 2018, Jeuxvideo.com in 2019, TikTok in 2020, LinkedIn in 2021 and finally Rakuten Viber and Twitch in 2022.
3. Trusted flagger, also known as "trusted reporter", is a status granted by digital platforms to selected CSOs, providing them with direct communication channels and, ostensibly, a more expedited handling of reported cases. While this status was institutionalised in the Digital Services Act, adopted in the EU in 2022, the practice has existed since at least 2012 as a result of a YouTube initiative, soon imitated by other companies (Gillespie, 2018, p. 131).

The Code of Conduct has caught the attention of researchers, who explored its impact on freedom of expression (Bukovská, 2019; Kuczerawy, 2017; Quintel & Ullrich, 2018), on the companies' activities in policing hate speech (Alkiviadou, 2019; Aswad, 2016; Cavaliere, 2019), on the intersection of both dimensions (Coche, 2018) or on the power balance between public actors and private digital companies (Michalon, 2024, p. 97). In contrast, the role of civil society organisations in the monitoring system appears to have been overlooked in scholarship so far.

This gap can be explained by the fact that the CSOs' most visible contribution to the Code's ME was limited to collecting data on the responses they received from the platform companies and transmitting it to the Commission, which could be regarded as merely instrumental at first sight. However, this role should not be underestimated for at least two reasons. First, the fact that these figures come from independent organisations specialised in combating hate speech online lends expertise (Facing Facts, 2022, p. 19; Klingvall, 2023a) and additional legitimacy (Gorwa, 2019, p. 13) to the entire evaluation mechanism. Second, the reports periodically published by the Commission fully rely on data supplied by CSOs. As these reports are key to determining whether the scheme is functioning as intended, the significance of the underlying data cannot be overstated, especially given that both the Commission and platform companies have a vested interest in delivering results that support the Code's effectiveness in producing the expected outcomes.

Beyond this formal and key role in the ME, CSOs have found ways to move past this form of participation. Notably, the International Network Against Cyber Hate (INACH), a network of over 30 CSOs, has gradually become more embedded in some logistical aspects of the ME operation, transcending a data-collection role. Additionally, CSOs have been openly advocating for improving the monitoring exercise that they know well from the inside thanks to their direct participation in it. In addition to making these calls certain groups have even taken concrete steps to support their criticisms by separately collecting empirical evidence of the ME's weaknesses and shortcomings.

All these aspects, which I elaborate on in this paper, provide compelling reasons to consider CSOs when studying the Code of Conduct. Including these actors not only offers a more comprehensive understanding of the scheme but also a more accurate one, challenging the common notion that this is fundamentally a bilateral scheme involving the Commission on one side and a few digital companies on the other. While this characterisation was valid during the drafting of the Code's content (Gorwa, 2019, p. 7), adhering to it overlooks the role of CSOs as an increasingly relevant third party in its implementation process.

This paper interrogates the CSOs' take and impact on the Code's monitoring exercise: first, how have civil society organisations been assessing its effects over time? Second, to what extent have they managed to play an active role within and beyond their data-collection role? The answers to these questions are meant to allow for broader inferences regarding the system of co-regulation of hate speech content in the European Union.

After outlining the theoretical background and methodology, this article analyses CSOs' perceptions on the Code of Conduct from a dynamic perspective by comparing survey results and interviews conducted in 2019 and 2022. It then examines the separate monitoring exercises implemented by various groups of organisations to offer an alternative account of the platform companies' performance in addressing hate speech. Further, this paper provides insights into recent developments that, by entrusting an umbrella group of CSOs with new key functions in the ME's operation, have decisively shifted the Code away from a two-sided model of cooperation. Finally, it concludes by qualifying the nature of the CSOs' role within this coregulatory framework.

Section 1 – Theoretical framework and methodology

1.1 The Code of Conduct as a co-regulatory mechanism

While the Code of Conduct has been developed by a few platform companies, the Commission was closely involved in promoting it in the first place, in the negotiation process over its content, and ultimately in designing the mechanism to monitor its implementation by the participating firms. Given that the Commission has played a key role on several key aspects of the Code, I argued elsewhere that the Code fundamentally constitutes a co-regulatory instrument (Michalon, 2024, p. 149) rather than a self-regulatory one as some authors (Quintel & Ullrich, 2018) or even the Commission (European Commission, 2019) have argued.

I identified two main threads of definitions for "co-regulation". On the one hand, some authors define it as a mechanism involving a regulator belonging to public authorities and private actors from the regulated sector (see for instance Hirsch, 2011, p. 441; Horowitz, 2024, p. 16; Kleinsteuber, 2004, p. 63). This is what I refer to here as a *bilateral* understanding of co-regulation. On the other hand, other researchers consider a broader range of participants, including actors from civil society (Finck, 2017, p. 15; Marsden, 2004, p. 80; Rubinstein, 2018, p. 504; Yasuda, 2016, p. 430), which Steurer (2013, p. 398) designated as tripartite co-regulation. I refer to it as a *multilateral* conception of co-regulation, so as to mark the contrast

with the bilateral approach mentioned earlier, and to allow for the potential inclusion of actors from more than three categories.

Notably, the threshold for being considered of a co-regulatory nature is lower in the bilateral view, as it is reached as soon as both public and private actors play a meaningful role in the regulatory scheme. While this approach does not discard participation from other actors, it does not *require* such additional participation for the system to be of a co-regulatory nature. However, it is also worth clarifying that the involvement of a third category of actors from civil society would not negatively affect its characterisation as co-regulation, as long as the fundamental condition of having both public and private actors playing a regulatory role is met. In contrast, the multilateral approach to co-regulation places the threshold higher, as the participation of actors of a third kind – typically, civil society, becomes a strict requirement to reach it.

By exploring the CSO's evolving role within the Code of Conduct and its limits, my research aims to demonstrate that this bounded involvement of civil society is indicative of the scope and nature of the co-regulatory mechanism under study.

1.2 An empirical method mostly based on primary sources

This research aims to gather empirical evidence on CSOs' perceptions of the Code's monitoring exercise, and on the scope and limits of their role within the evaluation scheme. To achieve this, I combine qualitative information from interviews with quantitative data collected through two online surveys.

Between January and April 2019, in the context of broader research, I conducted interviews with representatives from public authorities, digital companies, and CSOs, most of which took place in person. In May-June 2022, I held a second round of interviews, this time exclusively online and with CSOs representatives, aligned with the specific focus of this study. To respect the anonymity requested by several interviewees, I refer to all participants solely as members of their respective organisations.

TABLE 1: Interviews conducted in the 2019 and 2022 rounds

JANUARY-APRIL 2019 INTERVIEWS			MAY-JUNE 2022 INTERVIEWS	
Platform companies	- Google (French office) - Google (Belgian office) - Dailymotion (French office) - Facebook (French office)		CSOs	- Centre for Peace Studies (Croatia) - Active Watch (Romania) - Háttér (Hungary)

JANUARY-APRIL 2019 INTERVIEWS		MAY-JUNE 2022 INTERVIEWS	
Public authorities	- European Commission - European Commission - European Commission - DILCRAH (France)		- DigiQ (Slovakia) - LICRA (France) - Never Again Association (Poland) - Romea (Czech Republic) - INACH (international): three interviews
CSOs	- LICRA (France) - UNIA (Belgium) - INACH (international): two interviews, both online		

In addition to these direct conversations, I conducted an online survey between April and July 2019, followed by a second one in April-May 2022. In both cases, I sent questionnaires to any organisations meeting these conditions:

- Being listed among the participants in the ME, according to the latest report on the Code's implementation published by the Commission prior to each survey[4]
- Having reported at least five pieces of content according to the corresponding report on the Code's implementation
- Formally being a civil society organisation, as opposed to a specialised government agency[5]

On both occasions, I sent individual emails to the eligible CSOs to describe the purpose of the consultation and provide the hyperlink to the Google Forms-based survey. The questionnaire was intentionally short, with six questions in 2019 and seven in 2022, almost identical in both editions to allow for inferences from the observed evolutions. Without a response within a week, I issued a written reminder and then resorted to phone calls when necessary (see Table 2).

TABLE 2: Broken-down count of CSOs considered, contacted and participating in the surveys

	2019	2022
- Organisations of any kind involved in the latest ME	39	35
- Civil society organisations involved in the latest ME	32	31
- CSOs having reported at least five pieces of content in the latest ME	30	29

4. Respectively, the fourth implementation report issued in February 2019, and the sixth issued in October 2021 (European Commission, 2019, 2021).
5. A small number of organisations involved in the monitoring exercises were not CSOs but public bodies belonging to the state administration. Three organisations were in this case in 2022: PHAROS (a governmental platform that collects users notices in France), and the Spanish Ministry of Interior and the Spanish Observatory on Racism and Xenophobia (OBERAXE) (European Commission, 2021, p. 5).

	2019	2022
- Answers received	27	24
- Answers considered valid[6]	27	23

Thus, my surveys enabled me to collect responses from 90% (27/30) and 79.3% (23/29) of the eligible CSOs in 2019 and 2022 respectively. This slight decline may be attributed to the fact that, by coincidence, my 2022 survey partially overlapped with that year's ME. In this context, personnel may have been busier – especially in small organisations[7] – or reluctant to answer questions about an ongoing process.

Additionally to these first hand sources, I complemented my research with an analysis of public reports issued by the Commission and by CSOs, outlining the outcomes of various types of monitoring exercises.

Section 2 – CSO's views on the monitoring exercise: sceptical on outcomes, hopeful on process

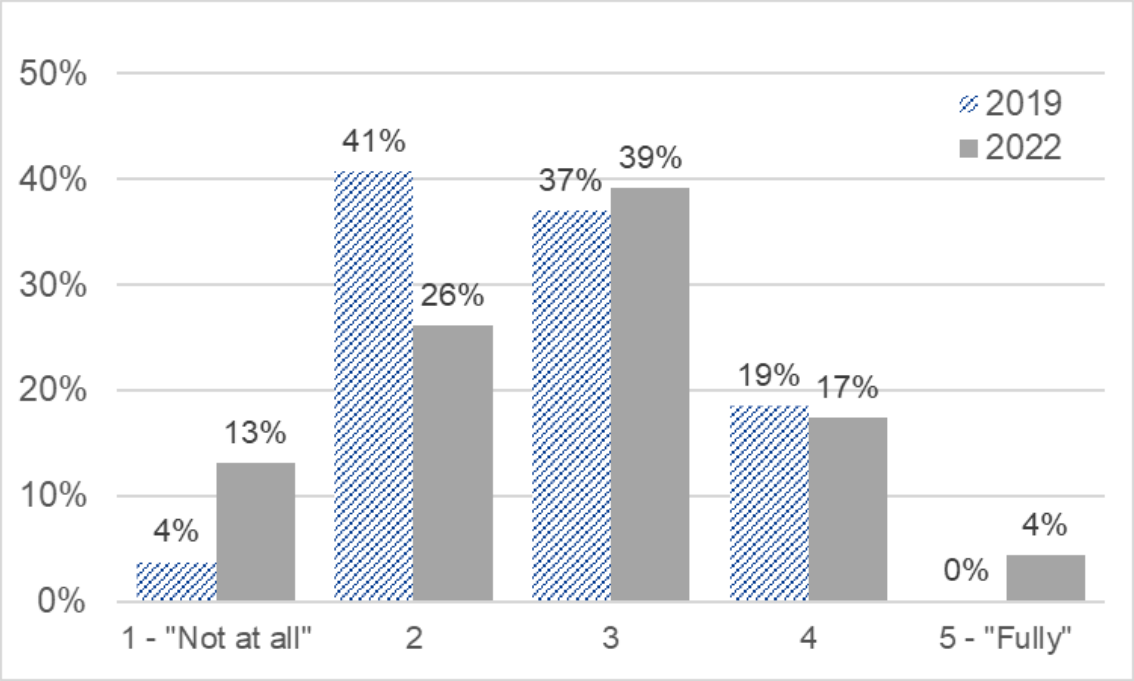
2.1 Questions and concerns regarding the monitoring exercise's representativeness

I define the ME's *representativeness* as the extent to which the figures obtained through this evaluation system reflect the platforms' actual and usual behaviour towards hate speech. To assess CSOs' perceptions of whether the ME properly captures how companies treat this type of content, I included the following question in both surveys:

Do you consider that the reports on the implementation of the Code of Conduct provide an accurate representation of the way hate speech is being dealt with on a regular basis by the platforms?

The responses consistently indicate that most CSOs perceive the monitoring exercise as having a medium to medium-low level of representativeness (see Graph 1).

6. In 2022, I had to set aside the input from a CSO because of a pattern of inconsistencies in its answers and for failing the "control question" that I had included to detect responses from unqualified people in the organisation.
7. As I explain in Section 4, CSOs overwhelmingly describe the ME as particularly demanding on their capacities.



GRAPH 1: Distribution of CSOs' assessments of the monitoring exercise's representativeness. Source: 2019 and 2022 surveys.

The average assessment of the ME's representativeness remained remarkably stable, staying below the middle range, since it only increased from 2.70 to 2.74 out of 5 between 2019 and 2022. However, this apparent stability at the aggregate level is not confirmed, at least not to the same extent, by a more granular analysis: out of the 15 CSOs whose responses could be tracked over time,[8] only one third maintained the exact same rating, while six of them increased it and four decreased it, all by just one point. Thus, this indicator is more volatile than it appears. Moreover, the standard deviation increased from 0.85 to 1.05, revealing that perceptions are gradually diverging.

While CSOs do not wholly reject the ME's representativeness, the average stance can be characterised by long-standing and widespread scepticism. This quantitative evidence was reinforced during interviews with CSO members, who recurrently expressed the view that companies' engagement with hate speech on their platforms varied depending on whether a monitoring exercise was taking place. As one participant put it in an in-person interview in 2019, "No one takes sick leave or goes on holidays [in the relevant departments at the European headquarters of the digital firms] during the testing period" (personal communication, February 19th, 2019). Although the Commission does not disclose the precise dates of the

8. I could not trace the evolution of the responses from CSOs that chose to answer my survey anonymously.

MEs to the companies, it does provide a broader timeframe, which an interviewed Commission official defended as a matter of transparency. Additionally, several CSO operatives emphasised, both in 2019 and 2022, that social media companies possess the technical means to detect unusual patterns of content flagging, which enable them to quickly deduce that a monitoring exercise is underway.

Thanks to this knowledge, companies have an opportunity to alter their behaviour when they are aware they are being observed, a phenomenon known as the Hawthorne Effect (Brannigan & Zwerman, 2001). In its 2017 annual report, the umbrella organisation INACH regretted that "[t]here was an undeniable bias that took place due to the fact that social media companies were informed about many details of the exercise, such as when it was going to take place and who was involved" (Berecz & Devinat, 2017a, p. 33). The same network later expanded these criticisms stressing that

social media companies are too involved in the development and organisation of the official monitoring exercises. This too deep involvement skews the outcome of these exercises and provides an environment that is too biased towards the interests of the companies. Thus, we repeat it here again, INACH recommends to the EC to keep social media companies in the dark, not just about the starting date of the exercises, but everything else related to the MEs […]. This is the only way to gain a representative and full picture of the efficacy of the CoC (Berecz, 2019, p. 15).

An interviewed Commission official conceded that the firms are likely aware of such monitoring, but pushed back on the notion that this knowledge could influence their handling of hate speech.

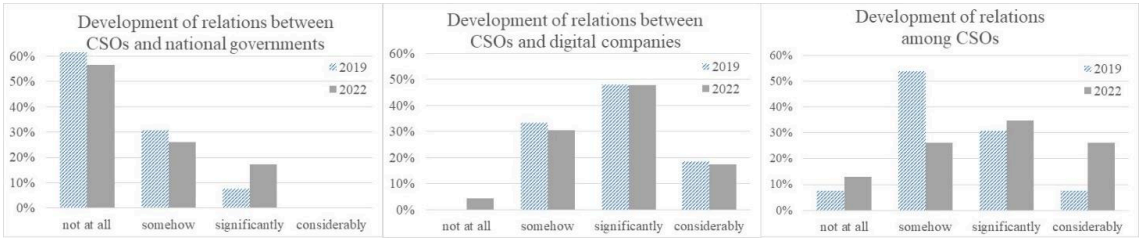
2.2 Recognising the intrinsic value of the process itself

Despite their persistent scepticism about the ME's representativeness, CSOs kept participating in the scheme year after year, with their numbers even increasing. This counterintuitive trend suggests that they have found value in other aspects of their involvement.

Drawing on insights from representatives of all three sectors during my first round of interviews, I hypothesised that, despite their disappointment with the results produced by the ME, CSOs valued the existence of the process itself. Therefore, I included in my survey a question to assess the extent to which the Code of Conduct (including its ME) had contributed to developing the relationship between CSOs and different classes of actors:

Regarding the issue of hate speech online, how much has the Code of Conduct contributed to develop the relations between your organisation and: a) the platforms b) the government of the country you are established in c) other civil society organisations concerned by the same topic?

A clear majority of CSOs responded that the Code had no impact at all on their relations with governments (see Graph 2). This outcome is unsurprising, as the Code's operation does not inherently involve public authorities, which played no active role in its design or implementation, even though they eventually endorsed the overall scheme.[9]As highlighted during interviews, the few responses indicating improved CSO-government relations were tied to national contexts where governments were already supportive of CSO efforts to combat hate speech. Conversely, some Eastern European organisations noted that their national authorities not only neglected hate speech mitigation but adopted an actively antagonistic stance towards the organisations championing this cause.



GRAPH 2: CSOs' perspectives on the evolution of their relations with various actors due to the Code of Conduct. Source: Author's compilation from the 2019 and 2022 surveys.

Even interviewees from public authorities and companies indicated that the Code of Conduct had fostered stronger interactions between CSOs and digital companies by creating both the incentives and the means to enhance dialogue and collaboration between the two sides.

This is not coincidental, since one of the industry's commitments contained in the document consisted precisely in "intensify[ing] their work with CSOs to deliver best practice sharing on countering hateful rhetoric and prejudice" (European Commission, 2016, p. 3). However, the Code of Conduct appears to have produced effects beyond this instrumental task. As one of my interlocutors at the Commission

9. The Commission's periodic reports are presented before the "high-level group on combating racism, xenophobia and other forms of intolerance", chaired by the European Commission and including experts from each member state (as well as representatives from CSOs, EU agencies and intergovernmental organisations). However, this group is not involved in the concrete operation of the Code of Conduct.

stated in 2019, "the function of the monitoring mechanism is not only to test", but also to serve as a "trust-building process among all actors".

CSO representatives underscored this dimension as well, noting that following the Code's adoption "the platforms came and looked for us", establishing (or easing) a line of communication. In 2019, a CSO member highlighted that, while they had been struggling to have their "trusted reporter" status restored by Instagram, the new dynamic created by the Code of Conduct suddenly cleared the path to this goal. In a 2022 interview, other operatives described the same pattern occurring with TikTok: after joining the Code in 2020, the platform proactively reached out to the CSOs involved in the ME to grant them this status and its associated prerogatives.

Starting in 2017, the digital companies began inviting CSOs to yearly meetings at their European headquarters in Dublin. Initially, the organisations perceived these events with scepticism: in 2019, several interviewees portrayed them as "old school", "public relations" exercises, citing their one-sided nature ("they pontificate their guidelines on us") and the platforms' control over the agenda ("the important topics are left aside").

These negative views on the event did not prevent CSOs from recognising that the Code had enhanced their relationships with digital companies, as reflected in the survey results for both years. This suggests that the strengthened ties were a result of the Code's operation itself, rather than being driven by isolated, platform-orchestrated events.

After two editions online because of the pandemic, the Dublin meetings returned to an in-person format in May 2022. As my second survey was conducted prior to this, the responses were unaffected by the upcoming gathering. However, two of my interviews, held *after* the event, described it as "fundamentally different", noting that the companies demonstrated more eagerness "to listen" and to foster a relationship of "partnership". One interviewee viewed the introduction of brief one-on-one talks between CSOs and platform representatives as indicative of this new approach.

In November 2022, the Commission released a one-page annex to the Code, described as a "[j]oint statement by trusted flagger organisations and IT companies for an action framework on enhanced cooperation" (European Commission, 2022b, p. 1). While Twitter initiated the proposal and the Commission facilitated the negotiations with the industry, the CSOs had a say – though a limited one – during the

final stages of the process.

The introductory paragraph of the document emphasises how the Code had "forg[ed]" cooperation between CSOs and platforms. The substantive section then outlines four areas of action to further develop this relationship, including regular meetings to "explore" specific issues, joint efforts to better consider "national contexts", and steps to "increase the visibility of their [CSOs'] efforts" (European Commission, 2022b). It is worth noting that, while the Code itself mainly lists commitments from digital companies (with only two concerning the Commission), its annex fully incorporates the CSOs as parties to the supplementary agreement.

One year later, an online "advocacy roundtable" brought together the CSOs participating in the ME and the platforms to discuss the reporting process and the treatment of the notifications, with the aim of repeating the experience in 2024 (INACH, 2023c).

The Code of Conduct also fostered closer collaboration among CSOs focused on combating hate speech in Europe. The involvement of a growing number of organisations in the ME, encouraged by the Commission to broaden its geographic coverage, expanded the scope of coordination among CSOs. As interviewees explained, participating in this EU-wide initiative in direct contact with major platforms incentivised CSOs to share best practices, particularly given the wide diversity within this group in terms of experience, resources, practices and even core concerns (racism, antisemitism, islamophobia, gender discrimination, homophobia…). In a 2021 report, INACH described how the ME provided a rationale for delivering training to organisations beyond its own members:

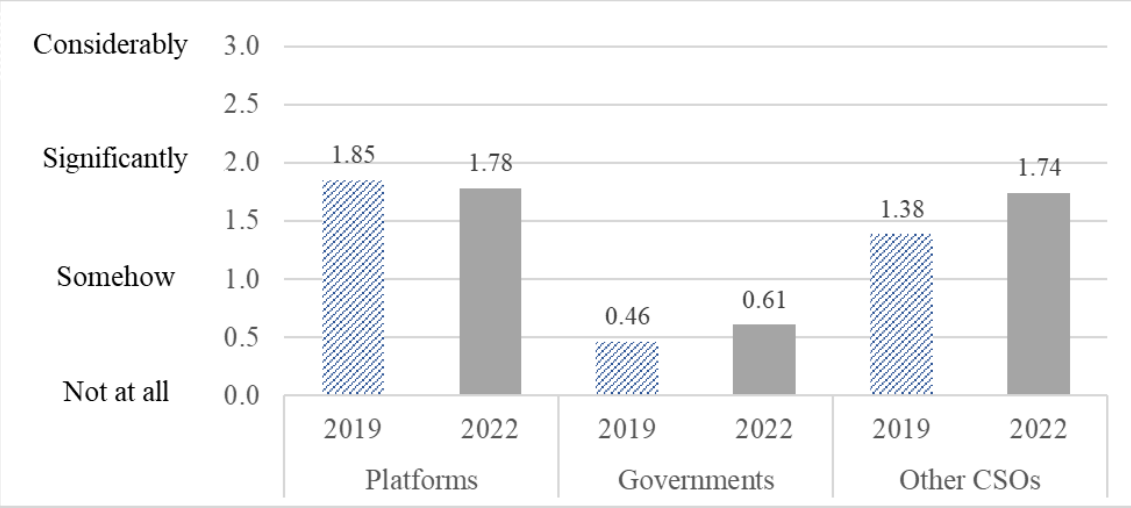
> INACH Secretariat and Licra, the two coordinating partners of the monitoring efforts, held two trainings before the main monitoring exercise. These trainings focused on the methodology of the ME, i.e., how long the monitoring period would be, how to record the data in our internal Excel sheets and the Commission's system, how to check the reported content and in what time increments these checks should be done. With these trainings we aimed to help NGOs that were new to this type of monitoring and also to make the monitoring process as uniform as possible, so the collected data would be as comparable as it could be (Feijoo & Berecz, 2021, pp. 4–5).

As noted in the same document, these ME-induced training sessions created an opportunity for participants to exchange views about the process: "Our participating members and partners also had valuable input on what could be changed in

the monitoring process and/or methodology to make the findings more representative or just even more in-depth" (Feijoo & Berecz, 2021, p. 5). These discussions enabled to point at the platforms' shortcomings as well: "Multiple partners mentioned the communication by the companies towards the reporters and their feedback given to complaints, or rather the lack thereof" (Feijoo & Berecz, 2021, p. 5). Consequently, a task primarily driven by practical considerations expanded into broader talks where opinions about the existing system were articulated and likely strengthened through the sharing of similar views and experiences.

Participating in the MEs also led to a growing institutionalisation of links among CSOs. As an INACH member explained in an email they sent to me in April 2019, "[The MEs] caused a boom in our membership. The [Code of Conduct] in itself would not have had these effects. The MEs were the events that brought in new members and strengthened the synergy and cooperation within the network" (personal communication, April 9th, 2019).

These initiatives and trends are reflected by the sharp increase in the perceived impact of the Code on intra-CSOs relationships between 2019 and 2022, as already illustrated by Graph 2. Presenting the answers collected from the earlier question[10]as average figures rather than as distribution series more effectively highlights that the Code's impact on CSOs' relations with their peers is now nearly on par with the impact on interactions with digital companies (see Graph 3).



GRAPH 3: Evolution of the CSOs' average perception of the development of their relations with other actors thanks to the operation of the Code of Conduct. Source: Author's calculations from the 2019 and 2022 surveys.

10. *Regarding the issue of hate speech online, how much has the Code of Conduct contributed to develop the relations between your organisation and: a) the platforms b) the government of the country you are established in c) other civil society organisations concerned by the same topic?*

Therefore, civil society organisations involved in the monitoring exercise can and indeed do simultaneously hold contrasted views on two crucial dimensions of the evaluation mechanism. Despite being critical of the outcomes in terms of reliable measurement of the platforms' behaviour, they acknowledge the value that the process itself generates in terms of interactions with both their peers and digital companies. These dual perceptions of the Code of Conduct' ME are powerful factors in explaining why CSOs were both willing and able to launch their own evaluation systems to measure the companies' actual commitment to tackling hate speech on their platforms on a day-to-day basis.

Section 3 – Alternative measurements by CSOs through their own monitoring exercises

Starting in 2018, various CSO coalitions initiated their own separate "silent" or "shadow" monitoring exercises (see for instance Berecz, 2019; OpCode, 2020a). While these adjectives implicitly acknowledge the centrality of the ME organised by the Commission, this section begins by recalling that certain CSOs had, in fact, pioneered this systematic and empirical evaluation of the platforms' actions towards hate speech *before* the Commission's own experiment.

3.1 Monitoring exercises prior to the one attached to the Code of Conduct

In 2016, the German Ministry of Justice and Consumer Protection formed a "Task Force" with YouTube, Facebook and Twitter to "immediately remove hate speech prohibited under German Law and provide user-friendly reporting mechanisms" (INACH, 2017, n.p.). The German CSO jugendschutz.net, an INACH co-founder, received federal funds to monitor the implementation of the platforms' commitments in three rounds: April-May 2016, July-August 2016 and January-February 2017. The results revealed disparate evolutions in the companies' handling of hate speech on their platforms (see also Gorwa, 2021, pp. 4–5).

Jugendschutz.net initially reported illegal content under the guise of regular users and, when necessary, as part of the organisation, making its *modus operandi* strikingly similar to the one applied, later, in the context of the Code of Conduct. Moreover, measurement focused on response time and removal rate, which were two of the four indicators that would subsequently be central to the Commission's MEs and annual reports (INACH, 2017). These resemblances draw a direct connection between this pioneering initiative and the Code of Conduct's ME, a parallel that was also spontaneously raised by several interviewees.

Another precursor to the Code of Conduct's monitoring mechanism was the "Research, Report, Remove" project, coordinated by INACH and funded by the European Commission under the Rights, Equality and Citizenship (REC) Programme. Six CSOs, including again jugendschutz.net, participated in six testing rounds between May 2016 and September 2017, which partially overlapped with the two first official MEs.[11] Its methodology was noticeably distinct from both the earlier initiative and the Commission's subsequent ME, especially in its inclusion of websites, forums and blogs in addition to platforms, as well as the use of several *sui generis* indicators (Berecz & Devinat, 2017b). However, the project's geographic coverage and its nature – featuring CSOs from different EU states – foreshadowed a defining characteristic of the mechanism the Commission set up by the end of 2016.

3.2 Alternative monitoring exercises, modelled from the Code of Conduct

In the years following the adoption of the Code of Conduct in 2016, different and often overlapping coalitions of CSOs launched their own monitoring exercises.

These shared some salient characteristics. First, they were supported by EU funds, delivered under financing programmes or tenders. Second, their promoters devised and promoted them with the explicit intention of enhancing the assessment of the digital companies' adherence to the Code and "unearth[ing] some issues that the official [exercises] could not" (Berecz, 2019, p. 5). Third, their methodology replicated that used for the official MEs and relied on the same four primary indicators: types of hate speech, removal rates, timely responsiveness and feedback provided to flaggers. Fourth, the abovementioned umbrella organisation INACH provided logistical support and/or coordination services for data collection.

Between 2018 and 2020, the French CSO Licra coordinated the EU-funded sCAN Project.[12] It included the implementation of two MEs of their own with the aim of "strengthening […] the monitoring exercises set up by the European Commission" (sCAN, 2020, p. 2). In addition to the nine sCAN members, a few other CSOs joined the initiative, further confirming the earlier point that the Code of Conduct has fostered cooperation among like-minded groups.

In their 2019 and 2020 annual reports, sCAN members presented indicators based on data they had gathered themselves, spanning two official MEs coordinated by

11. These were applied in November-December 2016 and May-June 2017.

12. Its full name is "Platforms, Experts, Tools: Specialised Cyber-Activists Network (2018-2020)" under the Project ID 785774.

the Commission (the fourth one in 2018 and the fifth in 2019) and two MEs conducted as part of the sCAN Project (one in 2019 and the other in 2020). They applied their own processing methodology to all the data, irrespective of its context of recollection.

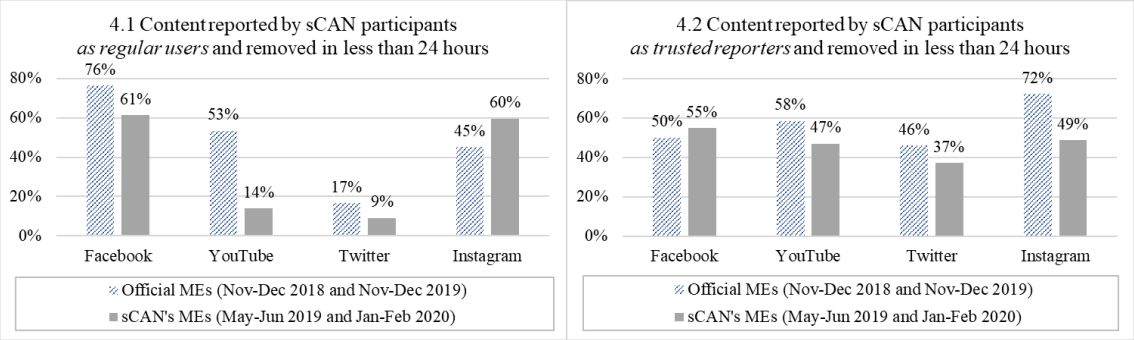
The sCAN methodology differed noticeably from that of the Commission. For instance, sCAN reports systematically broke down indicators into data obtained when participating CSOs posed as regular users versus when they openly acted as trusted reporters, a distinction the Commission reports applied only occasionally. Another telling discrepancy was that while the Commission focused on the percentage of reported content *assessed* within 24 hours, the sCAN Project indicated the proportion of content *removed* within this timeframe.

These divergences, that an interviewed leading actor in the sCAN Project explicated as "editorial choices", singularly undermine any attempt to compare the figures presented in these documents with the ones shown in the Commission's implementation reports. This state of affairs demonstrates that sCAN members did not seek to frontally question or disqualify the reliability of the Commission's reports by offering a comparable counterpoint.[13] In using their own metrics, sCAN participants rather aimed to point at specific shortcomings in the platforms' treatment of the notices, by fine tuning the presentation of their own results in a way that permits to highlight their central concerns.

Thus, the sCAN reports provide a useful way for comparing the platforms' timely responsiveness[14] as measured when sCAN members collected data in the context of the official MEs – where companies are suspected of being aware that they are underway – and in the context of the alternative ones, where firms are less likely to have such knowledge (see Graph 4).

13. Furthermore, sCAN members presented the figures from the official and alternative monitoring exercises in separate sections of its reports and did not include graphics comparing results from the different MEs. This approach confirms that sCAN participants did not intend to highlight the existing contrasts. I have designed **Graph 4** so to emphasize the differences between these results in a way that the sCAN reports do not.

14. "Timely platform responsiveness" refers to the proportion of flagged hate speech that the platforms addressed within 24 hours or less. The Code sets the goal of having a majority of reports being *reviewed* within this timeframe.



GRAPH 4: Reported contents removed within 24 hours by digital companies, in the official and the sCAN's monitoring exercises. Source: Author's calculations[15] from sCAN, 2019, 2020.

As shown in Graph 4, the sCAN figures support the CSOs' claims about the limited representativeness of the official MEs. Indeed, barring few exceptions,[16] the platforms' timely responsiveness is consistently higher, sometimes significantly so, during the official testing rounds compared to those conducted under sCAN.

In 2020-21, five CSOs from Estonia, Poland, Romania, Slovakia and Spain executed the OpCode Project.[17] They conducted three testing rounds in their respective countries to "verify the social media platforms' Code of Conduct compliance in various periods of time when IT companies are not scrutinised by the European Commissions' official Monitoring Exercises" (OpCode, 2020b, p. 2). Interestingly, their first cycle was synchronised with sCAN's second test, and the data collected were incorporated into the corresponding report as well (sCAN, 2020, p. 4), showcasing another instance of cooperation among different groups of CSOs.

Even though they relied on the same primary indicators as the other MEs, OpCode's three reports introduced variations to address their own concerns. Notably, they chose to present their data in a highly granular manner, breaking them down by platforms, types of hate speech, and country. This approach highlighted the significant disparity in outcomes depending on the national context: while platforms assessed 98% of reported content in Estonia (n=48) and Slovakia (n=122), this number was only 2% in Romania (n=100) (OpCode, 2020b, p. 7).

15. These percentages are the weighted averages of the data collected by sCAN participants in the context of the two official MEs, and in the context of the two MEs they applied independently as part of the sCAN Project.
16. Noticeably, in Graph 4, section 4.1, the results for Instagram deviate from the overall trend. This can be explained by the fact that the samples for this platform were noticeably smaller than those for the others – on the order of tens instead of hundreds.
17. Its full name is "Open Code for Hate-Free Communication" under the Project ID 850419.

Additionally, most of the data were expressed in absolute terms instead of percentages. This choice, along with the presentation of results primarily along national lines, complicates direct comparisons with the Commissions' figures. A contrasting exercise could even be considered pointless, given the small sample of countries covered in the OpCode MEs.

In their conclusions, the OpCode reports insist not only on the platforms' low responsiveness in general but also on the "disproportionate cross-national approaches regarding the hate speech phenomenon" (OpCode, 2020a, p. 2) and the effect of the pandemic-specific challenges (such as disinformation) in distracting the platforms from addressing hate speech content (OpCode, 2020b, p. 5). They also stress that, whenever possible, reporting as "trusted flaggers" very significantly increases the chances of achieving content removal (OpCode, 2021, p. 6).

Starting in 2021, Licra and INACH have been coordinating a "shadow monitoring exercise" involving rotating groups of five CSOs from different EU states and financed under a Commission's tender. In contrast to the other aforementioned CSO-produced reports, results from the shadow MEs were presented side-by-side with the numbers from the closest official ME, in graphs that facilitated comparisons (see Figure 1).

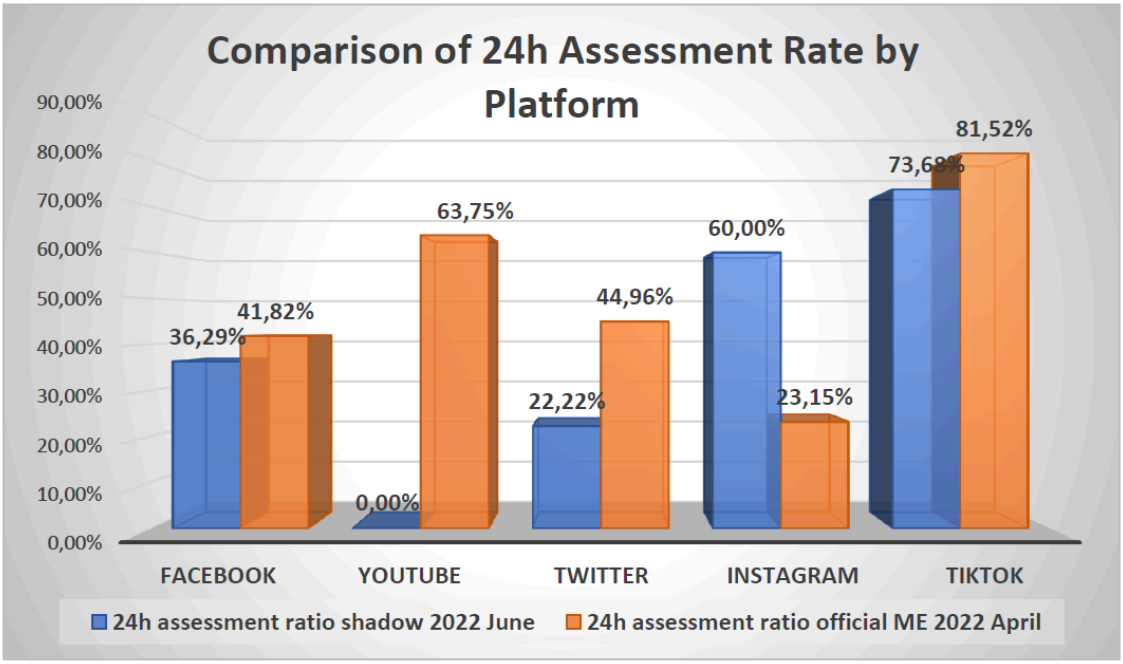


FIGURE 1: Graph from a Shadow Monitoring Report comparing results from two distinct MEs. Source: (Berecz et al., 2022, p. 4).

This enabled the CSOs to sharpen their criticisms regarding the differences in plat-

form behaviour during the official MEs and outside them.[18]For instance, one of these reports concluded that "[t]he findings of the first 2021 shadow monitoring clearly show the relevance and importance of continuing this type of monitoring activity. The companies did mostly worse in all indicators showed [sic] in this report" (Feijoo & Berecz, 2021, p. 7). These documents also highlighted the wide discrepancies in hate speech treatment depending on the country from which each case was submitted.

In January 2023, INACH and 21 CSOs launched the SafeNet Project,[19]supported by European funds for a two-year period. This initiative was noticeably more ambitious than its predecessors, not only due to the number of participants but also for its continuous nature, aiming to address the central flaw of the official system – namely, the ability of companies to adapt their efforts when a monitoring exercise is underway. The project covers five major platforms, adding TikTok to the four already covered in previous alternative testing mechanisms.

According to an interview with an INACH member actively involved in designing and implementing SafeNet, the Commission had been "strongly suggesting" the establishment of continuous monitoring and was "very pleased" that the project eventually materialised.

The intermediate results of the continuous ME are presented every two or three months as consolidated figures since the start of the programme. Each release includes a main document with aggregate figures from all the CSOs, alongside 19 other documents providing nationally-based statistics (INACH, 2024).

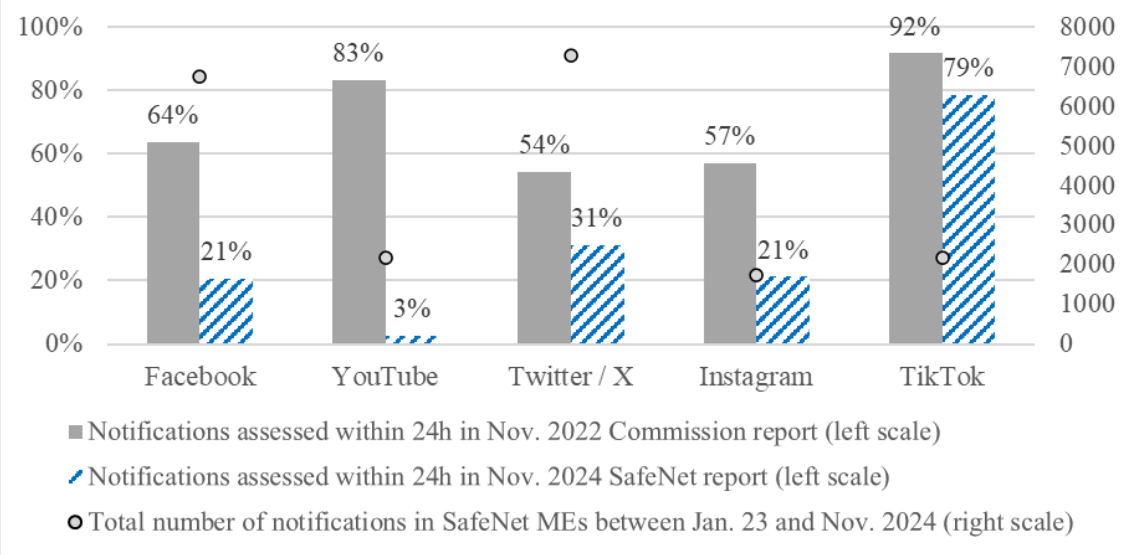
The reports, called "factsheets" like the Commission's own, are limited to four graphs – one for each key indicator. A fifth and final page is dedicated to a very short summary on the general trends observed. The figures are not broken down per regular user or trusted reporters, which simplifies comparisons with the official reports, although absolute values are provided instead of percentages.

The latest SafeNet report reveals that, during 2023-24, most digital companies have, on average, fallen significantly short of their commitments under the Code of Conduct. This trend is particularly striking given the sample size, which ranges

18. According to Figure 1, Instagram is the only platform performing better in the shadow monitoring exercises than in the official ones. The authors of the sCAN report suggest that "most likely it has more capacity in the countries that participated in the shadow exercise than globally within the whole of the EU" (Berecz et al., 2022, p. 5).

19. Its full name is "Monitoring and Reporting for Safer Online Environments" under the Project ID 101084457.

from over 1,700 to more than 7,200 cases per platform (INACH, 2023b) (see Graph 5).



GRAPH 5: Platform responsiveness to hate speech notices in SafeNet and Commission MEs. Source: (European Commission, 2022a, and author's calculations[20] from INACH, 2023b).

In summary, although reports from the four alternative MEs present data in varied formats, they consistently reflect the CSOs' expectations for platforms to fulfil their commitments on hate speech moderation, alongside their frustration over the significant room for improvement. Explicitly or not, these documents reinforce claims by CSOs that the figures produced by the Commission's MEs fail to accurately represent the digital companies' routine handling of such content.

In addition to developing these initiatives alongside the official monitoring exercises, CSOs have also in parallel formalised and systematised their involvement in the official MEs themselves.

Section 4 – A growing role for CSOs in the monitoring exercises, up to a certain point

Participating in official and unofficial MEs had a transformative impact on most CSOs, requiring them to channel energy and resources into tasks with which many of them were initially unfamiliar. For INACH, this involvement led to an increasingly prominent coordinating role, gradually taking on functions traditionally within the Commission's purview. In retrospect, certain developments in the earlier stages

20. The figures in the SafeNet reports had simply to be converted from absolute numbers to percentages.

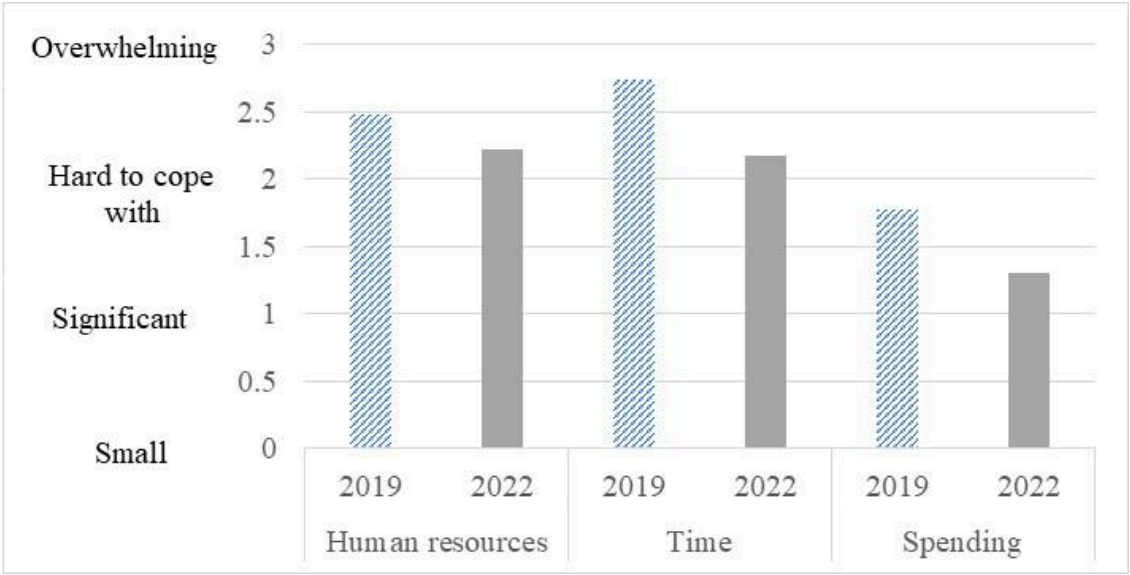
of the MEs can be seen as precursors to this evolution.

4.1 Novel and demanding additional functions for CSOs

CSOs have been experiencing a trend toward greater involvement in the MEs. For most, participating in these testing rounds was not a natural extension of their usual activities but rather a new and challenging endeavour. According to "Facing Facts", another network of CSOs,[21] adding this task to their agenda "led to a significant stretch in resources (time, human and financial)" and dented their capacity to continue executing their "core civil society activities such as capacity building and advocacy" (2022, p. 19). To assess how demanding the MEs are at several levels for the involved organisations, I included in my two surveys a question that asked:

Taking the normal functioning of your organisation as a reference, how would you evaluate the additional effort that is being required by your reporting to the European Commission of the answers (eventually) given by the platforms to your flagging of hate speech content, during the monitoring period, in terms of the following: a) Human resources b) Time c) Spending?

The collected responses confirmed that such involvement is costly at multiple levels, although the situation has reportedly slightly improved between 2019 and 2022 (see Graph 6).



GRAPH 6: Additional efforts required from CSOs for participating in MEs. Source: Author's calculations from surveys conducted in 2019 and 2022.

21. A high proportion of its 34 members are also INACH members and/or participated in the MEs (Facing Facts, 2024).

While initially participating in these exercises without economic compensation in return, some CSOs began receiving funding from 2018 through the aforementioned European projects. These funds supported the organisation of alternative MEs as well as their involvement in the official ones. While this financial backing alleviated a significant part of the economic burden, it also further entrenched their engagement in these tasks and institutionalised their role within the mechanism.

4.2 The early CSO involvement, stopping short of more ambitious plans

Unlike the development process of the Code of Conduct itself (Michalon, 2024, p. 109), there is scant information about how and by whom exactly the monitoring exercise was devised. It is known, however, that the Commission played a critical role in shaping the methodology of the exercise it introduced in October 2016, drawing on insights from the jugendschutz.net experience outlined earlier.

Interviews with key CSO members and two Commission officials revealed that some organisations contributed to defining the core features of the ME, such as the fact that content would be first reported as a regular user and then, if needed, through trusted reporter channels. INACH's input was especially significant, as it developed the Excel template that served as the technical basis for data collection in the first MEs, before being replaced by an EU survey tool. While the methodology's underlying logic has remained unchanged over the years, it has undergone gradual refinements, with participant CSOs reportedly providing feedback and input throughout the process.

Before validating the definitive ME methodology, the Commission had considered entrusting INACH with the tasks of collecting and processing data, as well as producing the reports following each monitoring exercise. However, the Commission ultimately reversed course and opted to retain direct control over most of these strategic dimensions, assigning INACH only the role of gathering data collected by its own members. According to my interlocutor from the umbrella group, this decision reflected the Commission's desire to maintain control over aspects critical to the overall success or failure of the Code of Conduct, especially in a context where trust among the parties still had to be built. Recent developments appear to confirm that this latter expectation has since then been met.

4.3 An ongoing transfer of responsibility, but tightly circumscribed to operating the ME

As described earlier, the Commission has been collaborating closely with INACH since the inception of the ME. This partnership was later formalised by a "Framework Partnership Agreement" covering 2018-2021, which was renewed for 2022-25 and led to the umbrella organisation gradually taking on more ME-related tasks.

There is no lack of evidence of INACH playing an increasingly important role. Some of these signs can seem to be confined to practical matters, such as jointly deciding the dates for the ME or serving as the designated intermediary between CSOs and platforms to address technical or other issues arising during the official monitoring exercises.

Other new practices touch upon fundamental traits of the ME. As a result of its increasing membership over the years, INACH is now coordinating more than half of the CSOs involved in the testing rounds. The call for participating in the 2023 ME was issued not by the Commission but by INACH itself, which would then select the applicants and allocate the available resources: "INACH will aim to evaluate the proposals by the 24th of March and will start notifying the chosen NGOs afterwards" (INACH, 2023a, p. 2). Consequently, the umbrella organisation was assuming a full-fledged decision-making role in a key operational dimension of the mechanism.

However, the Commission eventually called off the 2023 ME.[22] Instead, it invited INACH to organise a third shadow monitoring exercise in September-October, using the funds initially allocated for the official testing round. It repeated the same pattern in 2024. It remains to be seen whether this shift to CSO-managed MEs was merely a response to a contingent situation or foreshadows a lasting change. In any case, it suggests that the Commission is now considering delegating the task of monitoring the platforms' adherence to the Code of Conduct to the CSOs, either temporarily or permanently.

This approach is consistent with the trend observed in recent years, with the Commission progressively scaling down its involvement in the MEs. Its initial role as an intermediary between CSOs and digital companies in monitoring the Code of Con-

22. The Commission did not communicate publicly about this decision. Contacted in February 2024, my interlocutor at the EU institution explained it was due to the ongoing revision of the Code. On the same topic, a CSO operative indicated that the Commission, in a "diplomatic" move, did not want to add "extra pressure" on companies amid a negotiation process with them.

duct laid the groundwork for meaningful cooperation between these two sectors. According to several interviewed CSO members, these new conditions now allow the Commission to gradually hand over this burdensome coordinating function to civil society groups. The same persons also suggested that the entire operation of the official MEs, including the publication of the reports, could very well be left in the medium term to the "CSOs sphere", with a Commission's role limited to funding the system.

Nevertheless, this trend toward empowering CSOs in the monitoring area has not extended to the definition of the substance of the Code of Conduct itself. In March 2023, the Commission and the platform companies opened negotiations, which eventually ended by mid-2024, to revise the Code of Conduct in order to "improve its preventive capability" and align the document with the requirements of the Digital Services Act (DSA) (European Commission, 2023, p. 7).

Despite their repeated calls, CSOs have not been invited to participate in the process and have not been informed either about its development. According to CSO operatives, the Commission argued that this was already the case when the Code was first created. This factual remark, recalling a choice that had been criticised back then,[23]highlights that despite civil society organisations' increasingly relevant and effective role in the operation and management of the monitoring exercise, they were unable to gain a foothold in the negotiations that would define the new content of the Code of Conduct.

This restricted perimeter underscores the limited scope of the CSOs' role extension: their involvement is encouraged by the Commission and welcomed by companies as long as there is a functional motive to do so. In contrast, these same actors continue to set aside civil society members from the redefinition of the platforms' core commitments to combat hate speech online.

This exclusion runs counter Article 45 of the Digital Services Act (DSA), which enshrines codes of conduct into the governance of online activities and identifies a broad range of "relevant stakeholders", including CSOs, as qualified participants in the "drawing up" of such agreements. Since this legal disposition prompted the revision of the Code conducted between March 2023 and mid-2024, there was a compelling rationale to involve CSOs in updating its content.

The absence of CSOs from these discussions is paradoxical given that, according to

23. Access Now and EDRi, respectively a CSO and an umbrella group, had forcefully denounced this exclusive format (EDRi, 2016).

a Commission official, the aforementioned DSA provision on including civil society actors in drafting codes of conduct was itself inspired by the practice, pioneered in the context of the Code, of meaningfully engaging such actors (Klingvall, 2023b).

Conclusion

This research yielded findings at both empirical and theoretical levels.

On the empirical side, the study enhances our understanding of the operational dynamics of the mechanism adjunct to the Code of Conduct. First, I demonstrated that, despite their recurrent criticisms regarding the uneven implementation of the firms' commitments under the Code across time and national contexts, CSOs value the process for fostering closer relationships with both their peers and platform companies. Second, I outlined how different groups of CSOs have conducted their own monitoring exercises to support their claims and exert additional pressure on firms. I also showed that, by supporting these initiatives, the Commission has been consistent with its tendency to transfer more and more operational dimensions of the testing rounds to these expert and advocacy groups from civil society. While many of them were initially unfamiliar with such functions, this process has led to what Facing Facts (2022, p. 17) described as the "institutionalisation of the role of CSOs in monitoring the implementation of the Code of Conduct". Third, I raised the point that the exclusion of CSOs from the recent negotiations for a revised Code underscores the limits of this expansive logic, which remains confined to the specific domain of monitoring.

On a broader and more abstract level, the theoretical findings inform how the Code of Conduct and its monitoring mechanism fit into a co-regulatory model. During the design stage, only two types of participants – a public authority and a few private companies – were involved in negotiating the document. Therefore, recalling the distinction outlined in the first section of this paper between the two threads in defining co-regulation, the conditions in which the Code itself was crafted in the first place would only allow it to qualify as a co-regulatory endeavour under the "bilateral" perspective. Conversely, the development phase of the Code would not meet the threshold for co-regulation according to the "multilateral" approach, which requires the participation of a broader set of stakeholders.

Nevertheless, the Commission's decision to actively involve CSOs in monitoring the companies' adherence to their commitments allows to transcend any potential debate between advocates of the two views on whether the initiative constitutes a co-regulatory scheme, as it incorporated a third category of actors into the evalua-

tion mechanism. Moreover, both the formal responsibilities assigned to CSOs and their subsequent efforts to expand their role beyond these functions, as described in this paper, support the point that they had a meaningful contribution to the scheme, as opposed to a merely symbolic one. These developments strengthen the argument that the Code of Conduct, viewed holistically to include its monitoring exercise, satisfies the criteria for being of a co-regulatory nature.

That being said, this research offers nuanced conclusions regarding the scope of CSO involvement within the co-regulation of hate speech online in the EU. While their contribution to the implementation side is recognised, encouraged and supported, this does not extend to policy-making. In other words, CSOs are a key third party, yet their involvement remains confined to this status: a *third party* with a central role in the operational aspects of a governance system, without being blended into the decision-making processes that determine the system's overall shape and functioning.

Thus, this paper highlights both how a given system was developed along a co-regulatory logic, and how, within this logic, actors from civil society were able to expand their role, but only up to a certain point.

References

Alkiviadou, N. (2019). Hate speech on social media networks: Towards a regulatory framework? *Information & Communications Technology Law*, *28*(1), 19–35. https://doi.org/10.1080/13600834.2018.1494417

Aswad, E. (2016). The role of U.S. technology companies as enforcers of Europe's new internet hate speech ban. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2829175

Berecz, T. (2019). *The outliers: Addendum to INACH's joint monitoring report with the sCAN Project*. https://www.inach.net/wp-content/uploads/the_outliers_addendum_to_ME_report_2019.pdf

Berecz, T., & Devinat, C. (2017a). *INACH annual report 2016–2017*. https://www.inach.net/wp-content/uploads/INACH_Annual_Report_2016_FINAL.pdf

Berecz, T., & Devinat, C. (2017b). *Quarterly report on cyber hate (July – September 2017)* (p. 21). INACH. https://www.inach.net/wp-content/uploads/Quarterlyreport_July-September_2017.doc.pdf

Berecz, T., Feijoo, M., & Schoorl, A. (2022). *INACH final shadow monitoring report 2022*. INACH. https://www.inach.net/wp-content/uploads/ME-shadow-report-2022.pdf

Brannigan, A., & Zwerman, W. (2001). The real "Hawthorne effect". *Society*, *38*(2), 55–60. https://doi.org/10.1007/s12115-001-1041-6

Bukovská, B. (2019). *The European Commission's Code of Conduct for countering illegal hate speech*

online: An analysis of freedom of expression implications [Working Paper]. The Transatlantic Working Group. https://www.ivir.nl/publicaties/download/Bukovska.pdf

Cavaliere, P. (2019). Digital platforms and the rise of global regulation of hate speech. *Cambridge International Law Journal*, *8*(2), 282–304. https://doi.org/10.4337/cilj.2019.02.06

Coche, E. (2018). Privatised enforcement and the right to freedom of expression in a world confronted with terrorism propaganda online. *Internet Policy Review*, *7*(4). https://doi.org/10.14763/2018.4.1382

EDRi. (2016). *EDRi and Access Now withdraw from the EU Commission IT Forum discussions*. https://edri.org/edri-access-now-withdraw-eu-commission-forum-discussions/

European Commission. (2016). *Code of conduct on countering illegal hate speech online*. https://commission.europa.eu/document/download/551c44da-baae-4692-9e7d-52d20c04e0e2_en

European Commission. (2022). *Joint statement by trusted flagger organisations and IT companies for an action framework on enhanced cooperation – annex to the Code of Conduct*. https://commission.europa.eu/system/files/2022-12/Annex%20to%20the%20Code%20%E2%80%93%20Joint%20statement%20by%20IT%20companies%20and%20trusted%20flagger%20organisations%20to%20enhance%20cooperation.pdf

European Commission. (2023, December). *Joint communication to the European Parliament and the Council. No place for hate: A Europe united against hatred*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52023JC0051

Facing Facts. (2024). *Network*. Facing Facts. https://www.facingfacts.eu/members/

Feijoo, M., & Berecz, T. (2021). *Monitoring Report 2021*. INACH. https://www.inach.net/wp-content/uploads/First_FINAL_ME_Report_2021_FINAL-1.pdf

Finck, M. (2017). *Digital regulation: Designing a supranational legal framework for the platform economy* (No. 15; LSE Law, Society and Economy Working Papers). https://eprints.lse.ac.uk/87568/1/Finck_Digital%20Co-Regulation_Author.pdf

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gorwa, R. (2019). The platform governance triangle: Conceptualising the informal regulation of online content. *Internet Policy Review*, *8*(2). https://doi.org/10.14763/2019.2.1407

Gorwa, R. (2021). Elections, institutions, and the regulatory politics of platform governance: The case of the German NetzDG. *Telecommunications Policy*, *45*(6), 102145. https://doi.org/10.1016/j.telpol.2021.102145

Hirsch, D. (2011). The law and policy of online privacy: Regulation, self-regulation, or co-regulation? *Seattle University Law Review*, *34*(2), 439–480.

Horowitz, M. A. (2024). A balancing act: EU media regulation, co-regulation and self-regulation in the digital age. *Media and Journalism Research Center*. https://journalismresearch.org/2024/05/a-balancing-act-eu-media-regulation-co-regulation-and-self-regulation-in-the-digital-age/

INACH. (2017). *Deletion of hate speech on Facebook, YouTube and Twitter*. https://www.inach.net/deletion-of-hate-speech-on-facebook-youtube-and-twitter/

INACH. (2023a). *Call for proposals from NGOs to get subcontracted for the 8th monitoring exercise to*

monitor the Code of Conduct on tackling online hate speech. https://www.inach.net/wp-content/uploads/Call-for-Proposals-from-NGOs-to-get-Subcontracted-for-the-8th-Monitoring-Exercise-to-Monitor-the-Code-of-Conduct-to-Tackle-Online-Hate-Speech.pdf

INACH. (2023b). *Fact sheet* (No. 10). https://www.inach.net/wp-content/uploads/Fact-sheet-No.-10.pdf

INACH. (2023c). *First advocacy roundtable organized for the SafeNet project*. https://www.inach.net/first-advocacy-roundtable-organized-for-the-safenet-project/

INACH. (2024a). *SafeNet fact sheets*. https://www.inach.net/safenet-fact-sheets/

INACH. (2024b). *Podcast with Louisa Kingvall, policy officer at the European Commission*. https://twitter.com/INACHnet/status/1749761601324953909?s=20

Jourová. (2019). *Code of Conduct on countering illegal hate speech online – fourth evaluation confirms self-regulation works*. European Commision. https://commission.europa.eu/document/download/8fd53d79-46c3-442c-8ea3-c43688147afc_en?filename=code_of_conduct_factsheet_7_web.pdf

Kleinsteuber, H. J. (2004). The internet between regulation and governance. In OSCE (Ed.), *Self-regulation, co-regulation, state regulation*. https://www.osce.org/files/f/documents/2/a/13844.pdf

Klingvall, L. (2023). *Podcast with Louisa Kingvall, policy officer at the European Commission*. https://x.com/INACHnet/status/1752304615494668352?s=20

Kuczerawy, A. (2017). *The Code of Conduct on online hate speech: An example of state interference by proxy?* https://www.law.kuleuven.be/citip/blog/the-code-of-conduct-on-online-hate-speech-an-example-of-state-interference-by-proxy/

Marsden, C. T. (2004). Co- and self-regulation in European media and internet sectors: The results of Oxford University's study. In OSCE (Ed.), *Self-regulation, co-regulation, state regulation*. https://www.osce.org/files/f/documents/2/a/13844.pdf

Michalon, B. (2024). *Framing, shaping, and maneuvering: Exercising power in building online content regulation*. https://theses.hal.science/tel-04952440

OpCode. (2020a). *Monitoring and reporting illegal hate speech* (No. 1; Shadow Monitoring Report). https://www.inach.net/wp-content/uploads/Report_OpCode.pdf

OpCode. (2020b). *Monitoring and reporting illegal hate speech* (No. 2; Shadow Monitoring Report). https://www.nigdywiecej.org//docstation/com_docstation/172/monitoring_and_reporting_illegal_hate_speech._shadow_monitoring_report__2nd_e.pdf

OpCode. (2021). *Monitoring and reporting illegal hate speech* (No. 3; Shadow Monitoring Report). https://www.nigdywiecej.org/en/projects/project-reports/4787-%E2%80%9Cmonitoring-and-reporting-illegal-hate-speech-shadow-monitoring-report-third-edition%E2%80%9D

Quintel, T., & Ullrich, C. (2020). Self-regulation of fundamental rights? The EU Code of Conduct on hate speech, related initiatives and beyond. In B. Petkova & T. Ojanen (Eds.), *Fundamental rights protection online*. Edward Elgar Publishing. https://china.elgaronline.com/view/edcoll/9781788976671/9781788976671.00019.xml

Reynders, D. (2021). *6th evaluation of the Code of Conduct*. European Commision. https://commission.europa.eu/document/download/ff9f23ab-1846-4d4e-a2fd-3e6a7fa56860_en?filename=factsheet-6th-monitoring-round-of-the-code-of-conduct_october2021_en.pdf

Reynders, D. (2022). *7th evaluation of the Code of Conduct*. European Commision. https://commission.europa.eu/document/download/5dcc2a40-785d-43f0-b806-f065386395de_en

Rubinstein, I. S. (2018). The future of self-regulation Is co-regulation. In E. Selinger, J. Polonetsky, & O. Tene (Eds.), *The Cambridge handbook of consumer privacy* (1st ed., pp. 503–523). Cambridge University Press. https://doi.org/10.1017/9781316831960.028

sCAN. (2019). *Monitoring report 2018–2019*. http://scan-project.eu/wp-content/uploads/sCAN_monitoring_report_year_1.pdf

sCAN. (2020). *Monitoring report 2019–2020*. https://scan-project.eu/wp-content/uploads/sCAN_monitoring_report2_final.pdf

Steurer, R. (2013). Disentangling governance: A synoptic view of regulation by government, business and civil society. *Policy Sciences*, *46*(4), 387–410. https://doi.org/10.1007/s11077-013-9177-y

Yasuda, J. (2016). Regulatory governance. In *Handbook on theories of governance* (pp. 472–484). Edward Elgar Publishing. https://www.elgaronline.com/display/edcoll/9781800371965/9781800371965.00051.xml

Zhesko, M., & Heller, D. (2022). *Current activities and gaps in hate speech responses. A mapping report for the facing facts network*. Facing Facts Network. https://www.facingfacts.eu/wp-content/uploads/sites/4/2023/04/Facing-Facts-Network-Mapping-Report-v8.pdf

Published by

Alexander von Humboldt Institute for Internet and Society

CAIS Research for the Digital Age

in cooperation with

CREATe

cnrs centre internet et societe

UOC R&I IN3 Internet interdisciplinary Institute Universitat Oberta de Catalunya

University of Tartu Johan Skytte Institute of Political Studies