5th meeting of the Subgroup on countering hate speech online

Brussels, 12 March 2018

The meeting was chaired by the Commission and attended by Member States representatives (all except Cyprus and Malta), 31 civil society organisations including members of the network monitoring the implementation of the Code of Conduct, the IT Companies (Facebook, representing also Instagram, Google – YouTube and Google+ - Twitter, and Microsoft) and the Council of the EU.

The morning session, open to all stakeholders, was devoted to presenting the 1st March Recommendation on tackling illegal content online and the results of the 3rd monitoring exercise published in January 2018. In the afternoon, two parallel sessions took place: a) Member State authorities and IT Companies discussed how to ensure better cooperation on investigation and prosecution of online hate speech crimes as well as how to enhance the protection of human rights defenders; b) the network of civil society organisations gathered in parallel to discuss how to further improve the monitoring activities and the next steps on the implementation of the Code of conduct (including more broadly on education, counter narratives, etc.).

Morning session - *plenary*

Update on recent steps on countering illegal content online, including illegal hate speech

The Commission introduced the Recommendation on measures to effectively tackle illegal content online adopted on 1st March 2018. In particular, the presentation focused on the relevance of the Recommendation for the work on countering hate speech online and the interplay with the Code of conduct. Some delegates from national authorities (Spain, Slovakia) intervened to support the further steps taken by the Commission which provide for a stronger frame for action on countering hate speech. Some others questioned whether the Commission had not considered limiting the scope of the Recommendation to certain crimes and where the content was manifestly illegal. They also questioned whether the Recommendation *de facto* is leading to a further expansion of the scope (Sweden), including content that may not be illegal *per se* but lead to illegal behaviours (Netherlands). The Commission clarified that the Recommendation does not enlarge the scope and does not interfere with the E-Commerce Directive. As a recommendation it can only contribute to the clarification of responsibilities of online platforms. Neither the E-Commerce Directive nor the Recommendation define what illegal content is – and refer to national legislation in this respect. Google took the floor to support the positive collaboration of the sectorial dialogue as well as to express some concerns about some principles in the Recommendation (i.a. the 1-hour turnaround to respond to flags on terrorist content). The Czech NGO In Iustitita supported the reference in the Recommendation to the role of trusted flaggers but called for better protection when trusted reporters are exposed to threats for the work they carried out. France took the floor to mention initiatives in the pipeline to protect flaggers of illegal content referring to cases where their identity was disclosed to the actual authors of the hate speech content. The flaggers should remain anonymous to avoid threats and hate attacks.

The Commission provided an update on the forthcoming funding possibilities under the 2018 Rights, Equality and Citizenship Programme in the area of combatting hate speech online. CEJI – on behalf of the EU funded project "Facing all the Facts!" - presented the recently released e-learning on hate speech online. The initiative was welcome by the Commission and was suggested as an additional tool to build the capacity of trusted flaggers.

State of play on the implementation of the Code of conduct on countering illegal hate speech online

The Commission summarised the results of the 3rd monitoring exercise on the implementation of the Code of conduct which was carried out between 6th November and 15th December 2017. The results were

presented to the media on 19th January 2018. Main findings are the overall trend of progress on the swift review and removal of content notified while some challenges and gaps persist on both feedback to users and coherence of treatment of notifications depending on the channel used. Too often the assessment changed when the notification was transmitted through channels available only to trusted flaggers.

IT Companies took the floor for a quick self-assessment. Facebook expressed gratitude for the collaboration, including on sharing the raw data, which enabled to check where the response to notices was not adequate: this led to further training for the reviewers and enhanced bilateral dialogue with trusted flaggers. **Facebook** also announced new trusted flaggers enrolled and ten thousands more staff in the content management teams. **YouTube/Google** confirmed that the results of the 3rd monitoring mirror the records of their self-assessment. Progress is needed in particular in a specific group of countries. YouTube also mentioned the very positive impact for their team of reviewers of initiatives such as the workshop with trusted flaggers on online hate speech sponsored by the Commission and hosted by Google in Dublin in November 2017. Similar initiatives should be continued and a follow up event could be envisaged in the summer. YouTube also announced two important steps: a) transparency reports will be modified to include also data on referrals based on the community standards, including a breakdown on notices on hate speech, and will be published quarterly; b) all users will have a dashboard to track their flags and the outcomes. **Twitter** expressed satisfaction for the improvements achieved on removal and time of assessment with respect to previous monitoring exercises. Twitter also informed that an intense dialogue is ongoing to foster further improvements with many of the civil society organisations that participated to the monitoring. In particular, Twitter is investing on stepping up the quality of the information users receive, especially when behaviours are against rules and terms of services, to achieve an overall goal of increased healthy and civil conversations on the platform.

Three organisations (from Sweden, Croatia and Belgium) briefly presented their results and highlighted the persisting challenges from their local/national perspectives. The International Network against Cyber Hate (INACH) took the floor to inform about the recent signature of a Framework Partnership with the European Commission, an important step to enhance the capacity of the work of civil society on countering hate speech online in general as well as on the monitoring of the implementation of the Code. The Austrian organisation ZARA signalled the importance of looking at data broken down per company because often the average removal rate is due to low performance of one IT Company only. ENAR Ireland flagged the difficulties encountered by their members, small grassroots organisations, to report about hate speech content: often they lack the knowledge on reporting mechanisms and the fact of not benefiting from the status of trusted flaggers implies lower feedback which risks to discourage further reporting. Some civil society organisations also pointed to the need of addressing the issue of repeated offenders who continue to spread hate even if their posts are removed. CEJI flagged the importance of looking beyond social media to include also other websites and web platforms. Fundación Secretariado Gitano (Spain) expressed satisfaction from the progress achieved in terms of cooperation with the IT Companies, requesting however that the monitoring results better reflect the amount of hate speech content against Roma communities. Tell MAMA (UK) signalled that often IT Companies limit their action to withholding the content in one national context, but this may not be an adequate response, as the effects of hatred online can spread over the borders and affect victims and communities. On this point, IT Companies responded that certain content may be withheld in a given country typically when it violates local laws but not the company's terms of service. Member State delegates intervened to stress the importance of setting clear criteria for the appointment of trusted flagger to protect free speech (Poland), the importance of moving the attention to understanding the landscape of hatred (Italy) as well as the importance of education and counter narratives (Croatia and Spain).

Afternoon sessions

Session I *(Member States and IT Companies)*

- **Investigation and prosecution of hate speech online**

The session started with a short presentation by the Cyber Data Coalition which brings together a number of small and medium enterprises delivering highly technological services in the areas of automatic detection of illegal content online.

Following the presentation, the Commission introduced the discussion informing that the need to effectively investigate and prosecute hate speech has been addressed in several meetings of this expert group, including the last one in October 2017. Issues of practical difficulties in requesting information to IT platforms were often raised as well as issues related to how to optimise the role of national contact points. While there are high expectations that the forthcoming European legislation on e-evidence will help to ensure a better cooperation and information sharing on online evidence in support of investigation and prosecution of cyber crimes, the intention of the discussion was to focus on what information at the moment IT Companies can disclose and what should be the channels to be used by national authorities to receive effective responses.

Facebook, Twitter and YouTube took the floor for a presentation on their internal policies regarding requests from law enforcement authorities.

Twitter explained the system in place for legal and non-legal take down and information requests. For legal requests Member State authorities are referred to a "Legal request form" online which contains full information on how a request should be formulated. Attention was also drawn to the fact that on Twitter most information that the company has is already public as the accounts are open. In case the account has an associated email, IP address and phone number this information could be requested too. The company has a policy in place that allow them to voluntarily disclose emergency request that involve danger to a person. For non-emergency requests, an MLA should generally be used. Given the fact that this may take time it is advised to combine such a request with a preservation request. Twitter also explained that the user will be informed about the request made in relation to their account unless there is a request from the enforcement authority that such information should not be provided for a period of time, for reasons of investigation.

Google (YouTube) outlined the main request formats and conditions for requests. For preservation orders, they can be issued by any law enforcement authority. The only formality requested is that the demand is made on formal letterhead. The information can be preserved for as long as the investigation goes on. MLA procedures are a diplomatic channel for request whereby an EU law enforcement authority can demand to their counterpart in the US to submit a request on their behalf, provided that such a request would be permitted under US law. EU law enforcement authorities may also submit a direct request to YouTube in respect of non-content related information connected to a user. Such a request would require a simple police or a court order depending on national law. YouTube also responds to emergency requests from EU law enforcement authorities provided that they attest that the information is needed to prevent imminent threat or harm. YouTube also recalled that data disclosure laws including EU data protection laws set certain limits as to what can be disclosed and under which circumstances. Sometimes, if imminent threats are identified, YouTube may report directly to law enforcement.

Facebook explained the importance of transparency in the rules governing requests for access to user data and informed that they have in place special law enforcement request systems and guidelines. The company responds to over forty thousand requests per year only in the EU. All requests are reviewed by humans and disclosure is only provided after verification. In order to be able to respond 24/7 the company has teams across time-zones to deal with emergency request. Like Twitter, Facebook underlined the need

to request a preservation order in parallel to proceeding with the MLA request in order to be sure that the evidence is preserved.

Several Member States intervened after the presentation. Sweden underlined the effectiveness of single points of contact which works well and has shown an 90% response rate. The problem occurs when the IP address is located outside Sweden or where the content is too old or when the information upon which the request is based relates to content that is protected by the first amendment in the US. YouTube confirmed that if the requests cover content or an alleged offence that is covered by the first amendment it will not be released. Belgium questioned why notifications for removal of terrorist content on average leads to a 75 % removal rate while the corresponding figure for hate speech is significantly lower. IT companies advised that the data is sent to them so that they can do a proper assessment. UK underlined the need to be pragmatic and innovative to see how the systems that the IT companies have in place can protect the individual, for instance through security settings that the user selects themselves. Both Twitter and YouTube fully agreed and informed that they are taking measures to allow the user to prompt blockings and filtering of commentary fields based on keywords.

- **Protection of human right defenders engaged on countering hate speech online**

The Commission explained that, as heard during the morning session and in previous meetings, there seems to be serious concerns about the threats and intimidation against those that stand up against racism and xenophobia online, and report illegal hate speech to IT platforms and the police. While the protection against threats and intimidation remains a competence of national authorities, it is considered useful to have an exchange to better understand the dimension of the phenomenon in the various Member States, as well as existing practices which could be shared. UNAR (Italy) took the floor to share the initiative of a joint group initiated by the Ministry of Justice involving other authorities (e.g. the police) and several NGOs including grassroots groups to help activists that have been attacked online. France echoed that there is a problem that hate is increasingly directed towards those that defend minorities and try to tackle hate. The role of law enforcement is of course very important in these situations and the accessibility of e-evidence with the platforms is important to allow law enforcement authorities to offer protection to the victims. Poland indicated that since 7 years, they have a special point of contact in place for hate speech. The contact point has had very good experience when it comes to removal while requests for e-evidence has been more problematic.

The Chair concluded that the discussions revealed the need to focus more attention on the "eco system of hate" by mapping its origin and channels for spreading. This is a matter that the Commission is examining but it would be good if Member States could provide national studies or findings so to identify possible knowledge gaps and bring this issue to the agenda in the upcoming meetings.

Session II *(with civil society organisations part of the monitoring network)* **Debrief from the 3rd monitoring, fine tune of the methodology and next steps**

The Commission introduced the discussion highlighting that the results from the 3rd monitoring are encouraging because there is global trend of progress, across countries and organisations. However, some gaps remain and some organisations are still finding difficult to receive swift responses and to obtain the removal of the content they flag. Another challenge refers to the coherence of treatment of the notifications depending on the reporting channels. In many cases the use of the escalation became necessary because the initial notification was not followed up. Also, feedback to users is still quite problematic in particular for both Twitter and YouTube. Facebook is more systematic but their response is quite standard. The floor was open for a brainstorming discussion on a) what remain the key urgent problems in the response by IT Companies to hate speech notifications; b) what should be improved on the

methodology for the monitoring; and c) what initiatives could be helpful for preparing the ground for next monitoring and to foster improvements – e.g. next workshop with IT Companies. All NGOs took the floor to express their views and proposed actions. The main points raised were: importance of sharing information on concrete cases of content removed (e.g. a database); reviewers in the IT Companies should be trained possibly by trusted flaggers who know context, language and case law; the reporting tools made available to normal users are not sufficiently visible and known; the lack of feedback or – often – of action on the normal users notification is a key problem because it discourages the individuals (and often even the victims) to notify; while the reasons for a lower removal rates may be linked to a series of factors (language capacity of reviewers, context related aspects, unclear hate crime legislation), the monitoring activities could be more structured and ensure that a) certain portions of the notifications concern manifestly illegal content (inciting directly to violence and murder of groups) and b) each organisation concentrates on a balanced number of cases per IT Company to enhance comparability and reliability of the average removal rate. There was agreement amongst the group on the importance of monitoring the response by IT Companies also outside the official monitoring periods. Finally, all organisations stressed the importance of investing on education on the users and of the communities, in particular to boost content reporting.