Code of Conduct on countering illegal hate speech online: Questions and answers on the fourth evaluation

Brussels, 4 February 2019

Code of Conduct on countering illegal hate speech online: Questions and answers on the fourth evaluation

See IP/19/805

What is the aim of the Code of Conduct?

The European Commission launched the Code of Conduct in May 2016 together with four major IT companies (Facebook, Microsoft, Twitter and YouTube)and in an effort to respond to the proliferation of racist and xenophobic hate speech online.

The aim of the Code is to make sure requests to remove content are dealt quickly. When companies receive a request to remove content deemed to be illegal from their online platform, they assess th-is request against their rules and community guidelines and, where necessary, national laws transposing EU law on combatting racism and xenophobia. The companies have committed to reviewing the majority of these requests in less than 24 hours and to removing the content if necessary, while respecting the fundamental principle of freedom of speech.

Today, nine companies adhered to the Code, notably Facebook, YouTube, Twitter, Microsoft, Instagram, Google+, Dailymotion, Snapchat and Webedia (jeuxvideo.com).

How does the Commission evaluate the implementation of the Code of Conduct?

The Code of Conduct is evaluated through a monitoring exercise by a network of civil society organisations located in different EU countries. Using a commonly agreed methodology, these organisations test how the IT companies apply the Code of Conduct in practice.

They do this by regularly sending the IT companies requests to remove content from their online platforms. The organisations participating in the monitoring exercise record how long it takes the IT companies to assess the request, how the IT companies' respond to the request, and the feedback they receive from the companies.

What are the main takeaways of the 4th monitoring exercise?

- **Swift response to illegal hate speech notified by users to platforms**

The results of the fourth monitoring exercise, which also includes Instagram and Google+, show that about 89% of the notifications are assessed within 24 hours. The IT companies fully meet the target of reviewing the majority of notifications within 24 hours. Facebook has even reached 92.6% of notifications assessed within 24 hours.

On average, IT companies are removing almost 72% of illegal hate speech incidents notified to them by the NGOs and public bodies participating in the evaluation. Between 70% and 80% is estimated to be satisfactory removal rates, as some of the content flagged by users could relate to content that is not illegal. In order to protect freedom of speech only illegal content should be removed.

- **Consistent and scrupulous assessment of illegal hate speech content**

The average removal rate is higher for more serious cases of deemed illegal hate speech. Content that calls for murder or violent acts against certain groups is removed in 85.5% of cases. Similarly, content likely to be holocaust denial is taken down in 75% of cases. Content using degrading, defamatory words or pictures to name certain social groups or individuals belonging to such groups are removed in 58.5% of the cases. This suggests that the review made by the companies is done with due consideration of protected speech and that there is no sign of over-removal.

- **More efforts needed on transparency and feedback to users**

There are still some gaps in the information provided to users on the outcome of their notifications: the average of notifications which received feedback is slightly lower than last year (65.4% vs. 68.9%). Facebook is the only platform that provides systematic feedback to all users while the other platforms do not yet reach these levels (Twitter, 60.4%, Instagram 41.9%, YouTube 24.6%).

- **Promoting positive narratives of tolerance and pluralism**

Partnerships between civil society organisations, national authorities and the IT platforms have been established on awareness raising and education activities, to promote positive narratives of tolerance and pluralism. A major online campaign will be launched in the coming months at EU level as a result of joint efforts between companies and civil society organisations.

Table with key results:

Content removed	1st monitoring (Dec 2016)	2nd monitoring (May 2017)	3rd monitoring (Dec 2017)	4th monitoring (Dec 2018)
Facebook	28.3%	66.5%	79.8%	82.4%
YouTube	48.5%	66.0%	75.0%	84.5%
Twitter	19.1%	37.4%	45.7%	42.5%
Instagram	-	-	-	70.5%
G+	-	-	-	76%
Overall	**28.2%**	**59.1%**	**70.0%**	**71.7%**

% of notifications assessed within 24h	1st monitoring (Dec 2016)	2nd monitoring (May 2017)	3rd monitoring (Dec 2017)	4th monitoring (Dec 2018)
Facebook	50.0%	57.9%	89.3%	92.5%
YouTube	60.8%	42.6%	62.7%	80.9%
Twitter	23.5%	39.0%	80.2%	88.0%
Instagram				77.7%
G+				47.4%
Overall	**40%**	**51.4%**	**81.6%**	**88.9%**

Has the Code of Conduct delivered on its commitments?

The monitoring results show that since the adoption of the Code, IT companies have strengthened their reporting systems, making it easier to report hate speech, and have improved their transparency vis-à-vis notifiers and users in general. They have increased their staff of reviewers and the resources allocated to content management. This has driven improved responses to hate speech flags/notifications. In 2016, only 40% of notifications were assessed within 24 hours, while today it is 89%. Removals of hate speech content increased from 28% in 2016 to 72% in 2018.

In addition, IT companies have strengthened their cooperation with civil society organisations through dedicated partnerships and programmes, as well as through regular trainings, to ensure a better understanding of reporting systems, national context and legal specificities related to hate speech.

As regards transparency towards the general public, in 2016, IT companies only made information available on the number of law enforcement requests and rarely provided any detail on illegal hate speech as a specific ground for removal. Today, the removals of hate speech content are well presented, on a regular basis, in each of the IT companies' transparency reports. Further progress could still be made however, for example by including a more detailed breakdown.

In terms of feedback to users sending notifications, there is still room for progress. Despite notable differences among the companies, on average almost a third of the notifications does not receive feedback.

More information on the Code's achievements can be found here.

How does the Code of Conduct contribute to the wider work of the Commission on illegal content?

The results of the Code of Conduct monitoring feed into the wider work of the Commission on the role of online platforms in the prevention, detection and removal of illegal content.

On 28 September 2018, the Commission adopted a Communication which provides for guidance to platforms on notice-and-action procedures to tackle illegal content online. The importance of countering illegal hate speech online and the need to continue working with the implementation of the Code of Conduct is reflected in this guidance document.

A Commission Recommendation on measures to tackle effectively illegal content online was published on 1 March 2018. It contains two parts, a general part on measures applicable to all types of illegal

content and a specific part addressing the special actions that platforms would need to take to address terrorist content. In terms of the rules applicable to all types of illegal content the recommendation includes clearer 'notice and action' procedures, more efficient tools and proactive technologies, stronger safeguards to ensure fundamental rights, special attention to small companies and closer cooperation with authorities.

What is the definition of illegal hate speech?

Illegal hate speech is defined in EU law under the Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law as the public incitement to violence or hatred directed to groups or individuals on the basis of certain characteristics, including race, colour, religion, descent and national or ethnic origin.

Is the Code of Conduct the right solution to tackle hate speech online?

The self-regulatory approach set by the Code of Conduct has proven to be an effective policy tool to achieve fast progress by the businesses in facing a major societal challenge. One policy alone cannot be the only solution to tackle the proliferation of hatred online.

The Code of Conduct focuses primarily of notice-and-action and removals and thus helps to treat the "symptoms". The challenges posed by hate speech online need to be tackled in a comprehensive way. Swift response to notices must be combined with actions by:

- national authorities, which should step up enforcement capacity and ensure effective prosecution;
- the IT platforms which have to continue progressing;
- civil society, promoting positive narratives, education programmes and awareness-raising campaigns for tolerance and pluralism.

How does the Commission work with the different IT platforms?

The Code of Conduct is based on cooperation involving the European Commission, IT platforms, civil society organisations and national authorities. All stakeholders meet regularly under the umbrella of the High Level Group on combatting racism and xenophobia, to discuss challenges and progress. In addition to the regular monitoring exercises, the Commission engages in a constant dialogue with the platforms to encourage progress on all the commitments in the Code.

Workshops and trainings are also organised with companies and other relevant stakeholders. For instance, a workshop held jointly with Google in Dublin in November 2017 focused on increasing quality of notices by trusted flaggers to ensure a more effective response by the companies' content reviewers. This workshop was followed up by similar events co-organised with Facebook and Twitter in June 2018 and January 2019 respectively.

Does the Code of Conduct lead to censorship?

No. The Code of Conduct aims to tackle online hate speech that is already illegal. The same rules apply both online and offline. Content that is illegal offline should not be allowed to remain legal online.

In the Code, both the IT Companies and the European Commission also stress the need to defend the right to freedom of expression. The Code cannot be used to make IT Companies take down content that does not count as illegal hate speech, or any type of speech that is protected by the right to freedom of expression set out in the EU Charter of Fundamental Rights.

In addition, the results of a 2016 Eurobarometer survey showed 75% of those following or participating in online debates had come across episodes of abuse, threat or hate speech aimed at journalists. Nearly half of these people said that this deterred them engaging in online discussions. These results show that illegal hate speech should be effectively removed from social media, as it might limit the right to freedom of expression.

Isn't it for courts to decide what is illegal?

Yes, interpreting the law is and remains the responsibility of national courts.

At the same time, IT companies have to act in line with national laws, in particular those transposing the Framework Decision on combatting racism and xenophobia and the 2000 e-commerce Directive. When they receive a valid alert about content allegedly containing illegal hate speech, the IT companies have to assess it, not only against their rules and community guidelines, but, where necessary, against applicable national law (including that implementing EU law), which fully complies with the principle of freedom of expression.

Do all expressions of hatred qualify as illegal hate speech, e.g. "i hate you"?

Offensive or controversial statements or content might be legal. As the European Court of Human

Rights said, "*freedom of expression ... is applicable not only to "information" or "ideas" that are favourably received or regarded as inoffensive or as a matter of indifference, but also to those that offend, shock or disturb the State or any sector of the population*".

In the Code, both the IT companies and the European Commission also stress the need to defend the right to freedom of expression.

Assessing what could be illegal hate speech includes taking into account criteria such as the purpose and context of the expression. The expression 'I hate you' would not appear to qualify as illegal hate speech, unless combined with other statements about for example threat of violence and referring to race, colour, religion, descent and national or ethnic origin, among others.

How can we prevent governments from abusing the Code of Conduct?

The Code of Conduct is a voluntary commitment made by the IT companies that have signed up to it. It is not a legal document and does not give governments the right to take down content. The Code cannot be used to make these IT companies take down content that does not count as illegal hate speech, or any type of speech that is protected by the right to freedom of expression set out in the EU Charter of Fundamental Rights.

MEMO/19/806

Press contacts:

Christian WIGAND (+32 2 296 22 53)
Melanie VOIN (+ 32 2 295 86 59)

General public inquiries: Europe Direct by phone 00 800 67 89 10 11 or by email