Council of the
European Union

Brussels, 27 September 2019
(OR. en)

12522/19

LIMITE

ENFOPOL 422	**DROIPEN 149**
ANTIDISCRIM 34	**DIGIT 145**
TELECOM 309	**DAPIX 282**
SOC 634	**CYBER 267**
MIGR 155	**COSI 201**
JAI 993	**COPEN 374**
FREMP 134	**COHOM 108**
EDUC 390	**AUDIO 102**

INFORMATION NOTE

From:	European Commission
To:	Permanent Representatives Committee/Council
Subject:	**Assessment of the Code of Conduct on Hate Speech on line**
	State of Play

Delegations will find attached an Information Note on the item mentioned above provided by the Commission services for the JHA Council on 7/8 October 2019.

JHA Council 7 October

Information note

Progress on combating hate speech online through the EU Code of conduct

2016-2019

The Code of conduct on countering illegal hate speech online was signed on 31 May 2016 by the Commission and Google (YouTube), Facebook, Twitter and Microsoft hosted consumer services (e.g. Xbox gaming services or LinkedIn). In 2018 and 2019, Instagram, Google+, Dailymotion, Snap and Jeuxvideo.com have joined. This means the Code now covers 96% of the EU market share of online platforms that may be affected by hateful content[1]. This document provides an assessment of progress made since 2016, following the structure and the commitments set out in the Code of conduct. It is based on data collected by the Commission during regular monitoring exercises as well as on selected information received by the IT Companies at regular intervals.

In summary, the Code of conduct has contributed to achieve quick progress, including in particular on the swift review and removal of hate speech content (28% of content removed in 2016 vs. 72% in 2019; 40% of notices reviewed within 24h in 2016, 89% in 2019). It has increased trust and cooperation between IT Companies, civil society organisations and Member States authorities in the form of a structured process of mutual learning and exchange of knowledge. This work is complementary to the effective enforcement of existing legislation (Council Framework Decision 2008/913/JHA) prohibiting racist and xenophobic hate crime and hate speech and the efforts needed by competent national authorities to investigate and prosecute hate motivated offences, both offline and online.

The Code of conduct requests IT Companies to:

- **have rules and community standards that prohibit hate speech and put in place systems and teams to review content that is reported to violate these standards.**

All IT Companies that signed the Code now have and continuously revise terms of service, rules or community standards prohibiting users from posting content inciting violence or hatred against

[1] https://gs.statcounter.com/social-media-stats/all/europe

protected groups. Interestingly, both Jeuxvideo.com and Dailymotion have substantially reviewed their terms of service to include a more precise definition of hate speech as prohibited content, in view of their participation to the Code. Snap, which joined the Code in Spring 2018, has during the same year entirely revamped its Safety Centre, which now contains information on prohibited content, including hate speech, for individuals, law enforcement and educators.

All platforms have also significantly increased the number of employees monitoring and reviewing the content. Facebook reports having a global network of about 15,000 people working on all types of content review and across Google and YouTube there are more than 10,000 people working to address content that may violate the company's policies.

- **review the majority of the content flagged within 24 hours and remove or disable access to hate speech content, if necessary**

On average, IT companies are now assessing **89% of flagged content within 24 hours**, up from 81% one year ago. Instagram, which was tested for the first time in 2018, reviewed more than 77% of the notifications within a day. A few months after the launch of the Code, the number of notifications reviewed in 24 hours was 40%. Dailymotion and Jeuxvideo.com have not yet been tested as part of the Commission regular monitoring exercises: however, they report that above 90% of the notices received in 2019 were reviewed within 24 hours. Snap reports that the vast majority of the content flagged is actioned within a few hours, and all content on Snapchat anyways disappears within 24 hours.

The removal rate is now stable at more than 70% on average. In 2016, after the first monitoring exercise on the implementation of the Code of conduct only 28% of the content flagged was removed. The current average removal rate can be considered as satisfactory in an area such as hate speech, given that the line against speech that is protected by the right to freedom of expression is not always easy to draw and is highly dependent on the context in which the content was placed.

Some of the IT Companies which joined the Code more recently, report having achieved a significant decrease in hate speech notices (e.g. Dailymotion going down from 27 000 notices in the first semester 2018 to 17 000 in the same period of 2019) thanks to the strategies put in place to comply with the Code. Gaming services (e.g. Xbox or Mixer) have implemented measures to foster

human moderation of hate speech in chats and fora, which led to the identification and blocking of 20 million pieces of content in 2019, including hate speech.

- **provide regular training to their staff**

All IT Companies report that they are holding regular and frequent trainings, and provide coaching and support for their teams of content reviewers, including on the specificities of hate speech content. Dailymotion is holding bi-weekly trainings on hateful material for their staff. Facebook has established a Product Policy Forum that gathers all its experts around the world every two weeks, to discuss potential changes to the community standards and capture new issues, trends and developments. The minutes of these meetings are publicly available.

- **Engage in partnerships and training activities with civil society in order to enlarge their network of trusted reporters.**

The IT Companies reported a considerable extension of their network of "trusted flaggers" in Europe since 2016. They are engaging with them on a regular basis to increase understanding of national specificities of hate speech. Twitter has enrolled 73 new trusted flagger organisations since signing the Code. YouTube has a four-times bigger network of trusted flaggers specializing in hate speech today compared to 2016, moving from 10 to 46 non-governmental organisations (NGOs); Facebook has increased its network of 82% (from 9 partners in 2016 to 51 today).

Since the signature of the Code, Facebook/Instagram have organised a total of 51 training sessions on its community standards in relation to hate speech, for up to 130 civil society organisations operating as trusted flaggers. Out of 38 training sessions provided in 2018 by YouTube to NGOs on their content policy and trusted flagger programme, 18 were focused on hate speech and abusive content. In 2019, YouTube conducted an additional round of trainings with 15 NGOs across 8 countries.

YouTube also reports on the significant impact of this extended network on the number of notifications by trusted flaggers: these have doubled from the trimester October-December 2017 and April-June 2019. At Facebook, from the fourth quarter of 2017 to the first quarter of 2019, action taken for hate speech violations increased from 1,6 million to 4 million (150% increase).

- **work [with trusted flaggers] on promoting independent counter-narratives and educational programmes**

IT companies also work together with their "trusted flaggers" on campaigns for tolerance and pluralism online. Between 2017 and 2019, three workshops took place at the headquarters of YouTube, Twitter and Facebook to facilitate such initiatives. A fourth workshop is planned by the end of 2019. As a result of these, more than 40 NGOs during the European elections of 2019 launched an EU wide online campaign in 24 languages, focused on promoting healthy and tolerant conversations online under the hashtag #WeDeserveBetter. The campaign has reached over 6 million users on Facebook and Twitter and has received support by the IT Companies in the form of advertisement grants. A pilot exercise run in 2018 to test a campaign reached over 2 million users in several Member States.

Microsoft has started a partnership with expert think tanks like the Institute for Strategic Dialogue on counter speech to assist NGOs to surface and serve impactful counter narrative content via advertisements on Bing.

- **Designate national contact points for receiving notices, in particular by national authorities**.

All IT companies that subscribed to the Code of Conduct have established national points of contact to facilitate contact with the relevant competent authorities at national level. It is important to highlight that the work in the Code of Conduct complements legislation fighting racism and xenophobia (Council Framework Decision 2008/913/JHA), which requires authors of illegal hate speech offences - whether online or offline - to be effectively prosecuted. Twitter is organising annual Law Enforcement Trainings for national authorities and contact points in the Member States and has provided specific guidance on how to report or request information.

- **Promote transparency towards users as well as to the general public**

In 2016, IT Companies only made information available on the number of law enforcement requests and did not provide any detail on online hate speech as a specific ground for removal. Today, the removals of hate speech content are clearly presented, on a regular basis, in each of the IT Company transparency reports, for example see the transparency reports published by Facebook, Twitter and YouTube. Both YouTube and Facebook have in 2019 launched dedicated pages on their transparency reports to enforcement of community standards regarding specifically hate speech content, including break down of data e.g. on counter notices, and automatic detection. There is

however still a lack of granularity, the published figures do not provide information on the time of review of the notices or the geographical distribution of the hate speech content flagged.

Before the Code of Conduct was launched, users rarely received a response by IT companies when they notified hate speech content. In addition, the reporting or flagging function was often not very user-friendly. Twitter has developed several improvements to the users reporting system, including to allow multiple reports of tweets by the same account. YouTube and Facebook have now in place "dashboard" systems by which users can monitor the outcomes of each of their flags. According to the results of the monitoring exercises, on average around two thirds of the notifications receive a response detailing the outcomes and measures taken. The performance by IT platforms differs, and only Facebook and Instagram are systematically sending feedback to notifications (more than 95% of notices is responded to). Further progress is therefore expected in this specific area in the coming months.

Transparency and feedback are also important to ensure that users can appeal a decision taken regarding content they posted as well as being a safeguard to protect their right to free speech. Facebook reports having received 1.1 million appeals related to content actioned for hate speech between January 2019 and March 2019, and 130,000 pieces of content were restored after a re-assessment.

Beyond the commitments in the Code of conduct: the role of technology and automatic detection tools

As part of the efforts to improve the way hate speech content is detected and removed, IT Companies are making an increasing use of technology and automatic detection systems. Facebook reported that in the first quarter of 2019, 65.4% of the content removed was flagged by machines (with an increase from the 51.5% of the previous months). YouTube reports that, in 2017, 79% of the videos removed for violating their policies were initially flagged by automatic flagging systems and in the second quarter of 2019 this was 87%. A considerable number of the videos removed is taken down before receiving a single view by the users. By April 2019, through using technology, 38% of abusive content that is actioned by Twitter is surfaced proactively for human review instead of relying on users' reports. This marks a significant increase from the previous year , where 20%

of potentially abusive content was flagged by machines. It should be noted that all content surfaced by automatic detection system is assessed by the team of reviewers before being actioned (human-in-the-loop).

What do we know about the volumes of hate speech content flagged to the IT Companies?

From data reported by some of the IT Companies participating to the Code, the amount of notices on hate speech content seems to be in the range of 17-30% of the total[2]. Facebook reports having removed 3.3 million pieces of content for violating hate speech policies in the last quarter of 2018 and 4 million in the first quarter of 2019. In 2018, more than 6.2 million Twitter accounts were signalled for containing hateful conduct and the platform took action on approximately 536,000.

A study carried out by Vox POL for the Commission in 2018 made a comparative analysis of the activity of a group of about 175 "haters" in several Member States: while these produced 60,000 hateful tweets in 2016, their activity today is reduced to 7400 tweets.

The ecosystems of hate speech online and magnitude of the phenomenon in Europe remains an area where more research and data are needed.

[2] It is worth noting that this refers to flags received and does not correspond to the actual number of removals of hateful content. For example, it may happen that pieces of content are wrongly flagged by users as hate speech.