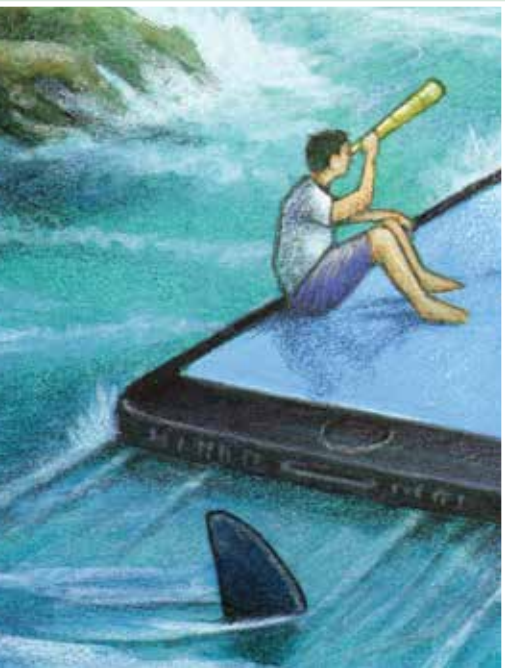
One in a Series of Working Papers from the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression



The European Commission's Code of Conduct for Countering Illegal Hate Speech Online

An analysis of freedom of expression implications

Barbora Bukovská
ARTICLE 19

May 7, 2019

TRANSATLANTIC WORKING GROUP

The Transatlantic Working Group Papers Series

Co-Chairs Reports

Co-Chairs Reports from TWG's Three Sessions: Ditchley Park, Santa Monica, and Bellagio.

Freedom of Expression and Intermediary Liability

Freedom of Expression: A Comparative Summary of United States and European Law
B. Heller & J. van Hoboken, May 3, 2019.

Design Principles for Intermediary Liability Laws
J. van Hoboken & D. Keller, October 8, 2019.

Existing Legislative Initiatives

An Analysis of Germany's NetzDG Law
H. Tworek & P. Leerssen, April 15, 2019.

The Proposed EU Terrorism Content Regulation: Analysis and Recommendations with Respect to Freedom of Expression Implications
J. van Hoboken, May 3, 2019.

Combating Terrorist-Related Content Through AI and Information Sharing
B. Heller, April 26, 2019.

The European Commission's Code of Conduct for Countering Illegal Hate Speech Online: An Analysis of Freedom of Expression Implications
B. Bukovská, May 7, 2019.

The EU Code of Practice on Disinformation: The Difficulty of Regulating a Nebulous Problem
P.H. Chase, August 29, 2019.

A Cycle of Censorship: The UK White Paper on Online Harms and the Dangers of Regulating Disinformation
P. Pomerantsev, October 1, 2019.

U.S. Initiatives to Counter Harmful Speech and Disinformation on Social Media
A. Shahbaz, June 11, 2019.

ABC Framework to Address Disinformation

Actors, Behaviors, Content: A Disinformation ABC: Highlighting Three Vectors of Viral Deception to Guide Industry & Regulatory Responses
C. François, September 20, 2019.

Transparency and Accountability Solutions

Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry
M. MacCarthy, February 12, 2020.

Dispute Resolution and Content Moderation: Fair, Accountable, Independent, Transparent, and Effective
H. Tworek, R. Ó Fathaigh, L. Bruggeman & C. Tenove, January 14, 2020.

Algorithms and Artificial Intelligence

An Examination of the Algorithmic Accountability Act of 2019
M. MacCarthy, October 24, 2019.

Artificial Intelligence, Content Moderation, and Freedom of Expression
E. Llansó, J. van Hoboken, P. Leerssen & J. Harambam, February 26, 2020.

www.annenbergpublicpolicycenter.org/twg

A project of the Annenberg Public Policy Center of the University of Pennsylvania in partnership with
The Annenberg Foundation Trust at Sunnylands and the Institute for Information Law of the University of Amsterdam,
with support from the Embassy of the Kingdom of the Netherlands in Washington, D.C.

TRANSATLANTIC WORKING GROUP

The European Commission's Code of Conduct for Countering Illegal Hate Speech Online

An analysis of freedom of expression implications[†]

Barbora Bukovská, ARTICLE 19[1]

May 7, 2019

This paper, developed within the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression (TWG) project, and informed by the discussions at the TWG's Ditchley Park Session, reviews the freedom of expression implications of the Code of Conduct for Countering Illegal Hate Speech Online that was developed by the European Commission in collaboration with major information technology companies in 2016. The analysis looks into the process that led to the adoption of the Code, the Code's legal basis, and the problems within the system it introduced. The paper also briefly outlines how the European Commission has assessed the implementation of the Code of Conduct and how the Code is reflected on the national level in some EU states. Because a number of countries are proposing new regulatory systems for content moderation, the paper suggests that the experience with the Code of Conduct and its implementation can inform the debates on both the effectiveness and the pitfalls of these proposed regulatory models.

Contents

[†] One in a series: A working paper of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression. Read about the TWG: https://www.ivir.nl/twg/.

Introduction

In recent years, there has been increased scrutiny over the content moderation practices of digital companies, with growing calls for tightening existing regulatory models. Pressure from governments toward online platforms to step up their efforts to remove illegal and harmful content and/or adopt tools to detect and prevent uploading such content automatically is not new. However, in recent months, a number of European countries (such as Germany, the UK, France and Ireland) stated their intentions to introduce statutory regulation in this area. Finding an approach that protects freedom of expression while preventing some of the more egregious abuses of digital communication channels is a challenge, with serious implications for society as a whole. The adoption of specific legislation and other regulations on content moderation by governments can lead to the creation of systems where private actors are tasked with applying criminal and other laws under short deadlines and under the threat of heavy fines. This further fragments legal obligations for social media companies, creates a situation where individual users have little or no remedy to address hastily removed content, and provides insufficient guarantees for the protection of individual freedoms.

Given the dangers of statutory regulation of content moderation, voluntary mechanisms between digital companies and various public bodies represent a less intrusive approach and preferred model. The Code of Conduct for Countering Illegal Hate Speech Online (the Code of Conduct) has been presented to be such a model. Launched on 31 May 2016, the Code of Conduct is the outcome of a series of discussions between the European Commission, Facebook, Microsoft, Twitter and YouTube (IT companies),[2] EU Member states and civil society organizations (CSOs).

According to the European Commission, the Code was developed following the October 2015 EU Colloquium on Fundamental Rights on "Tolerance and respect: preventing and combating anti-Semitic and anti-Muslim hatred in Europe," and the December 2015 EU Internet Forum. The European Commission stated that it was also motivated by an increase in discrimination and stigmatisation of minorities (in particular ethnic and religious minorities, migrants, LGBTQI persons and differently abled people) in Europe.[3] Věra Jourvá, the EU Commissioner for Justice, Consumers and Gender Equality, claimed that the Code of Conduct was inspired by "the need for clearer procedures to prosecute and take down "hate speech" on the internet," and was certain that the Code of Conduct could become a " 'game changer' in countering hate speech online."[4]

However, the Code of Conduct was not enthusiastically accepted by the civil society. The major criticisms focus on:

- The problematic *process of the development* of the Code;
- The *legal basis for the Code* which provides for overly broad definitions of "illegal hate speech";
- The *actual system introduced by the Code*, namely: delegation of enforcement activities from the state to IT companies; the risk of excessive interference with the right to freedom of expression; and a lack of compliance with the principles of legality, proportionality and due process. Some of these practices already have been put in place by the IT companies (such as the use of "trusted reporters"), however, the Code of Conduct appears to "codify" or formalise them;

- The *implementation* of the Code of Conduct and the monitoring mechanism within presents important challenges for evaluating its effectiveness and meeting its stated objectives.

Although it presently appears that the states have little interest in further pursuing this model, the experience of developing and implementing it can highlight the potential problems inherent in new statutory models. Hence, the story of the Code of Conduct and the evaluation of this project offers a useful lesson for all considering statutory regulation of online content moderation.

The TWG used part of its first meeting at Ditchley Park, UK, in February 2019 to discuss the strengths and weaknesses of the Code of Conduct's substantive basis and its mechanism, on the basis of an earlier version of this paper. Subsequently, the paper has been updated accordingly to reflect crucial insights from these discussions. Taking into account these discussions in the TWG, the central conclusions and recommendations are as follows.

Background to the Code of Conduct

Key findings:

- No genuine self-regulation approach;
- No multistakeholder process.

According to the information presented by the European Commission, the Code of Conduct was initiated by the European Commission and developed in consultation with IT companies, EU Member States and civil society.

In theory, the process of developing the Code of Conduct was based on the existing systems for regulation of the media industry in Europe (i.e., self-regulation for the press and co-regulation for the broadcast media).[5] Within these systems, the media industries and stakeholders have developed codes of ethics/standards and the industries subsequently commit to uphold these codes in their practices. These systems also provide a means by which people who feel aggrieved by particular media content can have their case heard without the need to go to a tribunal.

The principle of voluntary compliance is fundamental to models of genuine self-regulation: state authorities should play no role in adjudicating or enforcing the standards set, and those who commit to them do so not under threat of legal sanction but for positive reasons, such as the desire to further the development and credibility of their operations. Moreover, in order to ensure a broad sense of ownership and public trust in the system, the development of such codes should be consultative and transparent, including all stakeholders and the broadest possible representation of civil society.

The Code of Conduct's development was markedly different from these requirements. Although the Code was presented as "voluntary" (i.e., not binding or enforceable), it was developed at the behest of the European Commission under the threat of introducing statutory regulation. The European Commission also set up this system to monitor implementation of the Code.

Importantly, despite the European Commission's claims of a participatory process for developing the Code, available information shows that the Commission shared the text with the 28 Member States just a few days before it was revealed.[6] Hence, national authorities and stakeholders had no

opportunity to comment on the text. Further, despite several references to CSOs in the Code of Conduct, civil society was systematically excluded from the negotiations and there was apparently no involvement of free speech organisations in this process. Two digital rights organizations – EDRi and Access Now – walked out of the discussions on the Code due to the lack of transparency of the negotiations and subsequently stated that they did not "have confidence in the ill-considered Code of Conduct that was agreed."[7] This has severely undermined the credibility of the development of the Code and the Code itself.

Legal basis for the Code of Conduct

Key findings:

- Broad definition of "illegal hate speech";
- Focus on criminal law;
- No guidance for online communications.

The key problem with the Code of Conduct is its normative basis for defining "illegal online content." It specifically refers to the EU Framework Decision on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law (the Framework Decision)[8]. It specifies that the "illegal hate speech" should be understood as per the definition of this term under the Framework Decision[9] and national laws transposing it.

The Framework Decision is, however, a problematic document that has been criticised by civil society and academics for failing to comply with international standards on the right to freedom of expression.[10] The key concerns with the Framework Decision – and by extension with the standards that the Code of Conduct promotes – are as follows.

First, the offences outlined in the Framework Decision go beyond permissible restrictions on freedom of expression under international law. The Framework Decision requires criminalisation of "incitement to hatred," while under international law States are required to prohibit "incitement to discrimination, hostility or violence." The "advocacy of hatred" is the vehicle for incitement, but "hatred" is not, in itself, a proscribed outcome. "Incitement to hatred" makes the proscribed outcome an emotional state or opinion, rather than the imminent and likely risk of a manifested action (discrimination, hostility or violence). Further, the Framework Decision provides for "memory law,"[11] while holding individuals criminally liable for denials of historical events or for the expression of opinions about historical facts is at odds with international freedom of expression standards that call for the repeal of such legislation.[12]

Second, the severity threshold for criminalisation of speech is not specified in the Framework Decision. The individual provisions of the Framework Decision list various types of proscribed conduct[13] but provide little guidance to States on what is considered "particularly serious" to be sanctioned, and how to reconcile these limitations with the right to freedom of expression.[14]

Third, the Framework Decision exclusively focuses on criminalisation of speech, which should be an exceptional and last resort. It mandates the criminal prohibition of a number of speech-related offences and seemingly prefers custodial penalties as sanctions. This potentially violates the principle

of proportionality, as severe penalties are prescribed without requiring consideration of lesser sanctions in the criminal law or alternative modes of redress through civil or administrative law that would be less intrusive vis-à-vis the right to freedom of expression.[15]

Last but not least, the Framework Decision makes no provision on interpreting and implementing the obligations it contains in the context of online communications, giving no guidance to States on how to ensure that the right to freedom of expression should be protected in this context. Ultimately, this is likely to create more legal uncertainty for users and, worse, lead to the application of the lowest common denominator when it comes to the definition of "hate speech." This is concerning since many attempts by States to tackle "hate speech" online have been characterised as misguided.[16]

Issues with the content of the Code of Conduct

Key findings:

- Broad definition of "illegal hate speech";
- No commitment to freedom of expression;
- Lack of due process guarantees;
- Propensity to promote censorship.

Under the Code of Conduct, IT companies agree to take the lead on countering the spread of "illegal hate speech" online by:

- Having in place effective mechanisms to review notifications regarding "illegal hate speech" on their services so they can remove or disable access to such content;

- Having in place Community Guidelines clarifying that they prohibit the promotion of incitement to violence and "hateful" conduct;

- Reviewing the majority of valid notifications for removal of illegal hate speech in less than 24 hours, and removing or disabling access to such content, if necessary.

In particular, the Code of Conduct intends to strengthen notification processes between the companies and law enforcement authorities by channelling communications between them through national contact points on both sides. The role of CSOs as "trusted reporters" of "illegal hate speech" is also highlighted (at the end of the Code of Conduct), with the European Commission and Member States helping to ensure access to a representative network of CSO partners and "trusted reporters."

The Code of Conduct contains further commitments from the IT companies to educate their users about the types of content not permitted under their rules and community guidelines, to share best practices between themselves and other social media platforms, and to continue working with the European Commission and CSOs on developing counter-narratives and counter-hate speech campaigns.

While the Code of Conduct does not put in place any mechanism to monitor the signatories' compliance with it – and indeed is not binding or otherwise enforceable – the IT companies and the European Commission agree to assess the public commitments in the Code on a regular basis. In

addition, the Code of Conduct states that the European Commission, in coordination with Member States, will promote adherence to the commitments set out in the Code to other relevant platforms and social media companies (it however does not specify the process for this).

To some extent, the Code reflects the practices of IT companies that have been in place for some time. For example, Facebook, Twitter, Microsoft and YouTube have long had reporting or "flagging" mechanisms in place. These companies have steadily "tweaked" their Community Guidelines in the last year or so to reflect national legislation and concerns from Member States around hate speech and incitement to terrorism. YouTube or Facebook have been working with "trusted reporters" for some time, though these companies have so far not published any information about who these "trusted reporters" are and how they operate. That IT companies review removal notifications against their Community Guidelines and, where necessary, national laws, is also nothing new. Therefore, in practice, it appears that the Code of Conduct is primarily publicizing and formalising certain aspects of the internal processes that these IT companies already had in place prior to adoption of the Code to deal with complaints about certain types of content.

Simultaneously, the Code of Conduct is problematic in light of international freedom of expression standards.

First, the Code of Conduct encourages the removal of "illegal hate speech" – and the "tweaking" of Terms of Service – by referencing the EU's Framework Decision. As outlined above, there are concerns regarding the compatibility of the Framework Decision with international freedom of expression standards, which are replicated in the standards section of the Code. The Code fails to make it clear that any restriction on free expression should remain the exception rather than the rule, and contains no meaningful commitment to protect freedom of expression. These problems will be further exacerbated if IT companies rely on the Framework Decision, as their assessment of prohibited expression will not meet the international standards either.[17] Moreover, insofar as the Code promotes cooperation with "trusted reporters" or "CSOs," it makes no reference to the need to ensure that free expression groups are consulted in the implementation of the Code of Conduct.

Under the Framework Decision, States are accorded a degree of flexibility in transposing its provisions in national law, including making determinations about what severity threshold speech should meet before being criminalised. In other words, IT companies are encouraged to enforce, via their Terms of Service, widely different legal approaches to "hate speech" across the EU. Further, and in any event, the Code seems to encourage companies to go beyond the requirements of the Framework Decision because IT companies commit to make clear that they prohibit "hateful conduct," i.e., a vague term that could encompass mere vulgar abuse.

Second, the Code of Conduct is problematic because of the lack of due process requirements. It puts companies – rather than the courts – in the position of having to decide the legality of content. It allows law enforcement to pressure companies to remove content in circumstances where the authorities do not have the power to order its removal because the content itself is legal. Importantly, the Code does not require the adoption of any safeguards against misuse of the notice procedure and is silent on remedies to challenge wrongful removals. In particular, it does not include any specific commitments to provide access to an appeal mechanism or other remedy for internet users whose

content has been removed under this system. Content deemed as "illegal hate speech" is taken down within 24 hours and there is no possibility for the "offending" user to contest the removal.

Third, the Code of Conduct seems to indicate that the resources of law enforcement are increasingly devoted to the removal of content such as "hate speech" rather than the investigation and prosecution of those responsible for the allegedly unlawful conduct. In other words, States seem more concerned about the (in)accessibility of content rather than enforcement of the law. While in some circumstances, the removal of content may be a more proportionate alternative than criminal liability, nonetheless, it is indicative of the propensity of States to promote censorship rather than seeking to address the root causes of "hate speech" and the social problems at issue. In practice, it is also likely to be counter-productive as it gives an incentive to individuals engaging in "hate speech" to migrate to other platforms with less restrictive free-speech standards. In the case of suspected terrorists, this is likely to lead to a whack-a-mole game as companies suspend "terrorist" accounts only to see new supporters create new profiles on the same platforms.[18]

Hence, despite the Code's nonbinding character, freedom-of-expression and digital rights organisations – such as EDRi, ARTICLE 19 and the Center for Democracy & Technology[19] – warned that the Code could lead to more censorship by private companies and therefore a chilling effect on freedom of expression on the platforms they run.

Monitoring the implementation of the Code of Conduct

On 14 June 2016, an EU High Level Group was launched to lead the way to the implementation of the Code of Conduct. The group consists of "Member States authorities, key stakeholders including civil society organisations and community representatives… EU agencies, as well as international organisations active in this area."[20] Thus far, the European Commission has issued four reports on monitoring the implementation of the Code – on 1 December 2016 (first results on implementation), 1 June 2017, 19 January 2018, and 30 January 2019.

The *first report*[21] basically presented the results of a six-week exercise (from 10 October 2016 to 18 November 2016). According to the report, 12 organisations based in nine different Member States applied "common methodology" and notified the IT companies of alleged illegal hate speech online and recorded the rates and timing of responses. The details of the methodology were not provided. Some NGOs reported a "success rate" of almost 60%, while others reported only 5% and the "trusted flaggers" had a success between 29-60%. Overall, IT companies' reviewers did not seem to agree with the qualifications made by the flaggers, and asserted that the notified content was not illegal or that it complied with their Terms of Service. The first report did not offer sufficient indications as to why there was disagreement between the flaggers' and companies' qualifications. The report also showed that none of the IT companies responded to notifications within 24 hours, as required by the Code of Conduct.

The *second report*[22] (covering a seven-week period from 20 March to 5 May 2017 and involving 31 organisations and three public bodies from France, Romania and Spain) and third report (conducted in six weeks, with 33 organisations and two public bodies from all EU Member States, except for

Luxembourg) highlighted "significant progress," improvement of "efficiency" and "speed," and "higher quality of notifications."

The second report showed that "2575 notifications were submitted to the IT companies taking part in the Code of Conduct. This represents a fourfold increase compared to the first monitoring exercise in December 2016 ... Facebook removed the content in 66.5% of cases, Twitter in 37.4**%,** and YouTube in 66% of the cases. This represents a substantial improvement for all three companies compared to the results presented in December 2016, where the overall rate was 28.2%."

The *third report*[23] showed that "overall, IT companies removed 70% of the content notified to them, while 30% remained online. This represents a *significant improvement* with respect to the removal rate of 59% and 28% recorded in May 2017 and December 2016 respectively."

The *fourth report*[24] presents the results of a six-week exercise (5 November to 14 December 2018) undertaken by 39 organizations from 26 Member States, except Luxembourg and Denmark, consisting of sending notifications to the IT companies relating to hate speech deemed illegal. The report states that the exercise "used the same methodology as the previous monitoring rounds," although it does not provide any further information about this methodology. The Factsheet from the monitoring highlights as a success the fact that the "removal rate remains stable at around 70%, which is satisfactory as hate speech is not easy to define. Its illegality has to be balanced with the right to freedom of expression and the context."

However, the content of these reports is even more ambiguous than the first one; in particular, they do not provide any details on the improved methodology or on the types of content flagged as "illegal hate speech." It is difficult to assess the actual practices of the IT companies under the Code of Conduct owing to several reasons, in particular:

- First, the reports do not provide data on all removals of "unlawful hate speech" by IT companies. The reports only provide information about the responses by the IT companies to the requests submitted by cooperating organisations within the exercise during the limited time period. They provide no information on whether and how the companies actually adjusted their practices in this area and how they implement the Code of Conduct in general.

- Second, there is no information about the methodology under which the cooperating organisations report the content to the IT companies. It is rather troubling that there have been no commitments from the European Commission to provide further information and clarity on the assessment used, and no commitment to transparency and detailed, disaggregated data. For instance, it is unclear whether the flagging is done based on the flaggers' assessment of the compliance with the domestic criminal law or on their knowledge of and compliance with the respective Community Guidance. The report provides no information on whether the flaggers received training on the existing standards on freedom of expression and the need to balance the removals with the right to freedom of expression.

- Third, there is no information on how the IT companies evaluate the request from the cooperating organisations and how they explain the decisions to reject the recommendations/requests. Such information would be crucial in explaining the assessment

done by different companies, divergence in individual interpretations, and possible criteria used by different stakeholders.

- Fourth, the only criterion of "success" presented by the European Commission in the monitoring reports appears to be the speed and number of the removals. As noted above, the most recent 2019 report comments on improvement of the removal rates and states that only 28.3% of reported content remained online, while "this represents a small increase compared to the 70% one year ago." However, the rate of removals can hardly be considered an indicator of "success"; all it shows is the increase of consensus between the IT companies and the cooperating organizations on what content should be removed. This can be interpreted in several ways. For instance, it might show that the flagging organizations better understand (though the report provides no insight as to how) IT companies' policies and what content might not be acceptable under the respective Community Guidelines. Alternatively, it can indicate that the IT companies simply decided to respond positively to more requests to show good will and desire to comply within the exercise. Another possible interpretation is that over time the IT companies changed their content moderation practices and assess the content differently as compared to few years ago.

- Last but not least, the monitoring reports consist of mere presentation of statistics of removals and statistical information on what grounds was the content removed (e.g. sexual orientation, national origin, Afro-phobia, anti-Semitism and others), with no qualitative assessment whatsoever. There are no "case studies" and examples of the types of content removed and maintained. This is a significant shortfall, given that such information would provide more insight into the assessment and decision-making and changes within the existing process of the IT companies since the adoption of the Code of Conduct.

All in all, the only qualitative conclusion from the four monitoring reports is that there has been a steady increase of removals of the "hate speech" content – as vaguely and broadly defined in the Code of Conduct – within the specific time-period based on requests from specific organizations based on unspecified methodology. Overall, the monitoring reports provide very little information on the real effectiveness of the Code of Conduct system and what impact it has in protecting groups at risk of discrimination and hatred and ensuring that the right to freedom of expression is protected.

Importantly, the EU Member States have undertaken numerous measures to address "online hate speech" in their domestic policies and legislation and, while doing so, have sometimes referred to the EU standards, the Code of Conduct and the need to comply with the Framework Decision (in cases where their legislation provides standards more compliant with international law). For example:

- The Minister for Justice and Safety of the Netherlands recently announced his intention to substantially alter the existing regulations on online "hate speech" in the country; he announced plans to centralize referral and flagging activities into one entity in conformity with EU standards: "The European developments with regard to the tackling of illegal content require a reassessment of our approach to the tackling of illegal content, including the approach to hate speech. … Enabling our current referral units/hotlines to comply with the Commission's demands will require significant investments. At the same time, the handling of

removal requests should be streamlined and standardized. Therefore, the cabinet will focus on bundling the existing expertise into one organization, which can be designed in accordance with Europe's demands. It appears that this organization would not focus solely on hate speech offences, but all forms of illegal content including privacy torts and child pornography."[25]

- In 2016, the influential Irish Law Reform Committee issued a report on Harmful Communications and Digital Safety, in which it found that "action still needs to be taken to implement the 2008 Framework Decision."[26] Ireland lacks criminal-law provisions outlawing the public condoning denial or trivialization of genocide and the Committee recommended

- In the UK, the Digital Economy Act 2017 requires the creation of a Code of Practice with major platforms, which "will seek to ensure that providers offer adequate online safety policies" governing the removal of "inappropriate, bullying or harmful content."[28] Recently, on 31 January 2019, the UK Parliament's Science and Technology Committee published a report on the impact of social media and screen-use on young people's health. This report also mentions hate speech as one of the threats to children online, and argues that platforms should have a "duty of care" to protect children from such harms. It also offers a case study of the German NetzDG law[29] as a model to regulate online harms. Accordingly, they recommend that the Government should "introduce, through new primary legislation, a statutory code of practice for social media companies, to provide consistency on content reporting practices and moderation mechanisms. This should be accompanied by a requirement for social media companies to publish detailed Transparency Reports every six months. Furthermore, when content that is potentially illegal under UK law is reported to a social media company, it should have to review the content, take a decision on whether to remove, block or flag that item (if appropriate), and relay that decision to the individual/organisation reporting it within 24 hours, such as now occurs in Germany."[30]

It appears that the EU States are willing to blur even further the lines between voluntary arrangements and legal safeguards on freedom of expression.

Conclusion and recommendations

Given its problematic legal basis and unclear process of implementation, the Code of Conduct is a misguided policy on the part of the European Commission. For companies, it is likely to amount to no more than a public relations exercise. Despite its nonbinding character, the Code can lead to more censorship by private companies---and thus undermine the rule of law and create a chilling effect on freedom of expression on the platforms they run.

Because the European Commission highlighted the Code of Conduct as part of a series of approaches to address the problem of "online hate speech," it is hoped that this paper will be utilized by the European Commission in its further activities in this area. The European Commission and the IT companies should consider revising the Code of Conduct and ensuring that any similar projects fully comply with the international freedom of expression standards. They should consider all legal

questions and implications for freedom of expression under the Code of Conduct highlighted in this paper, such as the delegation of responsibility for determining what is "unlawful hate speech," vague and overbroad criteria, lack of due process, and redress mechanisms for violations of the right to freedom of expression.

The companies should consider the analysis outlined in this paper in their cooperation with the European Commission and beyond. To address these concerns, they should be more transparent about their content moderation practices, including providing some case studies, i.e., qualitative analysis of their decisions and detailed information about the tools they use to moderate content, such as algorithms and trusted flagger-schemes. The companies should also improve the internal complaints mechanisms, including those used for the wrongful removal of content or other restrictions on their users' freedom of expression. In general, individuals should be given detailed notice of a complaint and be provided with an opportunity for prompt redress. Internal appeal mechanisms should be clear and easy to find on company websites.

The European Commission should revise the Framework Decision and bring it into compliance with international freedom of expression standards.

Official documents

- Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law
- Code of Conduct on Countering Illegal Hate Speech online
- Factsheet summarizing the results of a first monitoring exercise to evaluate the implementation of the Code of Conduct
- Factsheet on the 2nd evaluation of the Code of Conduct
- The result of the 3rd monitoring exercise on the implementation of the Code of Conduct
- Factsheet - 4th monitoring round of the Code of Conduct and 2019 Factsheet - How the Code of Conduct helped countering illegal hate speech

Notes

[1] Dr. Barbora Bukovská is a Senior Director for Law and Policy at ARTICLE 19: Global Campaign for Free Expression, and international freedom of expression organization. She leads the development of ARTICLE 19 policies, including those related to digital technologies, and provides legal oversight across the organization. Contact: barbora@article19.org.
[2] According to the European Commission, in 2018, Instagram, Google+, Snapchat and Dailymotion announced "the intention to join the Code of Conduct," see European Commission, Countering illegal hate speech online #NoPlace4Hate, 15 January 2019.
[3] European Commission, Press Release: Speech by Commissioner Věra Jourvá at the launch of the EU High Level Group on Combating Racism, Xenophobia and Other Forms of Intolerance, 14 June 2016, available at https://bit.ly/1XU6wbC. See also European Commission, Press Release: European Commission and IT Companies announce Code of Conduct on illegal online hate speech, Brussels, 31 May 2016.

[4] Speech by Věra Jourvá, *Ibid.*
[5] For an overview of different models of regulation, see e.g., ARTICLE 19, Self-regulation and "hate speech" on social media platforms, 2018; available at https://bit.ly/2O3wztM.
[6] EDRi's Freedom of information request to DG Justice on the Code of Conduct against Hate Speech and responses from the Commission (Ask the EU, 28 April 2016 and 28 July 2016).
[7] EDRi, EDRi and Access Now Withdraw from the EU Commission IT Forum discussions, EDRi, 31 May 2016, available at https://bit.ly/2vDkPre.
[8] Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law.
[9] Article 1(1)(a) of the Framework Decision requires the criminal prohibition of "publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin." Article 1(1) limits the offence to intentional conduct, Article 1(1)(b) clarifies that the offence can be committed through the dissemination of any material, and Article 1(2) allows States to opt to punish "only conduct which is either carried out in a manner likely to disturb public order or which is threatening, abusive or insulting." Article 3 prescribes "effective, proportionate and dissuasive criminal penalties," with mandatory custodial sentences of between 1 and 3 years.
[10] For more information about the compliance of the Framework Decision with international freedom of expression standards, see, e.g., ARTICLE 19, Submission to the Consultations on the European Union's justice policy, December 2013, available at https://bit.ly/2ZOdbZ5.
[11] The Framework Decision requires States to criminalise "publicly condoning, denying or grossly trivialising" specific international crimes recognised under international humanitarian law.
[12] *Cf.* the UN Human Rights Committee, General Comment No. 34, CCPR/C/GC/3, at para. 49; the UN Committee on Elimination of Racial Discrimination, General Recommendation No. 35, *op. cit.*, para. 14; or Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/67/357, 7 September 2012, para 55. These standards stipulate that it is undesirable for States to interfere with the right to know and the search for historical truth by tasking itself with promoting or defending an established set of "historical facts"; and it should be the role of free and open debate to establish historical truths, and not the role of States. Moreover, any instance of incitement committed by way of condoning, denying or trivializing a crime committed against a protected group of people may, where necessary, be prosecuted through standalone provisions on incitement, or alternative provisions within the civil or administrative law. It should be also noted that the jurisprudence of the European Court of Human Rights on this topic is complex and often not consistent with these standards; *cf.,* e.g., *Garaudy v. France* (App. No. 65831/01, 24 June 2003), *Chauvy and Others v. France* (App. No. 64915/01, 29 September 2004) and *Lehideux and Isorni v. France* (App. No. 55/1997/839/1045, 23 September 1998.
[13] On the one hand, Article 1(1)(a) of the Framework Decision requires the criminal prohibition of "publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, color, religion, descent or national or ethnic origin." Article 1(1) limits the offence to intentional conduct, Article 1(1)(b) clarifies that the offence can be committed through the dissemination of any material. Article 1(2) of the Framework Decision allows States to choose to limit the scope of the obligation to prohibit incitement to circumstances where a public order disturbance is likely, or where the language at issue is threatening, abusive or insulting. At the same time, the Preamble of the Framework Decision provides that it is limited to combating "particularly serious" forms of racism and xenophobia.
[14] Set in Article 7 of the Framework Decision. This reveals how broad the obligation is under Article 1(1)(a) of the Framework Decision for States that do not exercise this option.
[15] The Human Rights Committee have stated that restrictions on the right to freedom of expression "must be the least intrusive instrument amongst those which might achieve their protective function", see General Comment No. 34, *op. cit.*, para 34.
[16] The 2012 Report of the Special Rapporteur on freedom of expression, *op. cit.*, para. 32.
[17] There is a growing body of recommendations from international and regional human bodies that social media companies have a responsibility to respect human rights. *Cf.* The Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework (which recognize the responsibility of business enterprises to respect human rights, independent of State obligations or the implementation of those obligation); reports of the UN Special Rapporteur on Freedom of expression to the Human Rights Council, May 2011 and June 2013; or Committee of Ministers of the Council of Europe March 2018 Recommendation on the roles and responsibilities of internet intermediaries.
[18] *Cf.* Brittan Heller, Combating Terrorist-Related Content Through AI and Information Sharing, TWG, April 2019.
[19] *Cf.* EDRi, Guide to the Code of Conduct on Hate Speech, 3 June 2016, available at https://bit.ly/1tbOF34; ARTICLE 19 EU: European Commission's Code of Conduct for Countering Illegal Hate Speech Online and the Framework

Decision, 20 August 2016, available at https://bit.ly/2o4Xdsb; and CDT, Letter to European Commission on Code of Conduct for "Illegal" Hate Speech Online, 3 June 2016, available at https://bit.ly/2DSN5e6.

[20] Each State Member has a designated authority, the EU Agency for Fundamental Rights participates as an EU agency, the European Commission against Racism and Intolerance, and the Office for Democratic Institutions and Human Rights participate as international organisations. The civil society members are Amnesty International European Institutions Office, European Network Against Racism, Open Society European Policy Institute, Platform of European Social NGOs, and the European Region of the International Lesbian, Gay, Bisexual, Trans and Intersex Association; available at https://bit.ly/2KDl6Bl.

[21] European Commission, Code of Conduct on countering illegal hate speech online: First results on implementation, December 2016. https://bit.ly/2I7qElo.

[22] European Commission, Code of Conduct on countering online hate speech – results of evaluation show important progress, 1 June 2017, available at https://bit.ly/2WliWLy.

[23] Results of Commission's last round of monitoring of the Code of Conduct against online hate speech, 19 January 2018, available at https://bit.ly/2GZe600.

[24] European Commission, How the Code of Conduct helped countering illegal hate speech online, February 2019, available at https://bit.ly/2HOWHLl.

[25] The letter of Minister Grapperhaus to the Lower House about the approach to online hate speech, 21 December 2018, available at https://bit.ly/2Se7WwO.

[26] Law Reform Commission, Report Harmful Communications and Digital Safety (LRC 116-2016), 2016, p. 119, available at https://bit.ly/2EliJ4F.

[27] *Ibid.*

[28] Government response to the Internet Safety Strategy Green Paper May 2018, P. 15, available at https://bit.ly/2IWsSZh.

[29] For more information about NetzDG, see Dr. Heidi Tworek and Paddy Leerssen, An Analysis of Germany's NetzDG Law, TWG, April 2019.

[30] The Science and Technology Committee, Social media companies must be subject to legal 'duty of care,' 31 January 2019, available at https://bit.ly/2SezSjY.