Barthélémy Michalon – Sciences Po

Power spotted through its outcomes: understanding the relative contributions of different stakeholders in the shaping of a governance scheme for illegal hate speech online

Abstract

The Code of Conduct against illegal hate speech online is an original scheme to regulate this phenomenon at the level of the EU. In addition to the written document itself, it also includes a "monitoring exercise" to assess its implementation. Beyond its effects on the matter, it is also a useful object to reveal the forms and dimensions of the power relationships that take place between several categories of stakeholders: the European Commission, (US-based) digital platforms and a growing number of civil society organizations.

Drawing from firsthand research on the field and primary sources, this paper analyses specific features of this scheme – from wording to data-collecting procedures – and interprets them as markers of the relative power capacity of each of the abovementioned stakeholders. This allows a more accurate depiction of how roles and powers are exercised in the incipient governance of illegal hateful contents online at the EU level.

Keywords

Power relationships
Code of Conduct
Hate Speech
(Self) Regulation
Governance
European Commission
Digital platforms
Civil Society Organization
Monitoring
Implementation reports

Introduction

In May 2016, four digital companies agreed on a “Code of Conduct against illegal hate speech online”, which is expected to set some ground rules on the regulation of such content within the European Union. Although they were the only signatories, they were not alone in the drafting of the document: the European Commission enticed them to work on it and pushed for adjustments in some key areas of the text. Later, the European Commission (EC) added a complementary mechanism to the agreement: a “monitoring exercise”, through which selected civil society

organizations (CSOs) from different EU member states would test whether the companies were fulfilling their commitments, and transmit the data to her so she could publish implementation reports on a regular basis.

Therefore, three main categories of players are engaged in the same game: the European Commission, digital platforms and CSOs specialized in the defense of certain rights in the digital realm. What is at stake there is the regulation of a certain kind of online contents: those which are close to the "hate speech" red line, or beyond it.

Apparently, preventing hateful speech from spreading online is a consensual objective shared across these three classes of relevant players. Beyond this superficial observation, it should be kept in mind that the participation of each actor in this initiative (i.e. the agreed text itself and its annexed "monitoring exercise") is driven by their own respective concerns and objectives. Additionally, each of them has a different set of resources to deal with the other players and achieve their goals.

My paper tries to answer the following research question: to what extent was each category of actors able to strengthen its relative power vis-à-vis the other players in both the framing and implementation processes of the Code of Conduct, and to broaden its scope of action beyond what had been formally established in the first place? Conversely, what shared overarching considerations acted as a restraint and prevented each of them from expanding their own power beyond a certain point?

After a presentation of the methodology applied in this paper and a brief statement on the chosen theoretical framework, the development of my argument is organized around two major phases: first, the drafting of the document itself, followed by the implementation phase of the mechanism it creates. In each phase, relevant signals are detected, analyzed and interpreted to draw inferences about the relative power of the actors involved.

Methodology

Because of the sources it relies on, my research is essentially of an empirical nature. It is mainly based on primary sources, which consist in written correspondence between relevant actors accessible online and in interviews I personally conducted with some of their representatives. Some secondary sources, in the form of official reports published by the Commission, were also mobilized.

Firstly, I extracted relevant information from dozens of emails that circulated between the Commission and the digital platforms during the writing process of the Code of Conduct from March to May 2016. These exchanges were made public thanks to a request successfully lodged by EDRi[1], a CSO which denounced that this conversation was taking place behind closed doors, leaving the civil society representatives aside. By the end of April 2016, it requested access to this

[1] European Digital Rights

information in order to overcome the opacity in which the process was developing. Such access was eventually granted in July 2016: these documents were not only delivered to the requesting party but, in accordance with the applicable transparency procedures in the European Union, were made public by posting them online on the EU dedicated website. They are organized in six "annexes", depending on the institutional actor who emitted the emails.[2] Throughout this paper, the information taken from these electronic exchanges will be signaled as such, with an indication of which annex it comes from.

The exploitation of the information contained in these 374 pages took as a starting point the identification of the positions defended by each party involved in the drafting process, in this case the European Commission and the three companies from the IT sector. Two specific documents allowed to have a clear vision of the key interests that both sides were pushing forward at the very beginning of the whole process. Those respective aims were then contrasted with the content of the first jointly written document and the subsequent updated versions of it, and finally with the substance of the definitive version of the Code of Conduct. In the analytic process, I identified a series of major stakes, which allowed me to track the evolution of the provisions related to every one of them in an organized fashion. The next step consisted in qualitatively assessing to which extent the final fate given to each stake was meeting the expectations of each side, as an indicator of its ability to transform its power capacities into tangible outcomes favorable to its interests.

However, the analysis of this dense paper trail was challenging for several reasons. First of all, it appears that not all the relevant documents were provided. The request was broadly formulated, by demanding access to "all drafts and documents in relation to this code of conduct against hate speech", "the exchange of e-mails with industry and the Commission with regard to the preparation of the code of conduct" and "the agenda and minutes of meetings with industry in relation to this code of conduct". This detailed wish list was only partially satisfied: not all the successive draft versions of the Code were included, and no formal minutes of meetings were provided. As to the emails that went back and forth between those five institutional actors (the European Commission and the four IT companies), some of them were apparently missing and the documents supposedly attached to them were not systematically included. Additionally, most of them were related to logistical details, and therefore of limited use for my research. On top of that, the names were almost always redacted (except in a few cases, probably involuntarily overlooked in the process), as well as other sections, in particular a three-page email that was fully hidden behind black bars, including all its metadata. However, the inconsistency with which the redaction was applied across the full range of emails brought some relevant insights: in some cases, a given email was (partially) redacted in a certain thread but not (or not in the same fashion) in another thread, which indicates which topics were considered as somehow sensitive, at least from certain perspectives. On a more practical note, the emails were not consistently arranged in a chronological order and the search through keywords was impossible.

[2] They are all available at https://www.asktheeu.org/en/request/code_of_conduct_against_hate_spe

Despite these obstacles, the close study of these exchanges revealed in what spirit and with which concerns and priorities the IT companies and the European Commission officials from DG Justice and Consumers took part in the development process of what would eventually turn into the final text. In particular, five draft versions of the Code of Conduct, where the edits and comments were visible, proved to be quite useful, and constituted the main source of information for this part of my research.

When it came to interpret this messy and incomplete raw material, I was fortunately able to mobilize the information I directly collected through a series of interviews that I had earlier with a dozen relevant participants in the Code of Conduct, as part of a field research conducted mostly in Brussels and Paris, and complemented through remote conversations, from January to April 2019. This firsthand information provided me with context that turned out to be valuable to make sense of this paper trail.

Additionally, in order to have a broader view of the perspectives that prevail within the civil society organizations involved in the monitoring of the implementation of the Code, I sent an online survey to 30 such actors and received answers from 27 of them, from 21 different countries, between April and July 2019. The aggregation and the processing of the answers provided me with a valuable outlook at their own assessment of the effectiveness of this whole mechanism in curbing the presence of illegal hate speech online.

Finally, some secondary sources were also consulted, specifically all four implementation reports that were published between December 2016 and February 2019. The figures they contain, as well as the information they do not disclose, are useful indicators of the outputs generated by the mechanism as a whole. More importantly for the purpose of this research, they can also be used to identify which actors succeeded in having their expectations and interests taken into account in the shaping and design of this public document.

My analysis centered on three key steps in the process: the writing and adoption of the Code of Conduct, the implementation of the monitoring exercise and the periodic reporting of the results. Such a breaking-down allowed me to identify in which phase each player had the greatest opportunity to curb the process in the direction more akin to its interests. Nevertheless, as exposed above, due attention was also paid to the reasons that, at each stage, impose a constraint on the exercise by each player of the full range of their potential powers.

Except in a few cases, this research regarded this group of private actors as a unified player: this treatment should not be interpreted as a claim that their interests were aligned, but as a methodological choice and necessity. First, it’s the result of a choice because, as stated before, my aim is to analyze the power relations between *categories* of actors, not between each of them taken individually. Second, it is also a pragmatic recognition of the need to adjust the scope of this research to the format

of an article, and an adaption to the fact that the disclosed emails were not including intra-industry communications.

Theoretical framework

The classical debate opposes power as a set of resources that are possessed by certain actors, with power as the capacity to move the behavior and decisions of others in a certain direction. My paper resolutely draws from the second conception: power exists as long as it is able to produce outcomes, in a way that is consistent with what the power holder intended.

The case of the Code of Conduct provides a useful opportunity to spot such outcomes produced by the power interaction that takes place between different categories of players and, to a lesser extent, within each category.

In doing so, I am choosing an angle that is similar to DeNardis and Hackl in their 2015 article "Internet governance *by* social media platforms", in which the authors stress the regulative power that digital companies exercise by pointing at specific rules and affordances (i.e. specific outcomes) that they respectively apply and allow in their respective platforms.

Discussion of the findings

As exposed earlier, this paper aims at highlighting the relative power of all three categories of actors involved throughout three decisive stages, starting with the drafting process of the Code of Conduct, continuing with the application of the system that serves to monitor its implementation and ending with the periodic presentation of the results. Unsurprisingly, the structure of this section follows the same logic.

1. *Drafting the Code of Conduct: A compromise between the Commission and the IT sector*

At first sight, the terrorist attacks that were perpetrated in Brussels, on March 22nd, appear to have triggered the adoption process of a set of rules to tackle hate speech, since no more than two days later, the Justice and Home Affairs Council instructed the Commission to "develop by June 2016 a code of conduct against hate speech online", through an intensified cooperation with the IT sector[3].

However, by then, the consultations between the European Commission and IT companies on the subject had already been under way for several months, according to two European Commission officials I talked with. The abstract notion of a set of principles to be applied on potentially dangerous content online was initially discussed in October 2015, in the broader context of the first colloquium on

[3] https://www.consilium.europa.eu/en/press/press-releases/2016/03/24/statement-on-terrorist-attacks-in-brussels-on-22-march/

Fundamental Rights that the European Commission organized that year, this time around the topic of the fight against antisemitism and islamophobia. During the event, participants from different sectors coincided, at least in principle, on the need to have a coordinated action in this area. Afterwards, several meetings took place on a regular basis between representatives from the firms and the Commission to address this specific subject. The creation of the EU Internet Forum in December 2015, gathering Member States, Europol and IT companies "to counter terrorist content and hate speech online", was a way to formalize cooperation is this field[4].

1.1. The selection of the players: the choice of a restricted field

On March 4th, 2016, a first "Coordination Meeting with Member States, IT companies and Civil Society in the context of the Dialogue with IT Companies on online hate speech" was organized. It served to formally launch the drafting process of the Code of Conduct[5] and, as its long name itself indicates, it involved a variety of actors.

However, as the name also reveals, the "dialogue" part, that would last for the next three months, implicated a much narrower club of players: the Commission, of course, and the IT sector, in this case limited to Facebook, Google and Twitter. As revealed by one of the emails threads, Microsoft was initially very reluctant to get involved: it decided to observe the whole process from the outside, and to formally join it only at the very end, without contributing to the discussions on its content.[6]

As to the CSOs, they were not invited to participate in any part of the drafting process. Two of them, EDRi and Access Now, denounced that "civil society was systematically excluded from the negotiations that led to the voluntary 'code of conduct' for IT companies – an official document that was presented today, despite the lack of transparency and public input into its content".[7] In order to give more weight to their protest, both organizations announced, the very same day the code was made public, that they would not participate anymore to the discussions within the EU Internet Forum, where they were sometimes invited on a case-by-case basis.

My own conversations with several CSO representatives confirmed that none of them were included either in the talks or in the electronic exchanges through which the Code of Conduct was developed. However, in one of the analyzed emails, a Facebook representative declared that "we are also getting input from NGOs but naturally this is taking some time".[8] This passing reference shows that, if any involvement of CSOs ever happened at some point, it was only with a very limited number of them and on the sidelines of the existing formal process.

---

[4] http://europa.eu/rapid/press-release_IP-15-6243_en.htm

[5] When formulating its request, EDRi did not specify any timeframe, but only to get access to the documents associated to the elaboration of the Code, so it is significant that the oldest correspondence delivered as a consequence of the request is precisely related to the meeting on March 4th.

[6] Information taken from the whole documentation available in Annex B.

[7] https://edri.org/edri-access-now-withdraw-eu-commission-forum-discussions/

[8] Information taken from Annex D, page 3.

The absence of the CSOs resulted from a deliberate choice from the Commission, which intended to characterize the document as an intra-industry agreement: In the interviews I had with officials, they made clear that they considered the text as having a "self-regulatory" nature, which is also confirmed by the wording of public statements on the matter.[9] However, given the place that the CSOs are recognized within the text itself,[10] and given the role granted to them by the code of conduct in the monitoring phase of the implementation of the agreed principles, it was not self-evident that they had to be wholly left outside of the drafting phase.

In fact, these organizations were invited to the second "Coordination Meeting with Member States, IT companies and Civil Society in the context of the Dialogue with IT Companies on online hate speech" in which, on May 31st, 2016, the existence and the content of the code of conduct against illegal hate speech online were made public[11]. While the title makes clear that only the industry participates in the dialogue with the Commission on the subject, one is left to wonder what can be "coordinated" with the other players, if their role is limited to be physically present in the couple of events that opened and closed the process.

1.2. The Commission's "set of expectations": four key points

The March, 4th 2016 "Coordination Meeting with Member States, IT companies and Civil Society in the context of the Dialogue with IT Companies on online hate speech" was described as a "discussion" on three "concrete points".

- "Firstly, how to reach a common understanding that IT companies beyond applying their own terms of services should commit, upon notification, to restrict or take down any [hate speech] content […], in accordance with EU and national law;
- Secondly, how to ensure that responsiveness to notices is prompt and equal and which target benchmark for "time to review and take-down" can be considered appropriate;
- Finally, how all participants can contribute to identify and support a representative network of civil society partners and trusted flaggers to help provide high quality removal referrals for hate speech content". [12]

As the public record shows, all three points were introduced in terms of "how to…". This formulation was suggesting that, by then, their eventual shape was supposed to be left largely open-ended.

---

[9] See for example the subtitle of the fourth implementation report (February 2019): "Fourth evaluation confirms self-regulation works".
https://ec.europa.eu/info/sites/info/files/hatespeech_infographic3_web.pdf

[10] They are explicitly mentioned 8 times in the final version of the document (as a comparison, "Commission" appears 11 times and "Member States" 10 times).

[11] Information taken from Annex E.2, p.37 or 51.

[12] https://ec.europa.eu/info/sites/info/files/press_hate_speech_meeting_04_03_2016_en.pdf

However, the emails that were made public thanks to the EDRi request provide a different account of these "concrete points": in the electronic exchanges that took place afterwards solely between the EC and the IT industry, these aspects were not anymore presented in a quasi-interrogative form, as if they were still waiting to be shaped by future discussions. On the contrary, in an email sent on March 13th,[13] the Commission formulated them in a much more specific and affirmative fashion. In this four-paragraph text, the Brussels institution was, in fact, listing the key features that, she expected, would shape the substance and logic of the future Code of Conduct:

a) IT companies should, *in addition to* IT companies' internal rules, assess notices of hate speech "against the legal standard set by EU law"
b) IT companies should commit to provide a "prompt" response to notices, that shall also be "equal throughout the EU": the review of notices and, where applicable, the removal of illegal contents shall be effective within 24 hours.
c) The communication between IT companies and Member States authorities should be improved, through swifter contacts between their respective "national contact points" and a standardization of notification forms.
d) CSOs should be involved "in all member states to help provide high quality removal referrals for hate speech contents". Additionally, the IT companies are expected to make public the criteria that such CSOs must meet in order to be considered as "trusted flaggers".[14]

The four points listed in this short document circulated by the Commission (which, let's remember, had no vocation to be made public) can be understood as specific answers given to different stakes, which can be classified as follows:

a) Normative basis for content assessment and removal
b) Deadline to assess notices and remove contents
c) Communication between Member states' and companies' national contact points
d) Involvement of CSOs

By then, the Commission appeared to be quite confident that those points, and its answers to the, would constitute the main substance of the whole Code of Conduct. Indeed, in an email sent on April 8th,[15] a policy officer from DG Justice referred to these four paragraphs as "our text" and invited the companies to share their drafting suggestions, "if any". This person aired a "*possible* need for a meeting" in the next ten days, in order to "*finalise* the common version of the commitments".[16] Although s/he admitted that the position from the IT sector had not made be known yet, the policy officer did not discard the possibility that "it could be settled by a simple exchange by mail" and expressed hope that it could be completed "quickly". All those elements of language reveal the assumption that the input from the industry would bring nothing but marginal changes. Therefore, it appears quite clearly that, in the

[13] Information taken from Annex B, page 2.
[14] Information taken from Annex B, p. 2.
[15] Information taken from Annex B, p. 7.
[16] In both cases, the emphasis is ours.

early stages of the drafting process, the EC presumed that the eventual document would be quite short and straightforward, and would mostly be its own creation.

In a conversation we had at the beginning of 2019, a representative from one of the companies referred to the fact the Commission had presented a "set of expectations" at the very start of the whole process, made of the key elements that should be included in the future document. Although he declined to be more specific, this designation was certainly referring to the four-paragraph document described above and was, therefore, quite revealing of the different approaches that both sides had about this initial text.

1.3. The first draft from the IT sector: A significantly different approach

On April 12th, a Facebook representative sent "a draft for further discussion" to the Commission, explicitly in the name of all three companies (the subject of the email was "Code of Conduct – Google/Twitter/Facebook input").[17]

This document was short (slightly more than one page) but highly valuable for analytical purposes, since it allows to spot the initial positions held by the IT sector at the beginning of the process. It was subtitled "suggested wording for legally non-binding political commitments", which highlighted not only its very preliminary status but also, and more importantly, the companies' insistence on their perception of the whole initiative as being essentially voluntary.

The upcoming analysis of the main elements contained in this document does not follow the chronological order in which they are presented. Instead, it is organized around the four stakes that were previously extracted from the EC's "set of expectations". The fact that it is possible to organize the content of both documents along the same lines shows that there was a consensus between the two groups of actors on the major stakes that the document should address. However, to structure the information this way also allows to highlight that both sectors had very different views on how to deal with each individual stake:

a) Normative basis for content assessment and removal

The notices would have to be reviewed "against [the companies'] terms and conditions", thereby ignoring the EC's expectations to *also* focus on national laws.

This company-centered approach is confirmed by the point that comes in the first place in their list of commitments: to include, within their own "terms and conditions", a specific section devoted to the prohibition of "content that expressly promotes or threatens violence". This initiative reflects the IT sector's intention to deal with hate speech from their own rules.

b) Deadline to assess notices and remove contents

[17] Information taken from Annex D, pp. 3-4.

Notices of hate speech should be reviewed within 48 hours, which was twice as much as the maximum amount of time set by the Commission. No time-bound objective in terms of effective removal was set.

A specific procedure was to apply for notices coming from law enforcement authorities and "trusted flaggers": only in these cases would the reported contents be analyzed against national law and the removals be effective "without delay". This part can be seen as a concession to the Commission's "set of expectations" but of a limited kind, since it would apply depending on who is producing the notice. Even if it was far from meeting the EC's expectations, it was explicitly signaled as tentative[18], which is revealing of the IT industry's reluctance to make such a step.

c) Communication between Member states' and companies' national contact points

The paragraph on this specific stake was visibly built from the Commission's specific demand on this point,[19] but a significant change was introduced: the improvement of communication was to be achieved through "education on the procedure for making reports", an effort that rests on the side of the national authorities. Nothing was said about the standardization of notification forms.

d) Involvement of CSOs

The fact that certain CSOs would be recognized as "trusted reporters" was acknowledged but there was no provision on making more transparent the process through which some of them would get this status, contrary to the Commission's expectations.

Moreover, a cooperation between the IT sector and CSOs was envisioned: civil society would provide insights on how to deal with hate speech, while the industry would help these actors increase the reach of their own campaigns.

Interestingly, the IT companies included a paragraph on improving, through "support and training", the CSOs knowledge and understanding of the "individual company Guidelines and Rules". This specific part echoes the intention, highlighted above, to train the national authorities in the reporting activity, along the companies' lines. In both cases, the digital firms' intention to structure the whole Code of Conduct around their own normative references appears quite clearly.

[18] The specific format applied to this specific paragraph (in italics and within brackets) as well as the straightforward mention "to be assessed by companies" left no doubt about it.

[19] Both the EC and the IT companies' paragraphs start with: "Member States and IT companies should designate and inform each other of their respective national contact points", but their views sharply diverge from there. Information taken from Annex B, p.2 and Annex D, pp. 3-4.

Therefore, the first document produced by the IT sector acknowledges the existence of the four stakes laid down by the Commission, but it provides, for each of them, an answer that directly contradicts the "set of expectations" presented one month earlier by the Commission. This fact indicates that the IT sector did not feel bound by what the European institution exposed in the March 4th meeting: instead of taking it as a starting point, it wanted to set its own baseline.

Additionally, the IT sector's first draft introduced elements unrelated to what was listed in the Commission's "set of expectations". This is quite logical, given the short size of the former document. Those new items are useful in our analysis in that they can be construed as markers of the companies' own interests and priorities in the negotiation process.

The companies' document opens in a particularly revealing way, by stressing, first and foremost, the importance of freedom of expression, that they "share a collective responsibility and pride in promoting and facilitating". They continue by citing a European Court of Human Rights decision that includes within the field of application of this freedom the expression of messages "that offend, shock, or disturb the State or any sector of the population".[20]

This emphasis is quite telling of how the three digital companies approached the issue of hate speech regulation on their platforms. Instead of setting hate speech elimination as their goal, and then recalling that this should take place in a way that is consistent with the right to freedom of expression, they have decided to first reassert the centrality of the principle of freedom of expression: The fight against hate speech would have to find for itself a place against this backdrop. The companies were therefore trying to prevent this upcoming effort against hate speech from having a broad impact on the functioning of their platforms. Freedom of expression, on which their very existence depends, should remain the rule.

Another element introduced by the digital firms' first draft was a series of measures that would be implemented out of their own initiative. Within the IT sector, staff would be trained on "current societal developments" and cooperation among the companies would be fostered. On the last point, a more specific provision was added later in the text: to "share best practices" on the inclusion of anti-hate speech provisions in their respective "terms and conditions". Finally, the tech sector also added commitments towards their users, by promoting "counter narratives" and "critical thinking", as developments that would be effective against hate speech. This is consistent with the position often defended by the tech sector which, framing its own role as an intermediary, insists on the need to develop skills in the community of its users.

Therefore, the list of stakes that was started above can be complemented as follows:

e) Hate speech and freedom of expression
f) Initiatives from the IT sector

[20] Information taken from Annex D, pp. 3-4.

The width of the gap between the Commission's set of expectations and the IT sector's first draft can be seen as the result of a negotiation technique: the companies may have set their own proposals far away from those from their counterpart with the aim to, finally, strike an agreement on some middle ground. Such a strategy should not come as a surprise from business players. Regardless of what their real intentions were (getting to actually include such provisions in the final document, or simply inserting them as nothing but a negotiation ploy), it nevertheless reveals that the companies engaged in this process with a high level of self-confidence.

It is also worth noting that the companies did not feel committed to using the same vocabulary as the European Commission: the expression "hate speech", as such, is nowhere to be found in the whole document, despite the fact that it was included six times under its literal form in the EC's four-paragraph long "set of expectations"[21]. Instead, they used a diverse group of expressions in their draft, which did not fully embrace the meaning of the original concept: from "hateful contents" to "online content that expressly promotes or threatens violence".

This lack of alignment in terms of content as well as linguistics reveals that, at least at that stage, the two groups of actors were entering the talks on hate speech with quite different objectives and mindsets.

1.4. A key step towards bridging the gap: the face-to-face meeting in mid-April 2016

The drafting reunion that took place on April 15th contributed to narrow the gap between the two categories of actors. The earlier document, developed by the companies, was used as the starting point, although many changes were eventually made to it. They were introduced either during the meeting itself or immediately afterwards by the Commission "according to [their] discussions".[22] Therefore, except in a few cases,[23] it was not possible to trace back, solely from the edits on the document, who in particular initiated each change.

Nonetheless, it is a particularly useful document since it allows to spot, for each of the stakes identified above, what common ground was found between the two sides, at least provisionally:[24]

a) Normative basis for content assessment and removal

[21] Information taken from Annex B, p. 2.
[22] Information taken from Annex E2, p. 7.
[23] Sometimes, the content of the amendments or some short accompanying comments leave little doubt about their origins.
[24] Information taken from Annex E2, pp. 9-11.

In accordance with the Commission's initial expectations, the reported contents would not be assessed only against the companies' terms and conditions but also against "national laws transposing the Framework decision 2008/913/JHA". At this point, *both* normative references are supposed to be mobilized, as evidenced by the connector "and" between them. Furthermore, the introductory part of the document was complemented by a full paragraph stressing the necessity to apply the existing legal rules on the digital platforms: "it is essential to ensure that the relevant EU and national laws are being applied in the online world as well as offline".

About this first stake, the companies' draft was respected on only one point, related to the inclusion of provisions on hate speech in their terms and conditions.

b) Deadline to assess notices and remove contents

The Commission's will to set a 24-hour deadline eventually prevailed on the proposal from the IT sector, which was twice as long. This maximum amount of time would apply on "illegal hate speech" in general: at this stage, there was no goal set in terms of percentage of notices to be assessed, or content to be removed within this timeframe.

Quite strangely, the specific provision that was supposed to apply only to trusted flaggers and law enforcement was not struck down in this version of the document, despite the fact that it was very similar to the newly agreed procedure, now applicable for all reported contents.

c) Communication between Member states' and companies' national contact points

The wording chosen by the IT sector was fully maintained including, thus, the idea that Member States authorities need to be "educated" to properly report contents. Nevertheless, this formulation was not incompatible with the Commission's initial insistence on improving the communication between national contact points, except for one point: the standardization of notification forms, that the companies' draft ignored, was not reintroduced.

d) Involvement of CSOs

Most of the companies' proposals on this stake were preserved, in particular the CSO's need to get training in understanding the platforms' "Guidelines and Rules" as well as the reporting process.

The only substantial change had to do with the reintroduction of the Commission's expectation to "make public the criteria/procedure used to select trusted flaggers/authorized reporters". However, this specific addition was seen as tentative: as indicated by the note that followed, the companies still had to "assess" this sentence.

e) Hate speech and freedom of expression

The IT sector's words on the cardinal value of the freedom of expression were maintained in the introduction, but the paragraph itself was largely amended – its size more than doubled – to emphasize, also, the importance of the fight against hate speech. A new paragraph, entirely dedicated to the "devastating effects" of hate speech, was subsequently added.

f) Initiatives from the IT sector

Staff training on "societal development" was maintained intact, although a brief note was added at the end calling for further quantification or specification in the future. Regarding intra-industry cooperation, the sharing of "best practices" was not limited anymore to the definition of "robust" terms and conditions against hate speech, as the IT's draft suggested, but it was mentioned in general terms. The Commission added to the companies' commitments to enhance their transparency, leaving the companies the opportunity to "check if fine".

Concerning the provisions on counter narratives and critical thinking, they were left almost intact: the only change consisted in adding that the Commission was also subscribing to the view that both elements were key to tackling hate speech.

In addition to those pre-identified stakes, other two were introduced in this new document:

First, the companies are designated as "taking the leadership [in] strengthening the fight against illegal hate speech online". Since the companies' draft was less than emphatic concerning the dangers associated to hate speech, it is unlikely that they were the ones who pushed to be recognized such a prominent role in the process.

Second, a new paragraph, which would turn out to be particularly crucial, was added at the very end of the document: these public commitments would be subject to a "preliminary assessment", to be presented to the "High Level Group on Combating racism, Xenophobia and all forms of intolerance" in the second half of 2016. No specification is given on the mechanism that would operationalize such assessment. Once more, the content of this addition makes it very likely that it was promoted by the Commission: as the initiator of the project, it has a clear interest in developing ways to monitor its eventual implementation, while the companies would logically be more favorable to self-report to what extent they have delivered on their promises.

Hereinafter, both stakes will be known as:

g) Assignment of a leadership role
h) Assessment of implementation

This April 2016 meeting between representatives of the two sectors was key in that it led to the constitution of a first joint document, from which further versions would be developed. Although such document was built from the IT companies' first draft,

this version was notably tilted towards the Commission interests: its "set of expectations", which the IT sector largely ignored in its own draft, had been almost fully reinstated, especially its most crucial points (legal basis and assessment within 24 hours).

The overemphasis that the companies had granted in their draft to the freedom of expression, leaving only limited room for the fight against hate speech, was not in line with the Commission's approach, but the very substantial additions to the introductory section managed to frame the relations between the two in terms of a balance to be found. The new document is also explicit and unequivocal concerning the negative effects of hate speech, while the IT draft was almost silent on this point. The other two stakes that the IT sector introduced out of its own initiative were not directly challenging the Brussels institution's objectives, and they were slightly adjusted – although sometimes only tentatively – in a way that was either neutral or beneficial to its aims.

In terms of format, it is also noticeable that the document was given a shape that was much more consistent with the Commission's standards. This appears clearly in the introductory part, that precedes the commitments themselves, where references to the 2008 Council Framework Decision or the Joint statement by the extraordinary Council of 24 March were inserted: such specifications (legal basis, landmark antecedents) are typically found in EU-produced documents. Additionally, a definition of hate speech was provided, and this very expression replaces all the variants that were used in the previous version.

The comparison between the two parties' initial texts with the – still provisional – joint document shows consistent evidence that the Commission was having the upper hand on this key moment in the development process of the code of conduct. Did the subsequent phases in the drafting process confirm this trend?

1.5. The gradual evolution of the Code of Conduct on the major stakes

The first joint document that came out this first meeting on April 18th was modified repeatedly in the following weeks. In this process, a second meeting organized exactly one week later allowed to make further progress in the consensus-building effort, such as reaching an agreement on nothing less than the name of the document itself.[25]

According to the exchange of emails that was revealed afterwards, at least five amended versions circulated electronically among the participants:

- two of them were sent to the Commission the very same day (April 22nd), the first by Google and the second one by Twitter, on their respective behalf

[25] The name was still left undefined by April 22nd (information taken from Annex A, p. 59, also present in Annex C, p.59) but the document sent by the Commission one day after the meeting and reflecting the changes agreed in it, was bearing what would eventually stand as its definitive name (information taken from Annex E2, p. 21).

- The Commission sent an updated version of the Code one day after the April 25th meeting
- Google sent another version on May 13th in the name of all three companies
- The Commission sent the last version, identical to the final one (except for very minor formal details) on May 20th, right after sharing it to the Member States

Several of those documents provide information on two accounts: since the edits are sometimes left visible, the same document reveals both the original version and the changes that its author(s) want to see introduced. Here comes a recapitulatory table:

Day the original version was written	Day the edited version was sent	Author of the latest version	Annex and pages
April 19th	April 22th	Google	Annex A, pp. 59-61
April 19th	April 22th	Twitter	Annex C, pp. 59-62
April 26th	No visible edits	Commission	Annex E2, pp. 21-23
May 3rd	May 13th	Approved by all three	Annex A, pp. 15-18
May 20th	No visible edits	Commission	Annex E2, pp. 9-11

The available information provides us with a good outlook at the main steps in the development of the previously identified stakes, and allows us to note the addition of some more. The evolution of each of those stakes is analyzed by comparing the first "joint starting point" (represented by the April 19th document) with the final version of the Code of Conduct, i.e. the one that can now be consulted online,[26] with an emphasis on those provisional versions of the document where major modifications were applied on this particular point. In order to facilitate the identification of instances of changes and continuity, each stake will be treated in the same order as in the previous sections of this paper.

a) Normative basis for content assessment and removal

This specific point was highly controversial, and therefore intensively discussed throughout the negotiation process of the Code of Conduct. A Commission official that I interviewed at the beginning of 2019 openly described it as the major source of debate and this fact was unequivocally confirmed in several of the emails made available thanks to the EDRi request.

It is worth noting that the Twitter suggestions sent on April 22nd were scraping any references to national laws, to fully leave the room to the companies' "Rules or Community Guidelines". In contrast, in its own series of comments sent earlier the very same day (and that Twitter had the opportunity to consult), Google had proposed to maintain both, but connected them by "or", instead of "and". This difference on such a major issue shows that the players from the IT sector were not always on the same page, or speaking with a single voice, throughout the development process of the Code of Conduct.

[26] https://ec.europa.eu/newsroom/just/document.cfm?doc_id=42985

The option put forward by Google was the one that made its way through the next provisional versions of the document, and eventually to its definitive shape.

Therefore, the compromise struck between the companies and the EC was that both the terms and conditions and the national laws would maintained as normative references, but they would be used alternatively, instead of simultaneously. This stands in contrast with the EC's original intention, which in its "set of expectations" had insisted on having one being used *in addition to* the other[27].

On May 3rd, Facebook sent to the Commission a version of the Code of Conduct, with the explicit mention that it had been updated jointly by all three companies. The accompanying email, also on behalf of the whole group, was entirely dedicated to building a "common understanding" and to justifying the sector's position on the very issue of which normative basis would be used when assessing contents reported under hate speech.[28]

In substance, the argument was that the companies' terms and conditions "to a very large degree overlap with national laws in the EU" against hate speech. Starting from there, it would be more efficient for the platforms to make their assessment from their own rules, especially given the short amount of time they are given to perform this action. The rationale was that, in practice, it is much easier for the companies to make decisions based on their own rules, uniformly applicable everywhere, rather than mobilizing experts specialized in each national legislation for "full legal review". The argument continues by conceding that contents would be assessed against national laws "when appropriate", meaning only in the few cases when they would be illegal while not violating the companies' rules: in other words, only in what we could qualify as fringe or residual situations.

As an additional clue of the importance of this subject from the sector's perspective, this sequence of ideas was carefully and formally worded, in contrast with the brief and more informal character of most exchanges.

The EC accepted this change on May 4th, i.e. only one day after. This rapidity does not mean there was no reluctance: on the contrary, three emails were exchanged in the 24 hours after Facebook's email, two of which (representing three pages in total) were fully redacted, including the information related to the sender and recipient(s).[29] This situation, unique in the 374-page long trove of released correspondence, is revealing of the high stakes associated with this particular issue.

Nevertheless, a "way out" for this visibly contentious topic was suggested by the Commission and eventually accepted by the IT sector: the contents should be

[27] Information taken from Annex B, p. 2. Our emphasis.
[28] Information taken from Annex D, p. 11.
[29] Information taken from Annex E2, pp. 40-42.

assessed against national laws "where necessary", instead of "as appropriate"[30]. This proposal was introduced with a formality that echoes the email from Facebook, and with the specification that only "with difficulty" had they managed to receive the green light from their hierarchy:[31] the issue of the relevant normativity was decidedly a tough one for both sides.

From the Commission's perspective, the new wording articulated around the notion of "necessity" was certainly expected to reduce the margin of appreciation left to the platforms when it would come to decide which normative basis to use. However, differing interpretations are likely to subsist: it would not be far-fledged to suppose that the companies would interpret these newly introduced keywords as meaning "only when necessary", while the Brussels institution's understanding of the same part would stand much closer to "whenever necessary".

Ambiguity is seldom part of the recipe to strike deals in a negotiation and this is not the exception: with this specific wording on paper, the companies can keep on focusing mainly on their own terms and conditions when assessing reports (which they do, as confirmed in interviews), while the Commission can pretend that the law is still being considered by the companies in their effort to counter hate speech. Tellingly, Commissioner Jourova explained in her answer sent in June 2016 to the Director of European Affairs of the US-based CDT[32] who had expressed concerns about the implications of the Code of Conduct on the freedom of expression, that "the Code expects IT Companies to take action *only against content that is illegal*",[33] leaving aside the fact that the normative basis used most often would be, actually, the companies' own rules.

The title of the document itself, that explicitly targets *illegal* hate speech online, is another evidence of the Commission's intention to reassert the importance of national law as a normative reference to guide content take-down decisions, contradicting and, possibly, trying to compensate for the fact that, in practice, legislation would only be used marginally by the platforms.

b) Deadline to assess notices and remove contents

Since the April 18th meeting, the 24-hour deadline was accepted by both sides, but by then it was devoid of any quantification. The update developed by Google on April 22nd filled this void by introducing the goal of having "the *majority* of illegal hate speech reviewed in less than 24 hours and removed, if necessary".[34] This proportion

[30] Between these two terms, "when applicable" had also been introduced during the April 25th meeting. Information taken from Annex E2, p. 34 and Annex A, p. 60.
[31] Information taken from Annex E2, p. 39.
[32] This CSO's initials stand for "Center for Democracy and Technology".
[33] Vera Jourova, letter to Mr. Jeppesen, June 21st, 2016. Accessed in November 2019. Available at https://cdt.org/wp-content/uploads/2016/09/Commissioner-Jourova-to-Mr-Jeppesen.pdf [Emphasis in the original]
[34] Information taken from Annex A, p. 60. Our emphasis.

was the one that the Commission had set in its "set of expectations" and it remained unchanged in the final version.

One point of contention came up though: In the April 18th meeting, the participants agreed that this 24-hour time limit was to run from the moment a "valid" removal request was addressed to the platform,[35] but a controversy subsequently arose about what criterion allowed to consider that this validity requirement was met.

The companies must have realized that, if the validity was understood too broadly, any request could trigger a 24-hour countdown during which they would have to review and possibly remove contents. Therefore, they attempted to raise the threshold of admissibility in the updated version of the Code they presented on May 3rd, by adding that the request had to be not only "valid" but also "legal".

The Commission pushed back by expressing, elsewhere in the already mentioned May 4th email, its concerns that this adjective could be understood as requiring that each notice was supported by a Court order. Moreover, an unknown proportion of the two heavily redacted emails mentioned above was related to this point as well.[36]

Eventually, the word "legal" was removed from the expected characteristics of the notices: instead, a couple of lines were added in the preamble to define validity as "not being insufficiently precise or inadequately substantiated", in replacement of the IT sector's own suggestion. This was proposed by the Commission on May 17th, in a context of "extreme urgency", as described elsewhere in the same email.[37] Therefore, this specific point was one of the last two issues that had to be overcome to complete the whole drafting process of the Code of Conduct. The other one is related to the next stake.

c) Communication between Member states' and companies' national contact points

The sensitive call for the Member States authorities to be "educated" about how to properly report contents was first suppressed after the April 25th meeting, but the idea was later reintroduced by Google on May 13th in a more delicate way by referring to the necessary effort, by the Member states authorities, to "familiarize themselves with the methods to recognise and notify the companies of illegal hate speech online".[38]

In the final version, the verb itself was conserved but this familiarization was not presented anymore as a commitment, but as the outcome of the improvement in the

[35] Information taken from Annex A, p. 60.
[36] Information taken from Annex E2, p. 39. That email, which starts addressing the issue of the "legality" of the requests, refers to "the explanation in the message from [redacted name of a company] below".
[37] Information taken from Annex E2, p. 72.
[38] Information taken from Annex C, p. 61.

flow of information between the companies and the Member states' respective national contact points.

As to the standardization of notification forms, which was an original expectation from the Commission, it never surfaced again in the written conversation.

d) Involvement of CSOs

In this field, the IT companies' position made its way to the final version without any significant changes: they would collaborate with CSOs considered as "trusted reporters", would provide them with the necessary training – thus being in a position to spread their own standards and proceedings. They expressed interest in benefitting from their expertise in the domain of countering hate speech.

In the final document, no requirement applies in terms of transparency about how the status of "trusted reporter" is granted and obtained, a step back for the Commission, which had tried to reintroduce this commitment little after the April 18th meeting, and explicitly insisted on it after the 25th meeting, where an agreement on this point appeared to have been reached.[39]

However, this provision was eventually struck down in the May 13th document supported by all three companies,[40] a blow to the Commission's expectations. This fact convincingly explains why the EC seemed to consider this particular issue as especially sensitive, as shown by the full redaction of the specific part that was dealing with it in the May 26th email.[41] The absence of public standards on this point allows the platforms to maintain a significant amount of control about who they would partner with.

e) Hate speech and freedom of expression

On this point, the content of the April 18th version remained mostly unchanged. Some marginal but symbolically significant modifications were nonetheless performed. For instance, in the first joint version of the Code of Conduct, the document started with a paragraph stating that the IT companies were sharing a "collective responsibility and pride" in both promoting free expression and condemning hate speech.[42] This wording was placing both actions at an even level of importance, but in was not anymore the case after the April 25th meeting: from then, the platforms' role in permitting free speech was emphasized first and, in a separate paragraph, they are

[39] "This issue of making the [sic] public the criteria for selecting the trusted flaggers has been *solved* by making reference the [sic] need 'to make public the procedure used to select 'trusted reporters'" (information taken from Annex C, p. 66). Our emphasis.

[40] Information taken from Annex A, p. 17.

[41] Information taken from Annex E2, p. 19. It was possible to know the content of the hidden lines since the very same email was also included in Annex C (p.66), this time without any redaction.

[42] Information taken from Annex E2, p. 9.

presented as "*also* shar[ing] the European Commission's and EU Member States' commitment to tackle illegal hate speech online".

Along the same logic, the ostensibly strong language about the impacts of hate speech in society, found in the April 18th document, was watered down by the IT sector by mid-May: the "devastating effects" were replaced by "negative effects" and its potential to "intimidate and silence" was morphed into less telling reference to "negative impacts".[43] While the Commission conceived hate speech as a phenomenon that was serious enough to justify significant restrictions on the exercise of free speech, the IT sector was adamant – and successful – in opposing this view.

f) Measures internal to the IT sector

The measures that the companies voluntarily introduced in their very first draft were all maintained, without any change in substance. The Commission's half-hearted[44] attempt to have them commit to increase transparency was short lived, since this provision was simply struck down from the April 22nd version.[45] Three days later, in the April 25th meeting, the concept was reintroduced but now in the closing paragraph, which deals with the assessment of the commitments, announces future discussions about "how to promote transparency", among other subjects:[46] quite clearly, the idea of transparency subsisted, but its implications and impact on the companies had been considerably lowered.

Unsurprisingly, the provisions on counter-narratives and critical thinking went across the full process with minimal alterations, except from the fact that the counter-narratives were still expected to be "independent", but not necessarily "effective" anymore.

g) Assignment of a leadership role

The statement according to which the IT sector would play such a prominent role in the fight against hate speech, added after the first joint meeting on April 18th, was maintained in the final version of the Code, although under a slightly diluted form, by less emphatically referring to a role consisting in "taking the *lead*". This new wording was introduced in the April 25th meeting[47] in substitution to the *leadership* that had been proclaimed exactly one week earlier.

[43] Information taken from Annex A, p. 15.
[44] "Check if fine" at the end of that line, in the April 19th email (Annex A, p. 60)
[45] Information taken from Annex A, p. 60. The Twitter comments emitted the same day (Annex C, p.61), did not eliminate this provision but, as indicated elsewhere, was not used as a starting point for the next exchanges.
[46] Information taken from Annex A, p. 18.
[47] Information taken from Annex E2, p. 28.

The preservation of such notion can be seen as the result of the Commission's insistence on this point, since as a symbolically relevant step allowing to portray the Code as primordially an intra-industry agreement, voluntarily developed and agreed by the players from the business sector.

On the contrary, it is noticeable that the companies themselves were promoting changes to the text in the other direction. As we have already highlighted it, several of their suggested amendments were rather heading towards a minimization of their responsibilities in this area. As an additional example, the April 25th meeting also ratified a change promoted by Twitter in its own document sent three days before, which consisted in replacing "the signatories want to respond to the challenges of ensuring that the Internet do not offer opportunities for extremism and intolerance"[48] by "the signatories support the Commission and EU Member States in their efforts to respond to the challenges…"[49]: such a formulation, that subsisted in the final version, grants a supportive rather than leading role to the companies.

Despite these contradictions the issue of "leadership" did not appear to generate a major controversy: according to the available written documents, the exchanges on this particular topic were quite limited. Such contradictory signals serve to explain this apparent consensus: both sides feel

h) Assessment of implementation

From its introduction to the provisional versions of the Code to the adoption of the document in its definitive form, this stake was not included within the list of commitments, signaled by bullet-points. Instead, it stands as a conclusive paragraph for the whole document:

"The IT Companies and the European Commission agree to assess the public commitments in this code of conduct on a regular basis, including their impact. They also agree to further discuss how to promote transparency and encourage counter and alternative narratives. To this end, regular meetings will take place and a preliminary assessment will be reported to the High Level Group on Combating Racism, Xenophobia and all forms of intolerance by the end of 2016".

Thus, the key provision regarding the reporting is maintained, as well as the timing of its presentation, hardly more than half a year after the approval of the document.

However, this part of the Code was changed on one crucial point: in the wording that came out of the April 18th meeting, this commitment was only directed at the companies, which would have left them free to put in place the conditions of the assessment of the implementation of the Code of Conduct and therefore to design the evaluation mechanism to which their own actions would be subjected. The final version sharply contrasts makes clear that this task would be conducted by both the

[48] Information taken from Annex A, p. 59.
[49] Information taken from Annex C, p. 59.

IT companies and the European Commission. This change was adopted in the April 25th meeting, or very close to that date.[50]

Along the same logic, a reference to the organization of "regular meetings" on the topic of assessment was added: from the Commission's perspective, such specification was giving weight to the idea that the assessment would be subject to a follow-up and to a process of incremental improvement. For the IT sector, it meant that it would be associated to the discussions on the development of the mechanism to be put in place.

What is striking however is the lack of precision about the concrete characteristics of the still-to-be-created assessment mechanism: while the issue of its existence was settled, as well as the reporting obligation, no further information was provided about its shape, functioning or periodicity. Implicitly, it meant that all those aspects were left to be defined at a later, yet still undefined stage.

1.6. Beyond the evolution of the major stakes, tweaks to the tonality of Code

In addition to actions related to the specific stakes identified above, the companies have been making a constant effort to modify the wording in ways that allowed them to reduce the scope of their commitments and to convey the idea that these commitments were in fact to be seen as the continuity of actions they were already performing against hate speech.

Firstly, in several instances apparently minor adjustments served the IT sector's purpose to minimize the practical implications of their pledges. After the April 18th meeting, they were expected to build partnerships with CSOs "all over the EU geographical area", then the term "all" was removed in the version promoted by Google[51] four days after and this section was finally rephrased, after the April 25th meeting, as "widening the geographical spread of such partnership",[52] a less specific goal expressed in terms of process instead of outcome. In the email that accompanied this version of the Code, the Commission alluded to this change and described it as the consequence of a "Twitter contribution".[53]

Furthermore, the IT sector has consistently tried to portray those commitments in terms of continuity: its purpose was visibly to show that this text was not launching something new, but officializing and somehow structuring actions that had already been under way for some time. In that sense, expressions such as "reaffirm their commitment", "intensify cooperation", "continue or […] increase the scale of their proactive outreach to civil society" were present since the very first version they presented to the Commission,[54] while wording with a similar meaning such as

[50] Information taken from Annex A, p. 61.
[51] Information taken from Annex A, p. 60.
[52] Information taken from Annex E2, p. 22.
[53] Information taken from Annex C, p. 66.
[54] Information taken from Annex D, p. 3.

"strengthening of the fight against hate speech"[55] as well as "continue their work" and "intensify their work"[56] were added afterwards.

2. The Code of Conduct in practice: Monitoring and communication of the results

2.1. The design of the monitoring exercise: The CSOs enter the game

For being a "voluntary agreement" among companies, no enforcement mechanism could have been built into the Code of Conduct. This is the reason why, instead, a mechanism to assess the outcomes was announced in the final part of the document, as an incentive for the companies to meet their own commitments. As highlighted above, its eventual existence was simply stated, without further details.

Such mechanism, designated by the actors involved as the "monitory exercise", was designed during the months that followed the adoption and official presentation of the Code of Conduct, on May 31th 2016, and implemented for the first time in October of the same year. The official documents published by the Commission on this subject are not explicit about who exactly developed it: the first public report on the results of the implementation of the Code of Conduct only indicates that it was *agreed*[57] on October 5th by the "sub-group on countering hate speech online", a structure that the Commission created in July 2016 to gather representatives from the Member States, the IT sector and CSOs.[58]

Thanks to direct talks with several players, I learnt that the development of such mechanism did not involve the same players as for the drafting of the document itself: The Commission clearly did play a central role in the designing process of the methodology but this time the companies were not associated to it, at least not formally. Instead, some CSOs played a significant role, specifically the ones that would eventually participate in the first implementation of the mechanism.[59] The coordination between the two sides took place through meetings and exchanges of emails. One representative from an influential CSO testified to me in an electronic exchange that their organization had been "deeply involved in the development of the exercise".

Some of these organizations already had a practical experience in testing the responsiveness of digital platforms to reports that are communicated to them. It was the case of Jugend Schutz, which at least since 2007 publishes statistics about the answers it receives from Internet actors after reporting illegal contents (hate speech

[55] Information taken from Annex E2, p. 9.
[56] Information taken from Annex A, p. 62.
[57] https://ec.europa.eu/newsroom/document.cfm?doc_id=40573
[58] https://ec.europa.eu/newsroom/document.cfm?doc_id=42830. Therefore, these are the same participants as in the *ad hoc* meetings of March 4th and May 31st which, officially, launched and closed the drafting process of the Code, respectively.
[59] There were twelve of them, from nine different EU member states.
https://ec.europa.eu/newsroom/document.cfm?doc_id=40573

being one of them)[60]. INACH, an umbrella association of CSOs, was another example through its project "Research – Report – Remove: Countering Cyber Hate Phenomena", implemented since mid-2016[61]. Such already existing processes were qualified as an "inspiration" by a Commission official met at the beginning of 2019.

In the implementation reports, the methodology, is briefly described as follows. To the best of our knowledge, there is no publicly available information that is more detailed than that on the matter:

- CSOs from several member states "volunteered to test the reactions of IT Companies upon notification of alleged illegal hate speech content and to record their response".[62] In practice, the number of participating organizations increased from 12 in 9 member-states for the first implementation at the end of 2016, to 35 (plus 4 public bodies) in 26 countries for the latest one, at the beginning of 2019.[63]
- The exercise takes place over a six-to-seven week period but no predefined periodicity is mentioned
- The organizations report the answers that they have received to the notifications they sent to the platforms. In practice, this sample raised from 600 in the first edition to almost 3850 in the most recent one.[64]
- Since the second monitoring exercise, the CSOs "often" report the content first as regular users and do it as trusted flaggers only if their first notification was not taken into account.[65]

2.2. The implementation of the monitoring exercise: The companies found their comfort zone

While the involvement of the IT companies in the designing process of the monitoring exercise was practically inexistent, the conditions of its implementation have resulted beneficial to them in practice for several reasons.

First of all, they are aware of the moment when such an exercise is conducted. Although they are not officially notified, they have many reasons to know this. To start with, in the name of "transparency", the Commission indicates an approximate period (a range of three months) during which the six week-long assessment will

[60] See its "Annual Report 2008", which includes a comparison with data from the previous year (http://www.jugendschutz.net/fileadmin/download/pdf/jsn_annual_report2008_web.pdf, page 15)
[61] http://www.inach.net/project-research-report-remove-countering-cyber-hate-phenomena/
[62] European Commission (2017). Factsheet – First monitoring round of the Code of Conduct https://ec.europa.eu/newsroom/document.cfm?doc_id=40573
[63] European Commission (2019). Factsheet – Fourth monitoring round of the Code of Conduct https://ec.europa.eu/info/sites/info/files/code_of_conduct_factsheet_7_web.pdf
[64] According to the two reports quoted above. However, the fourth report officially declares that "4392 notifications were submitted to the IT companies taking part in the Code of Conduct" (p. 1), while the sum of the reports made by each CSO corresponds to the lower number, that I provide in this paper. A similar and, so far still unexplained, discrepancy is found in the second and third reports.
[65] European Commission (2018). Factsheet – Third monitoring round of the Code of Conduct http://ec.europa.eu/newsroom/just/document.cfm?doc_id=49286

take place. Since it is usually organized at the very end of the year, the last two weeks can be discarded, leaving little uncertainty left. This uncertainty is further reduced by the pragmatic recognition that "Brussels is a small world", in the words of a CSO member I talked with, and by the fact that it is quite easy for the platforms to infer, thanks to the sudden increase of reports from trusted flaggers, that something unusual is happening.

Knowing that their responsiveness is being monitored allows them to display efforts they would not make otherwise. In more than one occasion, interviewed people from CSOs presented as self-evident that the companies are reviewing reports much more diligently during the implementation of the exercise, in comparison to regular times: this is something that they have been able to observe directly.[66]

Second, the CSOs expressed serious doubts on their ability to effectively report contents as anonymous users, especially when taking into account that they do it through platforms whose business model relies precisely on being able to develop a deep knowledge on their users. Therefore, they are fully aware that their behavioral pattern could easily be detected, and their reports treated accordingly. This lack of effective anonymity might partially explain why, according to the latest report, "the divergence in removal rates of content reported using trusted reported channels as compared to channels available to all user was only 4.8 %".[67]

This lack of trust from the CSOs regarding the representativeness of the sample of cases treated during the monitoring is reflected in the answers provided by 27 of the organizations involved in the implementation of the monitoring process.[68]

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? (n=27)		
Assessment	Number of answers	Percentage
1 (not at all)	1	3.7%
2	11	40.7%
3	10	37.0%
4	5	18.5%
5 (fully)	0	0%

Therefore, less than one fifth of the consulted organizations indicated that the reports published by the Commission on the implementation of the Code of Conduct could be considered as having a level of representativeness above medium.

[66] According to a written answer by one CSO representative: "most CSOs state that the removal rates and response times of social media are much worse in general than during MEs [monitoring exercises]".
[67] European Commission (2019). Factsheet - Fourth monitoring round of the Code of Conduct https://ec.europa.eu/info/sites/info/files/code_of_conduct_factsheet_7_web.pdf
[68] The poll was sent to 30 organizations from April to July 2019, through emails directly sent to each of them individually, with an exposition of the motives behind the invitation to answer it. It consisted in 6 different questions and the CSOs had to option to answer it anonymously (which 26% of them did).

The Code of Conduct designated both national laws (transposing the Framework decision 2008/913/JHA) and the companies' terms and conditions as the alternative options for determining whether a given content should be removed. Which of them finally prevailed in the implementation phase?
Most of the time, the CSOs use the law as the normative basis when reporting contents that they considered hate speech, as show in their answers below:

On what basis do you flag content that you consider as illegal hate speech online? (n=27)		
Normative basis[69]	Number of answers	Percentage
National law	19	71.1%
Each platform's Community Standards	1	3.7%
Both	6	22.2%
Either one or the other, depending on what is the most convenient for every case	1	3.7%
Other (to be specified)	(1)[70]	(3.7%)

Consequently, almost all CSOs (92.6% of them) use either the law or both the law and the companies' rules. This fact could be seen as meeting the Commission's expectations and the position it consistently defended throughout the drafting process of the Code. However, regardless of what normative basis is invoked by the reporting CSO, the companies usually make their decision on the ground of their own rules. This is what representatives from the IT sector openly exposed to me, arguing in substance that the companies are not jurisdictional bodies, therefore their role is not to enforce the law but merely to respect it.

To confirm this point, a CSO member showed me the answer that they had received after reporting a case of illegal hate speech online. Although the removal request mentioned nothing but the national law, the incriminated message was taken down for violating the companies' terms and conditions.

If the focus is set on the immediate outcome, it can appear to be equivalent: the content was taken down anyway. But this is not the case for two reasons.

First of all, this practice establishes an asymmetric dialogue between the parties: a request is made in the name of one specific normative reference, but its answer is based on another one. Each is speaking its own preferred language, which is a source of ambiguity and misunderstandings in an area – law – that is precisely expected to avoid it.

Second, the practical implications are not the same: if the content were to be qualified as hate speech solely under the law, it would have to be "geoblocked" instead of plainly removed: as a consequence, it would cease to be visible in the country whose law has been used as the normative basis, but it would remain

[69] The options were presented in the same order as they appear below. Only one answer could be selected
[70] The only one CSO that chose this option wrote "national law combined with EU and international obligations […]", therefore its answer has been added to the ones which selected "national law".

accessible from anywhere else in the world. On the contrary, any content found to be breaching the platform's rules would be taken down, regardless of any geographical considerations.

In other words: removing a content under a company's rules produces much larger effects than doing it under a given state's legislation. Therefore, it is in the CSOs' interest to keep on accepting that their notices based on a given text are eventually treated and resolved under another one – be it of a private nature. The Commission finds itself in a similar situation: in terms of geographic coverage, its efforts against hate speech are, in the current configuration, better served by private norms than the national laws. The very same national laws it tries to have the platforms apply.

When it comes to implementing a critical part of the Code of Conduct, the companies are able to mobilize their assets – technical control of the space to be regulated and global transnational reach – to preserve their interests in this particular stake. In contrast, the Commission is hindered by its clear limitations on both accounts, which gets translated into a weaker position and the obligation – so far – to accept the current state of affairs in the implementation phase.

2.3. Public communication about the outcomes: A common interest in portraying it as a success

Four monitoring exercises have taken place so far, with reports – in fact labelled "fact sheets" – published a few weeks after the completion of each assessment: in December 2016, June 2017, January 2018 and February 2019.

The public availability of such information is key to the whole process, since the public nature of the figures is expected to put some pressure on the companies, so they get ever closer to their commitments – or maintain their efforts at the expected level. For companies that have been under the spotlight in the past years for their responsibility (or lack thereof) for the content that circulated on their platforms, or for the improper sharing of their users' data, the preservation of a positive public image is an objective taken quite seriously.

However, it is worth noticing that, in the Code of Conduct's final paragraph, the commitments from the companies were qualified as "public" whereas the assessment was not: the only requirement was to report it to the High Level Group, without any specification about whether its content should be made available for the public in general. Therefore, the official presentation of the findings of the monitoring exercise developed out of practice rather than from the text itself.

This fact may partially explain why the information disclosed through those "fact sheets" turns out to be quite controlled. The main points are provided, following an order that has been evolving from an edition to another. In the last two deliveries, they were presented as follows:

1. Notifications of illegal hate speech: number of CSOs involved, from how many countries, overall number of notifications, then broken down per the channel

used (as trusted flaggers or general users) and per company to which they were addressed.
2. Time of assessment of notifications: percentage of cases meeting the 24-hour deadline, and those treated in less than twice this time, overall and broken down by company.
3. Removal rates: percentage of times a notice led to a removal, overall and broken down by company and by kind of hateful content. Also, a comparison of removal rates depending on the channel used (as trusted flaggers or general users) is provided. Two graphs visually show how this rate has evolved across editions, broken down by company and by country.
4. Feedback to users and transparency: percentage of cases when the users receive information on the decision made on the case they reported, broken down by company and by channel used.
5. Grounds for reporting hatred: breaking down of contents depending on the motive for which they were reported (ethnic origin, xenophobia…)

The sixth and last page is dedicated to exposing the methodology, which is summarized in six short paragraphs and presented as constant since the first exercise. The document ends with a list of the participating CSOs, with the number of cases they have reported during the corresponding exercise.

Therefore, while the information presented in the fact sheets cannot be qualified as minimalist, it is still insufficient since the Commission is the one who decides which data should be included in each delivery, therefore conserving the option not to disclose some figures if deemed unfavorable – except for the central ones, whose absence would be noticed and considered abnormal. For instance, in the second delivery, a specific section was created to compare the removal rate depending on the user, broken down by company.[71] However, this information is not provided anymore in the next two fact sheets, at least not in this broken-down form. This absence can certainly be explained by an "innocent" motive, such as the administration of a limited physical space on the PDF document but it could also rely on less acceptable reasons, such as the willingness not to expose figures that could turn out to be embarrassing for one or several companies – or for the Commission itself.

This suspicion could be easily avoided by offering the possibility to download the databases of unprocessed data, as it is often the case in webpages providing statistical information. Similarly, the forms that the CSOs must fill and transmit to the Commission to record the fate received by their notices could be made public, so the general public can be informed about the kind of information collected. Even a direct request to have access to such document was denied.

[71] For instance, a graph indicates that in mid-2017 Twitter removed 31.5% of the content reported by the CSOs as general users, but 48.5% when the CSOs did the same as trusted flaggers, which in both cases represent a sharp increase in comparison to the previous exercise.
European Commission (2017). Factsheet – Second monitoring round of the Code of Conduct
http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=71674

This way, the EC conserves the latitude to select some of the data that will be presented, a virtual possibility that runs against the objective of transparency and which allows it to preserve the image of the Code of Conduct as a functioning mechanism, an idea that is beneficial for all the parties involved.

Following the same logic, in the fact sheet itself the Commission is more insistent on the positive developments than on the deficiencies. For instance, the results of the first monitoring exercise were described by several of my interlocutors as highly disappointing, since all three companies fell short of their commitments. This was in particular the case for one of the central objectives of the Code, related to the time spent to review a reported content:

> Data recorded show that in 40% of the cases IT Companies reviewed the notification on the same day (less than 24h) and in 43% of these cases on the day after (less than 48h).[72]

In this case, the information is simply delivered, without any explicit reference to the 50% benchmark. In contrast, there *is* such reference when this goal is reached:

> The target of reviewing notifications within one day is now met by all IT Companies and there has been a steady progress compared to the previous monitoring exercises.[73]

Instead of openly signaling and admitting that the results were below the expectations, the report for the first monitoring exercise appears to try to minimize this fact, by stating that

> "The monitoring exercise is a continuous process. These initial data constitute a baseline and a first valuable indication of the current situation. A second monitoring cycle will be carried out during 2017 to observe trends".[74]

It is hard to guess, from this reading, that the organization of the second monitoring exercise was in fact programmed six months earlier than planned – which explains why two such exercises took place in 2017 – in order to rapidly be able to produce and report better outcomes.

No doubt that, over time, much more satisfactory results were obtained, especially if we compare the numbers presented in the first and the latest reports, respectively by end-2016 and beginning of 2019. For instance, the overall percentage of cases reviewed within the 24-hour deadline raised from 40 to almost 89%, while the removal rate, although not a formal objective in and by itself, climbed from 28.2 to 71.7%.

---

[72] European Commission (2017). Factsheet – First monitoring round of the Code of Conduct, p.1. https://ec.europa.eu/newsroom/document.cfm?doc_id=40573

[73] European Commission (2018). Factsheet – Third monitoring round of the Code of Conduct http://ec.europa.eu/newsroom/just/document.cfm?doc_id=49286

[74] European Commission (2017). Factsheet – First monitoring round of the Code of Conduct, p.1. https://ec.europa.eu/newsroom/document.cfm?doc_id=40573

However, such encouraging trends and absolute values can have to do at least in part with the companies having successfully developed practices, such as the ones described in the previous section, allowing them to "game" the monitoring exercise and hit the expected numbers, rather than making progress solely thanks to their increased capacity to tackle hate speech on their platforms.

Out of the three steps described – drafting of the code, implementation of the monitoring exercise, and monitoring of the results – the last is the one where the Commission arguably enjoys the widest margin for maneuver, for a simple reason: it is the only one in charge of processing and finally presenting the data.

This power can be illustrated by one sentence in particular, found in the methodology section of the latest three implementation reports, in which it asserts that "the organizations *only* notified the IT companies about content deemed to be 'illegal hate speech' *under national laws*" [our emphasis in both cases]. Although we saw earlier that the CSOs tend to use the existing legislation more often that the companies' terms and conditions, this statement is not an accurate picture of what is really happening in practice: first, the CSOs are free to choose the normative basis of their choice and second, decisions are usually rooted in the companies' rules, rather that the states'. Here, the wording reflects more the Commission's expectations than the actual reality.

However, when drafting the implementation reports, the Commission consistently tries not to harm the image and reputation of the companies – the very same elements thanks to which the Commission was supposed to put pressure on them in the context of a voluntary agreement. This can be explained by its intention to cultivate a positive relationship with the actors of the IT sector, in order not to compromise the prospects of successful results in the longer run. As a complementary explanation, the Commission may also be aware that, if the companies fail to meet the commitments set in the Code of Conduct, the failure will also be seen as its own. Therefore, the Brussels institution is bound on two accounts: bound to the success of the process it fostered, and bound to maintain a positive working relationship with the IT companies.

Nonetheless, this assessment must be nuanced by the fact that the Commission still has the possibility to switch from this voluntary and negotiated mechanism, coupled with this soft monitoring of fulfilment, to a more classical regulation, based on hard law and dissuasive enforcement measures. This option was not openly raised in the reports themselves or in the written press releases that accompanied them, but it was verbally mentioned by Commission officials and reported by the press: " 'The good results do not mean the companies are off the hook, we will continue to monitor their efforts... If efforts slow down, we will consider some kind of regulation', [Vera Jourova] warned".[75]

[75] Kayali, Laura (2019). "Brussels: Facebook, Google, Twitter improved action against online hate speech". https://www.politico.eu/article/european-commission-facebook-google-twitter-improved-action-against-online-hate-speech/

A hard regulation against online hate speech at the European level is not a mere abstract possibility: some member states have already approved full-fledged regulation on the subject or are in the process of doing so, while the EU is already discussing a regulation on terrorist content. Therefore, the IT companies also have strong incentives to play the game that the Commission has started, but this does not preclude them from trying to play it in a way that maximizes their interests.

[CONCLUSION PENDING]