# Self-Regulation of Fundamental Rights? The EU Code of Conduct on Hate Speech, related initiatives and beyond[†]

*Teresa Quintel and Carsten Ullrich*[*]

## I. Introduction

There is no formal and straightforward definition on what constitutes illegal hate speech, however, hate speech might be classified as targeting minority groups in a way that promotes violence or social disorder and hatred.[1] The use of social media and online platforms to spread illegal content and hate speech has increased progressively during recent years, as content may be disseminated anonymously and further shared by other users. Therefore, the timely removal or blocking of access to illegal content is essential to limit the wider dissemination and harm of individuals targeted by hate speech. However, the blocking or removal of content may interfere with the fundamental rights of online users and must be assessed diligently and on a case-by-case basis.

The prominent role of online platforms in revolutionising modern communication and as influencers of the public opinion has increasingly come to the attention of policy makers. Since online platforms provide an important stage for phenomena such as fake news, hate speech or disinformation, the pressure to take more responsibility for content hosted by them has grown. Yet, concerns regarding possible limitations of the right to freedom of expression and information, where online platforms providers are required to adjudicate on the legality of content by blocking or removing it, often done through the use of automated systems, are intricate, particularly in the context of hate speech. Other rights, such as the rights to privacy and data protection may be affected where personal data are being processed by automated means or disclosed to third parties.

The EU Commission took action via several attempts to set certain rules for online intermediaries, mostly relying on non-binding agreements, often in the form of self-regulatory measures, such as codes of conduct, guidelines and recommendations.

In May 2016, the Commission published a Code of Conduct on countering illegal hate speech online.[2] Under the Code, several IT companies committed to the establishment of community guidelines regarding measures to be taken to tackle illegal hate speech. According to the Code, the companies should review notifications on illegal content received by users within a certain timeframe and should set up procedures to block or remove such content. In addition, the companies should contribute to raising the awareness of their staff and intensifying cooperation between themselves and other platforms. The community guidelines should provide information concerning the companies' rules on reporting and notification processes for blocking or taking down illegal content.

Other instruments for combatting illegal hate speech online are already existent. The Code of Conduct builds upon the Framework Decision on Racism and Xenophobia,[3] Article 14 of the E-Commerce



[†] This is article is an advance publication. It will appear as a Book Chapter in: Fundamental Rights Protection Online: the Future Regulation of Intermediaries, Edward Elgar Publishing, forthcoming Summer/Autumn 2019, Bilyana Petkova & Tuomas Ojanen (eds)

[*] Teresa Quintel is a FNR funded PhD Candidate at the Université du Luxembourg and at Uppsala University under the supervision of Prof. Mark D. Cole and Assistant Prof. Maria Bergström. Contact: teresa.quintel@uni.lu. Carsten Ullrich is a PhD candidate in law under the supervision of Prof. Mark Cole at the Doctoral Training Unit on Enforcement in Multi-Level Regulatory Systems (DTU REMS), Faculty of Law, Economics and Finance (FDEF), University of Luxembourg. Contact: carsten.ullrich@uni.lu.

[1] Thomas Davidson, 'Automated Hate Speech Detection and the Problem of Offensive Language', ICWSM, 11 March 2017.

[2] 'Code of Conduct on Countering Illegal Hate Speech Online' (2016) <http://ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf> accessed 9 March 2017.

[3] 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, OJ L 328, 6.12.2008.

Directive[4] (e-Commerce Directive) provides the basis for takedown procedures, and the scope of the proposed recast of the Audio-Visual Media Services Directive (AVMSD)[5] was extended to include video-sharing platforms (VSPs) in the fight against hate speech.[6] As a binding instrument open to signature, the additional protocol to the Council of Europe Cybercrime Convention[7] envisages the criminalisation of acts of a racist and xenophobic nature committed through computer systems.[8]

With the Code, the Commission opted for a model that is based on self-regulation. Whether it is a wise approach to entrust private companies with the task of checking posts against national laws on illegal content and bestow them with the responsibility to decide on the removal of controversial content is questionable.

The Code of Conduct was complemented by a Communication in September 2017[9] and a Recommendation[10] that the EU Commission published on 1 March 2018. Both initiatives originally target the fight against all kinds of illegal content on platforms on a horizontal level. Albeit being more detailed and explanatory than the Code of Conduct, both documents encourage self-regulatory measures, just like the Code. The Communication on disinformation from April 2018, which was complemented by a Code of Practice later that year will be a further example of self-regulation, albeit covering a different field than the aforementioned measures. These four instruments will be discussed in more detail below with regard to their implications on the fundamental rights of internet users.

This chapter will give a brief overview of EU legislation encouraging self-regulation, particularly in the form of codes of conduct (section II.). In section III., the Code of Conduct on illegal hate speech online will be illustrated and further commented on in relation to the Communication of September 2017 and the Recommendation that was issued by the Commission on 1 March 2018 (sections III.1. and III.2.). Section III.3. will address the Commission's Communication on disinformation from April 2018, followed by a brief discussion of the Code of Practice on disinformation that was published in September 2018 (section III.3.1). The interim conclusion will demonstrate both the benefits of the current self-regulatory options and the shortcomings of the non-binding rules that have been put forward by the Commission. Section IV will briefly comment on attempts by national legislators to impose binding legal rules in an area so far characterised by industry agreements. Before concluding, Section V will propose an alternative approach towards fighting illegal content on online platforms, which ventures squarely into co-regulation. The current initiatives, discussed in the preceding sections, it is argued, are ill-fitted to ensure the transparent protection of fundamental rights while effectively combating illegal hate speech and infringing content in general on the internet.

## II. EU legislation promoting self-regulation through codes of conduct

Self-regulation has emerged as the dominant form of regulation in the online environment today. There are various reasons for this. The EU defines self-regulation as "the *possibility for economic operators, the social partners, non-governmental organisations or associations to adopt amongst themselves and*

[4] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market, OJ L 187 2000.
[5] Proposal for a Directive of the European Parliament and of the Council amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audio-visual media services in view of changing market realities, COM(2016) 287 final 2016.
[6] Proposed AVMSD amendment (n 5), pp. 5, 9.
[7] Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems, Strasbourg, 28.1.2003.
[8] Article 1 of the Additional Protocol to the Cybercrime Convention.
[9] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, 'Tackling Illegal Content Online: Towards an enhanced responsibility of online platforms', COM(2017) 555 final, Brussels, 28.9.2017.
[10] Commission Recommendation of 1.3.2018 on measures to effectively tackle illegal content online, C(2018) 1177 final, Brussels, 1.3.2018.

*for themselves common guidelines at European level (particularly codes of practice or sectoral agreements)"*.[11] There are several reasons for the prevalence of self-regulation as a governance tool of the internet today, which shall not be discussed in detail here. It shall be sufficient to state that the revolution in information communications technologies (ICT), of which the internet is but one product, have profoundly challenged traditional approaches to jurisdiction, enforcement and regulatory skill sets.[12] Since the early days of the internet, regulators worldwide - including in the EU - took a more hands-off approach towards regulating this new sphere. Meanwhile, the EU also developed its own approach towards self- and co-regulation and came to adopt these forms of Governance into its wider regulatory toolset. It sees both tools as being in compliance with the Treaties and as a contributor to simplifying and improving the EU regulatory environment in line with the Lisbon Agenda.[13] Co-regulation is applied throughout a variety of areas, such as product safety standards, food safety, environmental protection, financial services, or data protection.[14] During the 1990s, steps towards building an overarching approach of codes of conduct on the EU level were developed in the area of the protection of minors within the broadcasting sector. In a Green Paper on the Protection of Minors and Human Dignity in Audio-visual and Information Services[15] from 1996, the Commission called upon national Governments to enhance cooperation with the industries by drawing up codes of conduct.[16] The first formal instrument to advocate the production of rules incorporated into codes of conduct within a national self-regulatory framework was Council Recommendation 98/560/EC.[17] Under No. 2 of the Annex to the Recommendation, codes of conduct were to be implemented on a voluntary basis by the operators concerned and aimed at covering the protection of minors[18] and the protection of human dignity[19].

The e-Commerce Directive[20] seeks to approximate the national provisions on information society services relating to the internal market through the establishment of codes of conduct.[21] These codes of conduct shall contribute to the proper implementation of the rules provided for by the Directive and shall, *inter alia*, define guidelines on the protection of minors and human dignity. The e-Commerce Directive is also relevant in terms of service providers' liability regarding the online content hosted. It affords these companies wide-reaching immunity from liability for content hosted by them if they remove it expeditiously after gaining actual knowledge of it. It also prohibits Member States from imposing general monitoring obligations on these online platforms.[22]

11 Interinstitutional agreement on better law-making, OJ C 321/01 2003, para 22.

12 For a more in-depth discussion see for example: Julie E Cohen, 'The Regulatory State in the Information Age' (2016) 17 Theoretical Inquiries in Law 369; Frank Pasquale, 'Platform Neutrality: Enhancing Freedom of Expression in Spheres of Private Power' (2016) 17 Theoretical Inquiries in Law 487; Christopher T Marsden, *Internet Co-Regulation: European Law, Regulatory Governance and Legitimacy in Cyberspace* (Cambridge University Press 2011); Carsten Ullrich, 'A Risk-Based Approach towards Infringement Prevention on the Internet: Adopting the Anti-Money Laundering Framework to Online Platforms' (2018) 26 International Journal of Law and Information Technology 226.

13 EU Commission, 'European Governance - A White Paper, COM(2001) 428 Final' 20; EU Commission, 'Action Plan "Simplifying and Improving the Regulatory Environment" COM(2002) 278 Final' 11–12 <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2002:0278:FIN:EN:PDF> accessed 31 August 2018.

14 LAJ Senden and others, *Mapping Self-and Co-Regulation Approaches in the EU Context'': Explorative Study for the European Commission, DG Connect* (European Commission 2015) <https://dspace.library.uu.nl/handle/1874/327305> accessed 19 September 2017.

15 Commission of the European Communities, 'Green Paper on the Protection of Minors and Human Dignity in Audio-visual and Information Services', COM(96)483 final, Brussels, 16.10.1996.

16 Ibid, p. 4, 24.

17 Recommendation 98/560/EC[17] on the development of the competitiveness of the European audio-visual and information services industry by promoting national frameworks aimed at achieving a comparable and effective level of protection of minors and human dignity.

18 Point 2.2.1. of the Council Recommendation.

19 Point 2.2.2. of the Council Recommendation.

20 Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market, OJ L 187 2000.

21 Articles 1(2) and 16 of the e-Commerce Directive.

22 Articles 14 and 15 of the e-Commerce Directive.

In the 2012 Cloud Computing Communication,[23] the Commission stated that it would '*work with industry to agree a code of conduct for cloud computing providers to support a uniform application of data protection rules'*.[24] In the area of data protection, codes of conduct were also encouraged by the 1995 Data Protection Directive[25] and now under the regime of the General Data Protection Regulation (GDPR),[26] which repealed the Directive from 1995 in May 2018.

Moreover, the amended Article 4(7) of the proposed recast of the Audio-Visual Media Services Directive (AVMSD)[27] encourages self-regulation through the setting up of codes of conduct within the scope covered by the Directive. The AVMS Directive is interesting insofar as it also contains provisions on the protection of individuals against content containing incitement to violence or hatred. Article 28a(b) of the Directive requires Member States to ensure that VSPs take appropriate action to protect all citizens from content containing incitement to violence or hatred.

In May 2016, the Commission published the abovementioned Code of Conduct on illegal hate speech online that was followed by a Communication towards an enhanced responsibility of online platforms to tackle illegal content online and a Recommendation on measures to tackle such content effectively.

On 26 September 2018, the Commission released a Code of Practice[28] on Disinformation, aiming to achieve several objectives that were set out in a Communication on tackling online disinformation from April 2018[29]. With the Code, the Commission seeks to fight online disinformation by facilitating the verification of ad placements, ensuring transparency for political advertising and issue-based content, and by enabling users to identify promoted content better. Moreover, platform providers should detect bot-driven interactions and close fake accounts. According to the Code, users should have access to various news sources in order to recognise disinformation and to receive different viewpoints. Finally, researchers should be granted access to platform providers' relevant data for monitoring purposes.[30]

Codes of Conduct and self-regulation are promoted elsewhere. On the level of the Council of Europe (CoE), the Committee of Ministers adopted a Recommendation on self-regulation concerning cyber content and user protection against illegal or harmful content,[31] in September 2001, encouraging the establishment of codes of conduct in the field of media, in particular on new communication and information services.

The CoE Policy guidelines on integrated national strategies for the protection of children from violence[32] make reference to the 2001 Recommendation and encourage "*all institutions, services and facilities responsible for the care and protection of children*" to adopt codes of conduct "*incorporating*

[23] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, 'Unleashing the Potential of Cloud Computing in Europe', COM(2012) 529 final, Brussels, 27.9.2012.
[24] https://ec.europa.eu/digital-single-market/en/news/data-protection-code-conduct-cloud-service-providers
[25] Article 27 of Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, [1995] *OJ L 281*, p. 31–50. No longer in force, date of end of validity: 24/05/2018. Repealed by Regulation (EU)2016/679.
[26] Under Article 40 of the General Data Protection Regulation, national supervisory authorities, the European Data Protection Board and the Commission are supposed to formulate codes of conduct to ensure the proper application of the data protection rules under the GDPR. Article 40 refers to the preparation of Codes of Conduct *inter alia* with regard to the fairness of processing, the exercise of data subject rights, the protection of minors, or the communication of data breaches.
[27] Proposal for a Directive of the European Parliament and of the Council amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services in view of changing market realities, COM(2016) 287 final 2016.
[28] https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation
[29] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, 'Tackling online disinformation: a European Approach', COM(2018) 236 final, Brussels, 26.4.2018. (Hereafter 'Communication on Disinformation').
[30] The final Code of Practice is due to be released in September 2018.
[31] Recommendation REC (2001) 8 of the Committee of Ministers on self-regulation concerning cyber-content.
[32] Council of Europe Policy guidelines on integrated national strategies for the protection of children from violence, November 2009, https://rm.coe.int/168046d3a0.

*the prohibition, prevention and rejection of all forms of violence against children*".[33] Moreover, Recommendation CM/Rec(2009)1 on electronic democracy[34] proposes self-regulation for implementation and reviewing purposes.

## III. The Code of Conduct on countering illegal hate speech online

The Code of Conduct was published on 31 May 2016 in order to counter the spread of illegal hate speech online and was committed to by four IT companies, namely Facebook, Twitter, YouTube, and Microsoft.[35] With the Code, the companies agreed upon reviewing the majority of notifications received by online users in less than 24 hours and to implement procedures to remove notified content when considered illegal. Moreover, the companies committed to putting in place community guidelines in which they should set forth the prohibition of incitement to violence and hateful conduct on their platforms. According to the Code, notified content should be evaluated along these guidelines as well as relevant national legislation (where necessary) on hate speech and assessed by dedicated reviewers who would then decide on whether to remove or block the notified content.

In addition, the companies should contribute to raising the awareness of their staff, intensifying cooperation between themselves and competent authorities and providing information concerning their rules on reporting and notification processes for illegal content. Sharing information on the procedures for submitting notices with Member State authorities should improve the effectiveness of communication channels set up between the companies and competent national authorities and contribute to familiarising the latter with methods to detect illegal hate speech online. Civil society organisations should undertake the role of so-called "trusted reporters" who would, based on their particular expertise, notify the platform providers of potentially illegal content. According to the Commission, the knowledge of trusted reporters could contribute to higher quality notices and faster takedowns of illegal content.[36] Further, the companies should contribute to raising awareness among their users by promoting counter-narratives, delivering effective counter-hate speech campaigns and supporting educational programs to encourage critical thinking.

The Code of Conduct is built upon the Framework Decision on Racism and Xenophobia from 2008, which defines hate speech as "[a]*ll conduct 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"*.[37]

However, the definition of hate speech used under the Framework Decision is very broad and gives Member States wide discretion when transposing the provisions. This has led to a different threshold concerning the content that is to be criminalised in the Member States, which complicates any harmonised application of the Framework Decision or the Code of Conduct. Legal provisions prohibiting hate speech may be interpreted loosely and applied selectively, so that there are diverging legal requirements for online platforms in the different Member States. Thus, worldwide operating companies have to consider different national provisions, traditions and societal circumstances, which renders the entire process very complex and shifts the tasks of national courts over to private actors.

In addition, it is rather unclear what constitutes a "valid notification". The Code solely defines such a notification as not "insufficiently precise or inadequately substantiated". Where the companies receive a notification to remove allegedly illegal content, they may review the notification against their

[33] Ibid, p.23.
[34] Recommendation CM/Rec(2009)1 of the Committee of Ministers to Member States on electronic democracy (e-democracy) (Adopted by the Committee of Ministers on 18 February 2009 at the 1049th meeting of the Ministers' Deputies).
[35] Following the announcement by Instagram and Google+ on 19 January 2018, on 7 May 2018, Snapchat also announced the intention to participate to the Code of Conduct on countering illegal hate speech online. On 27 June 2018, Dailymotion announced the company's participation to the Code of Conduct. It is the first Europe based IT Company joining this work.
[36] COM(2017) 555 final, p. 8.
[37] Article 1(a) of the Framework Decision.

community guidelines and shall only "where necessary" review the notification against the national laws transposing the 2008 Framework Decision. Hence, where content is being removed based on the community guidelines, a review against the law could become superfluous. Consequently, notified content would not have to be forwarded to national competent authorities whenever it has been blocked or deleted based on the community guidelines rather than as a criminal offence, as the companies are merely "encouraged" to report illegal content.

Moreover, the Code does not mention appeal mechanisms where users' content is being removed, even if it has not been proven illegal. Although some of the online platforms provide for counter-notices, through which users are being notified when their content has been taken down and may subsequently file a complaint to restore their content, these counter-notices mainly refer to copyright infringements.[38] Thus, under the Code, affected content providers of alleged hate speech have limited choices for seeking a remedy once their content has been blocked or removed.

The Code of Conduct neither mentions safeguards for the freedom of expression nor for the protection of personal data of users where algorithms are used for the detection of illegal content. The diligence and efficiency with which illegal content may be assessed and reviewed is often dependent on the financial capacity and resources of the companies. Whereas companies such as Facebook or YouTube have put in place user notification mechanisms and make available reports on requests for removals and content that has been taken down, smaller companies (although not included in the Code of Conduct) might not have the financial capacity to put such systems in place. Moreover, companies that do not have sufficient resources nor staff at their disposal might struggle to provide an adequate review of notices and counter-notices.

With the time limit of 24 hours suggested by the Code of Conduct, companies might either be unable to detect illegal content on time and therefore, not comply with their blocking obligations, or choose to "over-block" in order not to be held liable and to avoid penalties.

Since, according to the Code, it is the companies that are "taking the lead" on countering the spread of illegal hate speech online, public authorities are only marginally (if at all) involved in the decision-making process. Unlike the judiciary, evaluations by private intermediaries are opaque and their reports often lack the insights that would be needed for a thorough evaluation of the notice-and-takedown procedures. The outsourcing of the assessment on what constitutes illegal content and hate speech to private actors might not only undermine the fundamental right to freedom of expression, but also the right to privacy and data protection, or the right to an effective remedy and to a fair trial.

The third evaluation[39] of the Code of Conduct was released in January 2018,[40] showing that, on average, 70% of content that was notified as being illegal had been removed within 24 hours by those IT companies that had agreed to commit to the Code. However, the Commission had to admit that there was still a lack of users' feedback and that further improvements regarding transparency were needed. Yet, transparency on notice-and-takedown procedures should be regarded as one of the most crucial elements in assessing the effectiveness of the Code of Conduct. Without having an insight into the companies' internal procedures for blocking and removal measures, the evaluation becomes somewhat ineffective.

This shows that the self-regulation of fundamental rights by large corporations and the delegation of enforcement activities from state authorities to private companies may lead to issues where the community guidelines of these companies are elevated above the level of the law. Hence, where content

[38] See, for instance, the guidelines for counter-notifications of Facebook, Twitter or YouTube.
[39] A first evaluation by the Commission took place on 7 December 2016 and a second on 1 June 2017.
[40] European Commission Fact Sheet, 'Code of Conduct on countering illegal hate speech online: Results of the 3rd monitoring exercise', January 2018.

is being blocked or deleted based on community guidelines, the risk of an excessive interference with the right to the freedom of expression and information increases: where companies may be held liable for not removing illegal content within a certain timeframe, they might decide to block content that is not illegal in order to be "on the safe side".
The non-binding nature of the Code of Conduct permits online platforms to remain vague and to rely on agreements with little regulatory oversight. Requiring online platforms to remove illegal content might lead to private censorship where such an assessment is normally carried out by independent courts that provide for appeal mechanisms. In addition, and with regard to rising user awareness, the simple removal of a post or the temporary blocking of an account seems to be less effective than imposing a fine or court proceedings.

## III.1. The Communication on Tackling Illegal Content Online

In its Communication from 28 September 2017, the Commission acknowledges that a harmonised and coherent model to remove illegal content does not exist at present in the EU and that a more aligned approach would be needed in order to make the fight against illegal content more effective.[41]

The Communication thoroughly explains how illegal content may be detected by online platform providers, sets out how such content may be removed expeditiously and reported to law enforcement authorities and clarifies that removed content should be prevented from being uploaded again.

For the detection and removal of illegal content, the Communication suggests, *inter alia*,[42] the implementation of "proactive measures" by the online platforms, namely the use of algorithmic filtering of content stored by them.[43] According to the Communication, the liability exemption under the e-Commerce Directive would solely apply to service providers that merely host information of third parties without actually having control over the content.[44] It is argued that, with the use of proactive measures, the platform providers would obtain knowledge about illegal content and may therefore be excluded from the liability exemption under Article 14 of the e-Commerce Directive.[45] However, that would only be the case if, once they have gained that knowledge, the platforms would not remove the content in question expeditiously. The Commission therefore assures companies that if proactive measures against illegal content are voluntarily implemented, this would not necessarily lead to the loss of their liability exemptions.[46] To that end, online platforms "*should do their utmost to proactively detect, identify and remove illegal content online*" voluntarily and invest in the development and use of automatic detection technologies.[47]

Full automatization is also suggested for deletion measures where illegal content is being notified to the online platforms by law enforcement authorities or in cases where illegal content has been removed previously and is, thus, known by the respective platform(s).[48]

For the removal of particularly harmful content, the Communication suggests expeditious deletion in accordance with specific timeframes, thereby referring to the Code of Conduct and its 24-hour removal period. Moreover, notifications received by trusted flaggers would authorise quicker removal, as the quality of the notice would be higher due to the expert knowledge of these flaggers.[49] In order to prevent

[41] COM(2017) 555 final, p. 5.
[42] In addition to orders by Courts and competent authorities, notices by trusted flaggers and notices by users.
[43] COM(2017) 555 final, p. 10 and 12.
[44] Citing Recital 42 of the e-Commerce Directive and *Google France*, 114 and 120; Judgment of 12 July 2011, Case C-324/09, *L'Oréal v eBay*, para. 113.
[45] COM(2017) 555 final, p. 12.
[46] Ibid. p.13
[47] Ibid, p. 13.
[48] Ibid, p. 14.
[49] Ibid.

the re-appearance of illegal content, the Communication suggests automatic upload filters, for instance in the form of a "database of hashes" and automatic stay-down procedures.[50]

The Communication puts a specific focus on measures to enhance transparency in order to explain the policies for blocking or removing content and the possibility to contest removal decisions. In that context, the Commission encourages the online platforms to publish, at least once a year, transparency reports containing information about the amount of notices received, the measures that were taken and the time needed for action.[51]

Compared to the Code of Conduct, the Commission Communication from September 2017 may be seen as a "step-up". The Communication is much more detailed than the Code, as it clarifies and extends the latter's content. For instance, the Communication encourages cooperation between online platforms and competent authorities in order to verify whether content was notified legitimately. Moreover, emphasis is put on reporting by platforms to increase their accountability towards their users and to prevent over-removal. Where legal content has been blocked or removed, the content provider should have the possibility to submit a counter-notice and the removed content should be restored.

However, just like the Code of Conduct, the Communication is a non-binding instrument and therefore not enforceable against online platforms. Although the Communication puts forward measures to prevent over-removal, the adherence to fundamental rights and the compliance with data protection standards, the proposed measures depend on the will of the online platforms to take action, as non-compliance will not lead to sanctions. Moreover, out-of-court dispute resolution is given preference over traditional court proceedings, which may lead to similar issues as those mentioned above with reference to the Code of Conduct.

Online platforms maintain wide discretion for blocking or removing content according to their own community guidelines and, at the same time, are exempt from liability where (illegal) content is blocked or removed. This could evidently encourage online platforms to block or takedown content as a "preventive measure", particularly where platforms do not have the resources to review all notifications thoroughly. On the other hand, business incentives might confine over-blocking, as will be illustrated below.

Out-of-court settlements may, albeit having the advantage of being much faster than conventional court proceedings, be opaque and do not provide for an effective appeal mechanism. As a consequence, measures taken by the online platforms may not only have an impact on users' right to freedom of expression and to information, but they may affect other fundamental rights, such as the right to an effective remedy and to a fair trial enshrined under Article 47 of the EU Charter.

## III.2. The Recommendation on measures to effectively tackle illegal content online

The Code of Conduct and the Communication were further complemented by a Recommendation that the Commission published on 1 March 2018.[52] The document is divided into four chapters, focussing on purpose and terminology, general recommendations relating to all types of illegal content, a chapter on specific recommendations concerning terrorist content, and Chapter IV that refers to the provision of information and to the monitoring of the measures put in place by service providers.

The Recommendation seemingly provides for certain clarification regarding the sometimes ambiguous and opaque language of the Code of Conduct and further elaborates the Communication from 2017. For instance, the Recommendation seeks to improve the mechanisms on user feedback where notices and counter-notices are submitted to service providers. When encountering illegal content, users should be

[50] Ibid, p. 19.
[51] Ibid, p. 16.
[52] Commission Recommendation of 1.3.2018 on measures to effectively tackle illegal content online, C(2018) 1177 final, Brussels, 1.3.2018.

able to notify service providers via user-friendly mechanisms, allowing the notice provider to give an explanation on why content should be blocked or removed. Where notice providers indicated their contact details to the service provider, the latter should give feedback on the decision that was taken concerning the notified content.[53] Where notice providers submit their contact details, data shall be processed in accordance with the GDPR. This is a further step-up from the Code of Conduct and the Communication.

Moreover, service providers should inform content providers of any removed or disabled content and provide them with the possibility to contest a decision in cases where it is not clear that the content concerned is illegal content.[54] In such cases, content providers should be able to submit counter-notices to be reviewed by the service providers and, in cases were the content that was disabled or removed is to be considered *not* illegal, service providers should reverse the decision on removal.[55]

With regard to enhanced transparency, the Recommendation encourages service providers to publish information about removed or disabled content, the number of notices and counter-notices submitted, including the time needed for taking action.[56] However, this means that the abovementioned issues concerning transparency would also remain under the Recommendation, as online platforms are solely "encouraged" to provide more insights to their internal processes, but they are not further instructed on what exact data they should provide for scrutiny.

In addition, the Recommendation also suggests the use of proactive measures, taking as an example the use of automated means, to detect illegal content.[57] In accordance with EU data protection law, such automated detection should be accompanied by appropriate safeguards, human oversight and review.[58] Here, online platforms will have to show compliance with the GDPR, which sets high standards for the protection of personal data.

Altogether, the Recommendation provides for clarity concerning some of the ambiguities that the Code of Conduct lacked and requires more responsibility from online platforms with regard to transparency and accountability towards users. Yet, provisions on evaluating the implementation of the measures are missing, *for instance*, where the Recommendation solely suggests counter-notice procedures for content that is deleted on the basis of illegality, not for content removed under terms of service or community guidelines.

Moreover, despite being considerably more extensive than the Code of Conduct, the Recommendation only vaguely refers to the fundamental rights that might be impacted by notice-and-takedown measures or the conditions that must be satisfied to limit such rights (i.e.: proportionality, necessity and adequacy). Evidently, the EU Charter only applies regarding measures that were implemented by Member States when implementing EU law, but not to measures that are to be followed "voluntarily" by private companies. Yet, the EU Charter, as well as the ECHR,[59] may have an interpretive effect and provide guidelines, standards and conditions for voluntary actions by these companies. Beyond that, in accordance with Article 51(1) of the EU Charter, the EU institutions and, thus, the European Commission is bound by the Charter when proposing, together with companies, Codes of Conduct. The Commission must respect and observe the principles under the Charter, shall promote its rights and shall not interfere with any of such fundamental rights. Therefore, even if the Charter is not applicable

[53] Ibid, points 5-8.
[54] Ibid, points 9-10.
[55] Ibid, points 10-13.
[56] Ibid, points 16 and 17.
[57] Ibid, point 18.
[58] Ibid, point 20.
[59] Article 52(3) clarifies that the meaning and the scope of the rights protected by both the ECHR and CFEU should be the same.

to private companies, it is even more important that the Commission ensures that any measure that may interfere with the right to freedom of expression is accompanied by adequate safeguards.

## III.3. The Commission's Communication on disinformation from April 2018

Only one month after the publication of the Recommendation on measures to tackle illegal content online, another self-regulatory measure was issued by the Commission: A Communication on tackling disinformation online. Although the latter covers a different matter than the one that the Code of Conduct countering illegal hate speech refers to, both instruments are comparable in terms of their self-regulatory nature and the difficulties regarding the enforceability of the proposed measures.

With the Communication on Disinformation, the Commission reacts to a Resolution of the European Parliament from June 2017,[60] in which the Parliament had called on the Commission to analyse both the current situation and the legal framework with regard to fake news, and to verify the possibility of legislation to limit the dissemination of such content.[61] Following the resolution, the Commission included an "initiative addressing online platform challenges as regards the spreading of fake information" into its 2018 Work Programme,[62] launched a public consultation on fake news and established a high-level expert group (HLEG) representing academia, online platforms, news media, and civil society organisations in order to define a clear, comprehensive and broad-based action plan to tackle the spread and impact of online disinformation in Europe.[63] The Commission's Communication on Disinformation from April 2018 takes into account the HLEG's Recommendations, which were published in March 2018,[64] as well as the results of the public consultation[65] and the Eurobarometer opinion poll,[66] both published at a similar time as the HLEG Recommendations.

Despite the European Parliament's suggestion to consider legislative options, with the Communication on Disinformation the Commission opted, once again, for a self-regulatory measure, as "self-regulation can contribute to the efforts to tackle online disinformation, provided it is effectively implemented and monitored".[67] According to the Communication, online platforms should step up their efforts to counter disinformation through key principles developed in a Code of Practice, which would, *inter alia,* aim to (1) improve the scrutiny of online advertisement, (2) ensure transparency about sponsored content online, (3) intensify and demonstrate efforts to close fake accounts, (4) facilitate users' assessment of online content, (5) dilute the visibility of disinformation, (6) establish clear marking systems for bots, (7) facilitate content discovery and access to different news sources, (8) ensure that online services include, by design, safeguards against disinformation, and to (9) provide trusted fact-checking organisations and academia with access to platform data.[68] Moreover, the Communication seeks to foster media literacy of citizens,[69] highlights the importance of quality journalism[70] and encourages strategic communication and awareness-raising by public authorities to counter disinformation.[71] All

[60] European Parliament, P8_TA(2017)0272, 'Online platforms and the Digital Single Market European Parliament resolution of 15 June 2017 on online platforms and the digital single market', (2016/2276(INI)).
[61] Ibid, point 36.
[62] European Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Commission Work Programme 2018. An agenda for a more united, stronger and more democratic Europe, COM(2017) 650 final, Strasbourg, 24.10.2017.
[63] European Commission, 'Fake news', https://ec.europa.eu/digital-single-market/en/fake-news
[64] Report of the independent high level Group on fake news and online disinformation, 'a multi-dimensional approach to disinformation', March 2018.
[65] https://ec.europa.eu/digital-single-market/en/news/synopsis-report-public-consultation-fake-news-and-online-disinformation.
[66] http://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/survey/getsurveydetail/instruments/flash/ surveyky/2183.
[67] Communication on Disinformation, p. 8.
[68] Ibid, pp. 7-8.
[69] Ibid, p. 13.
[70] Ibid, p. 15.
[71] Ibid, p. 16.

this should be achieved, *inter alia,* by deploying new technologies, such as AI or cognitive algorithms to identify and combat disinformation.[72]

Consequently, the Commission views the responsibility of tackling online disinformation and fake news mainly on the side of the online platforms but leaves the option open to propose regulation in case of non-compliance with the Code of Practice.[73]

Yet, other than the content under the Code of Conduct countering hate speech, the Communication includes content that is not in itself illegal or defined in relevant national legislation such as the Framework Decision on Xenophobia to which the Code of Conduct makes reference.[74] Although the Communication excludes from its scope reporting errors, satire and parody, or clearly identified partisan news and commentary,[75] it will be challenging to make a distinction between disinformation and parody or partisan news.

Moreover, the threat to impose regulatory measures might increase the pressure on online platforms to remove content without any valid legal basis. Although the Communication emphasizes that the proposed measures should strictly respect the freedom of expression and acknowledge the commitment to an open, safe and reliable internet, it does not clarify how this should be achieved where content is selected, blocked or removed based on the platform's assessment and not based on the assessment of independent fact-checkers. Due to a lack of clear definitions and the fear of regulation, online platforms might discretionally classify content as disinformation and disproportionally censor valid information.

### II.3.1. The Code of Practice from September 2018

On 26 September 2018, five months after the publication of the Communication, the Commission issued a Code of Practice on Online Disinformation,[76] under which the signatories[77] recognise the need to improve the scrutiny of advertisement placement[78] and the transparency about political advertisement,[79] the importance of closing fake accounts and clarifying rules on bots,[80] enabling privacy-compliant access to data for fact-checking activities,[81] and the need to dilute the visibility of disinformation.[82]

The Code specifically covers five of the nine domains mentioned in the Communication – (1) scrutiny of ad placements, (2) transparency about political advertising and issue-based advertising, (3) integrity of services and closing of fake accounts, (4) empowering consumers by diluting the visibility of disinformation and (5) empowering the research community – related to which the signatories commit "to a wide range of actions, from transparency […] to the closure of fake accounts and demonetarization of purveyors of disinformation" to tackle disinformation.[83] Within these five areas, the signatories commit, *inter alia,* to put in place processes to disrupt the misrepresentation of essential information through advertisement,[84] to enable public disclosure of political advertising, to establish policies regarding the misuse of automated bots,[85] to invest in tools and technologies to help people to make

[72] Ibid, p. 11.
[73] Ibid, p. 9.
[74] The Communication defines disinformation as 'verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public and may cause public harm.
[75] Ibid, p. 5.
[76] European Commission, 'Code of Practice on Disinformation', 26 September 2018, https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation.
[77] For now, Google, Facebook and Mozilla are some of the signatories of the Code.
[78] Code of Practice on Disinformation, p. 4.
[79] Ibid, p. 3.
[80] Ibid, p. 6.
[81] Ibid, p. 8.
[82] Ibid, p. 7.
[83] European Commission Statement, 'Statement by Commissioner Gabriel on the Code of Practice on Online Disinformation', Brussels, 26 September 2018, http://europa.eu/rapid/press-release_STATEMENT-18-5914_en.htm.
[84] Code of Practice on Disinformation, p. 5.
[85] Ibid, p. 6.

informed decisions and to improve media literacy,[86] and to encourage research into disinformation and political advertising on their platforms.[87]

In order to measure the Code's effectiveness, the signatories pledge, with respect to their respective commitments, to writing annual reports, in which they shall demonstrate their work including the measures taken by them and the progress made in improving transparency regarding their work on tackling disinformation. The commitments shall be measured along a non-exhaustive list of "key performance indicators"[88] and the reports shall be reviewed by a third party, selected by the signatories.[89]

The Code of Practice, like the Communication on disinformation, was met with criticism, mainly due to its lack of a comprehensive approach, its meaningless and unenforceable commitments, its immaterial objectives and the unaccountable monitoring and implementation process.[90]

Additional points of criticism would be that accountability decreases where the annual reports do not have to be made available to the public and where the signatories to the Code may themselves select the third party to review their reports. In any event, reports will only include and an assessment will solely take place in line with those commitments that the signatories agreed upon. Furthermore, signatories may decide to withdraw from either specific commitments or the entire Code at any point in time,[91] which encourages "cherry picking" the points that the signatories fulfil already.

Transparency through making available relevant information is indispensable to demonstrate that there is a true desire to tackle disinformation, as the fight against disinformation and fake news cannot be measured according to blocking or removing content or accounts, nor by disabling automatic bots. Under the Code of Conduct countering hate speech, this is not (yet) achieved, as the evaluation reports assessing the Code of Conduct measure the effectiveness of the measure according to the number and speed of takedowns and not according to the illegality of removed content.

Although the Code of Practice on Disinformation does not impose any enforceable sanctions against signatories, it is more straightforward than the Code of Conduct in requiring certain steps by the signatories. Yet, the area covered by the Code of Practice is more delicate and may pose an even higher risk to the freedom of expression, as content that might be censored is neither illegal nor defined in binding legislation. Moreover, the Code of Practice neglects important aspects, for instance regarding the spread of disinformation: albeit mentioning the danger of automated bots and fake accounts, the Code does not refer to the automated processes and self-learning algorithms that disseminate disinformation systematically based on the users' preferences. Rather than providing "general information on algorithms", the Code could have stipulated a much more straightforward approach to require signatories to demonstrate accountability concerning their internal processes. After all, it might be undesirable to block or remove disinformation as this violates freedom of expression. Certainly, media literacy and raising awareness of users might be a first step to achieve credibility. However,

[86] Ibid, p. 7.
[87] Ibid, p. 8.
[88] For instance, policies and enforcement activities in relation to reducing monetisation opportunities for providers of disinformation, measures to improve the visibility to consumers of instances of political advertising, measures to integrate and roll-out policies in relation to the integrity of their services in the context of Disinformation, measures to empower consumers with products, tools, technologies, and programmes, measures to improve the ability of researchers and civil society groups to monitor the scope and scale of political advertising or to encourage training of people in critical thinking and digital media and skills. See: Code of Practice, pp. 8-9.
[89] Ibid, p. 9.
[90] See, for instance: Joint Press Statement, 'The Sounding Board of the Forum on Disinformation issues Their unanimous final Opinion on the so-called Code of Practice, https://www.euractiv.com/wp-content/uploads/sites/2/2018/09/Joint-Press-Statement-Sounding-Board-Issues-Opinion-on-Code-of-Practice-EMBARGO.pdf.
[91] Code of Practice on Disinformation, p. 10.

limiting the spread of fake news and disinformation through algorithms will be relevant where online users neither understand nor initiate the mechanisms behind the spread of certain information.

## Interim Conclusion

With the Code of Conduct, the Commission opted for a model that is based on non-binding self-regulation. Whether it is a wise approach to entrust private companies with the task of checking posts against national laws on illegal content or to bestow them with the responsibility to decide on the removal of controversial content is questionable.

Ultimately, the Code aims to tackle online hate speech that is *already* punishable. Is the Code therefore necessary or merely a political statement to demonstrate action is taken in an area that might require much more than a self-regulatory approach to resolve complex issues? For the companies it is an excellent tool to demonstrate their commitment to contribute voluntarily to counter illegal hate speech without actually having to abide by binding rules that might restrict them in their current data management strategies.

Questions concerning the prerogative of interpreting and limiting the freedom of expression, regarding the democratic legitimacy of removal procedures, the accountability of automation influenced by commercial interests and the lack of effective appeals mechanisms remain.

Both the 2017 Communication and the Recommendation from March 2018 should be evaluated as a step-up from the Code in terms of scope, clarifications and explanatory details. However, particularly the use of proactive measures that are mentioned in both documents remains unclear.

The unchecked use of algorithms in recognition and filtering technologies, which analyse and remove content generated on platforms, causes clear risks. Algorithms fed by an uncertain combination of commercial interest and legal parameters may remove content where it fits company interests rather than fundamental rights principles. Although at this stage the use of algorithms to monitor content would still make use of the "human-in-the-loop principle",[92] the diligence and efficiency with which illegal content can be reviewed is also dependent on the financial capacity and resources of each company. After all, the implementation of review and mediation mechanisms does not provide business incentives for the companies to create sophisticated communication channels with users or national regulators.[93]

Due to time pressure on reviewers, the three instruments might increase the danger of over-blocking, leading to legal content being taken down erroneously, or the removal of content that does not fit the agenda of the companies. Here, one could ask whether it is desirable to grant private companies the possibility to judge what is illegal content and what not and whether a profit-driven company should be given the task to decide on the scope of the right to freedom of expression.

Similar concerns arise with regard to the Communication on disinformation and its Code of Practice, although it might even be argued that, in the area of disinformation, the risk of a disproportionate interference with the right to freedom of expression is even higher than in the context of the Code of Conduct, the Recommendation and the March 2018 Communication, as – other than hate speech – the spread of disinformation is not a criminal offence and the blocking or removing of such content might

[92] The human in the loop principle requires human intervention for automated computing systems. Such human intervention is also required in the area of data protection for automated individual decision-making. See also the Google NthzDG Transparency report: Google, 'Removals under the Network Enforcement Law – Google Transparency Report' <https://transparencyreport.google.com/netzdg/youtube?hl=en> accessed 28 August 2018.

[93] For the German Netzwerkdurchsetzungsgesetz, Bitcom calculated additional costs of more than 500 million Euro for the implementation of complaint management systems. In: Bitkom, 'Stellungnahme zum Regierungsentwurf eines Netzdurchsetzungssetz',
https://www.bitkom.org/noindex/Publikationen/2017/Positionspapiere/FirstSpirit-1492755732142201704 20-Bitkom-Stellungnahme-zum-Regierungsentwurf-NetzwerkDG.pdf. 20.04.2017.

be even more discretional than under the Code of Conduct on hate speech. Although disinformation and fake news may be disproved by tangible facts, a certain margin of bias might nevertheless remain. Moreover, it is doubtful that the receivers of such information will be easily persuaded even if confronted with facts.

It could, however, be argued that the Code of Practice puts more pressure on the signatories to demonstrate compliance with their commitments that the Code of Conduct on hate speech, as it is more detailed and sets more distinct guidelines to measure the Code's effectiveness. Moreover, the Code of Practice involves the Commission in the monitoring activities and requires signatories to demonstrate transparency by providing access to relevant data. Nevertheless, both the Communication on Disinformation and the accompanying Code of Practice remain non-binding self-regulatory instruments without any genuine enforcement mechanisms from which the signatories can withdraw at their discretion. Yet, the hope remains that the signatories will abstain from withdrawing from the Code, as any withdrawal would probably have undesirable consequences on a company's reputation.

## IV. National legislation

As elaborated above, the Code of Conduct is a largely self-regulatory model, relying on private sector actors to thrash out the details of determining the illegality of content and removing it. It is controversial, however, whether these purely self-regulatory efforts have brought the desired effects so far. Some EU Member States did not think so and have started separate initiatives and legislative projects. Apart from the German network enforcement law (*NetzDG*)[94], , the UK initiated a Parliamentary inquiry into whether online platforms should be made more accountable for the content hosted on their sites.[95] Likewise, France[96] announced plans to beef up legislation aimed at making social media platforms step up enforcement against hate speech online. These initiatives occur against the backdrop of a rise in xenophobic, racist or extremist content in the wake of the *migrant crisis* in the EU and against the backdrop of terrorist threats and propaganda online. It is also not rare that legislative action is stirred prior to national elections, such as in Germany, Italy[97] or in the UK.

Germany had adopted its own industry code of conduct[98] by end of 2015. The code, also referred to as *Task Force against Illegal Hate Speech* (Task Force hereafter), was similar to the one concluded by the EU Commission one year later. The German Government agreed with major social media companies, Facebook, Twitter and Google's VSP YouTube, that they facilitate user notifications of hate speech, ensure expeditious verification and deletion of illegal content within 24 hours and invest in specialised staff that would deal with notifications according to German law.

Federal elections were held in 2017 in Germany. The run-up to the elections was accompanied by a surge in populist sentiment fuelled by the *migrant crisis* and the threat of terrorism, which was increasingly vented through social media. The Government's monitoring exercises, published in March

[94] Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken 2017 (BGBl I S 3352 (Nr 61)).

[95] 'The Internet: To Regulate or Not to Regulate? Inquiry' (*UK Parliament*) <https://www.parliament.uk/business/committees/committees-a-z/lords-select/communications-committee/inquiries/parliament-2017/the-internet-to-regulate-or-not-to-regulate/> accessed 24 August 2018.

[96] 'Plan antiraciste : Edouard Philippe cible la «cyberhaine»' (*leparisien.fr*, 19 March 2018) <http://www.leparisien.fr/societe/plan-antiraciste-edouard-philippe-cible-la-cyberhaine-19-03-2018-7616820.php> accessed 24 August 2018.

[97] 'Italy: First Attempt to Regulate the Online Political Propaganda' (*Portolano Cavallo*) <http://www.portolano.it/en/2018/02/italy-first-attempt-to-self-regulate-the-online-political-propaganda/> accessed 1 October 2018.

[98] Bundesministeriums der Justiz und für Verbraucherschutz, 'Together against Hate Speech - Ways to Tackle Onl Ine Hateful Content Proposed by the Task Force against Illegal Online Hate Speech' <http://www.bmjv.de/SharedDocs/Downloads/DE/Artikel/12152015_TaskForceErgebnispapier_eng.pdf?__blob=publicationFile&v=2> accessed 30 March 2017.

2017,[99] found that the processes put in place by the participating companies under its 2015 Task Force were still insufficient and that two of the three companies failed to delete the majority of illegal content reported to them. Consequently, the German Government proposed a law[100] which would impose obligations on social media networks and user-generated content providers on how to deal with illegal hate speech notified to them. This proposal became law *as* the *NetzDG* in June 2017, just before the September general elections. The NetzDG has been analysed in detail in a dedicated chapter in this book by Thomas Wischmeyer. It should therefore be sufficient to state for the purposes of this chapter that the *NetzDG* is an example of a Member States undertaking legislative action, precipitously, due to a perceived lack of traction from purely self-regulatory codes of conduct. But, as demonstrated by Thomas Wischmeyer, the national legislation merely fixes the non-binding measures previously agreed upon with industry into law. This concerns mainly the *ex-post* removal procedures of notified, allegedly illegal hate speech. The law could therefore be seen as an attempt to spell out procedural requirements of already existing obligations of hosting service providers under the ECD.[101] Member states have the discretion to do this because the notice and takedown obligations remain rather general within the ECD. However, the national legislator shunned more proactive duties to detect and remove illegal content, arguably because of the limitations imposed by the ECD.

## V. Beyond self-regulation of speech

Within a span of 20 years, social media networks and internet platforms in general have revolutionised the way we exchange, create and access information and do business. The regulatory response by the EU and Member States to the emergence of these massive transformations has focussed on self-regulation. This is not surprising. Self-regulation lends itself to environments characterised by fast technological change, creative disruption and overburdened and constantly adjusting regulators.[102] Arguably, today's internet giants may have prospered thanks to light touch regulation across the globe. Today, social networks like Facebook or Twitter, or user generated content sites like YouTube or Instagram, play essential roles in facilitating communication, expression and creativity. The large user base and global reach and the ensuing network effects mean that these platforms are likely to be the principal means of online and public communication for a significant number of people today.[103] These platforms are therefore increasingly seen as *quasi* public fora, which enable access to information and facilitate freedom of expression.[104]

There have been more and more calls to regulate these platforms according to their characteristics as digital utilities[105] or public spaces.[106] According to these approaches, the responsibilities of those intermediaries would be defined along the public interest, which would focus on fundamental rights, such as freedom of expression, privacy, property, dignity, or the protection of minors. Regulating platforms as entities with public functions would entail, however, an overhaul of the current regulatory framework applicable to these internet hosts.

The current wide-reaching liability exemptions, which also apply to the large platforms targeted by recent EU initiatives, hinge on the active or passive role these platforms play in the act of hosting

[99] Bundesministerium für Justiz und Verbraucherschutz, 'Löschung rechtswidriger Hassbeiträge bei Facebook, YouTube und Twitter - Ergebnisse des Monitorings von Beschwerdemechanismen jugendaffiner Dienste' (2017) <https://www.fair-im-netz.de/SharedDocs/Downloads/DE/News/Artikel/03142017_Monitoring_jugendschutz.net.pdf?__blob=publicationFile&v=3> accessed 27 August 2018.
[100] NetzDG (n 97).
[101] ECD (n 21). Article 14 (2)
[102] Ullrich (n 13) 12.
[103] Jonathan Peters and Brett Johnson, 'Conceptualizing Private Governance in a Networked Society' (2016) 18 NCJL & Tech. 15, 49.
[104] Pasquale (n 13) 512.
[105] ibid 490.
[106] William Perrin and Lorna Woods, 'Reducing Harm in Social Media through a Duty of Care' (*Carnegie UK Trust*, 8 May 2018) <https://www.carnegieuktrust.org.uk/blog/reducing-harm-social-media-duty-care/> accessed 28 August 2018.

information. It has been stated repeatedly[107] that, in the context of *Web 2.0*, this concept is becoming increasingly blurred and does not adequately reflect the revolutionary changes that have happened in content management, internet business models and diversification in the platform economy. *Pasquale* demonstrated, for example, how a big platform like Google has opportunistically characterised itself as a speech conduit (passive) and content provider (active), depending on the offence concerned and the commercial interest.[108] In the EU, Google was recently charged with a hefty competition fine for manipulating the search rankings of its search engine to favour its own products.[109] This also helps to demonstrate the obsolescence of the active/passive dichotomy.

If the platforms were considered public spaces, they would be judged by what they do to prevent the violation of fundamental rights to and by the visitors of their virtual spaces. Therefore, slapping proactive infringement prevention duties on private actors without installing public oversight, as attempted through the Copyright and Audio-Visual Media Services[110] Directive proposals and the recent EU Recommendation,[111] appears to be misguided. It is indeed in violation of the general monitoring prohibition of Article 15 of the e-Commerce Directive which, however, in itself seems like an outlandish concept considering the amount of information these companies analyse, manage and turn into money on a daily basis. Entities operating along commercial interest criteria, but providing important public functions, it is submitted here, should be subject to stricter regulatory oversight with regard to the exercise of fundamental rights.

There are by now a number of proposals and attempts which try to take into account and adjust the tilted balance between the power of global internet platforms, on the one hand, and users and their fundamental rights, on the other. Many of these attempts advocate increased responsibilities for platforms along the concepts of duty of care,[112] moral responsibility,[113] or economic law arguments relating to the cheapest cost avoider principle.[114] The concept of duty of care may for the moment be the most established one among these approaches. In addition, duty of care is linked with both considerations of moral responsibility[115] and economic models[116] of liability. It is already widely applied throughout a variety of regulatory fields in both common and civil law and has also found its way into CJEU case law.[117]

The important difference of such an approach to the proposals made by the Commission are that social media platforms, once recognised as public utilities or public spaces, would be regulated along closely

[107] Bertin Martens, 'An Economic Policy Perspective on Online Platforms' (Institute for Prospective Technological Studies 2016) Digital Economy Working Paper 2016/05 JRC101501 34; Peggy Valcke, Aleksandra Kuczerawy and Pieter-Jan Ombelet, 'Did the Romans Get It Right? What Delfi, Google, EBay, and UPC TeleKabel Wien Have in Common', *The responsibilities of online service providers* (Springer Berlin Heidelberg 2017).
[108] Pasquale (n 13) 494–497.
[109] 'European Commission - Press Release - Antitrust: Commission Fines Google €2.42 Billion for Abusing Dominance as Search Engine by Giving Illegal Advantage to Own Comparison Shopping Service' <http://europa.eu/rapid/press-release_IP-17-1784_en.htm> accessed 28 August 2018.
[110] Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market, COM(2016) 593 final 2016. Article 13, Proposed AVMSD amendment (n 28). Article 28(a).
[111] EU Commission, 'COMMISSION RECOMMENDATION of 1.3.2018 on Measures to Effectively Tackle Illegal Content Online, C(2018) 1177 Final'.
[112] Augustin Waisman and Martin Hevia, 'Waismann Theoretical Foundations of Search Engine Liability' (2011) 42 International Review of Intellectual Property and Competition Law 785; Perrin and Woods (n 109).
[113] Mariarosaria Taddeo and Luciano Floridi, 'The Debate on the Moral Responsibilities of Online Service Providers', *The responsibilities of online service providers* (Springer Berlin Heidelberg 2016) 1583 <http://link.springer.com/10.1007/s11948-015-9734-1> accessed 17 February 2017.
[114] Martens (n 110) 34–35.
[115] Anton Vedder, 'Accountability of Internet Access and Service Providers – Strict Liability Entering Ethics?' (2001) 3 Ethics and Information Technology 67.
[116] Emanuela Carbonara, Alice Guerra and Francesco Parisi, 'Sharing Residual Liability: The Cheapest Cost Avoider Revisited' (2016) 45 The Journal of Legal Studies 173.
[117] Herwig CH Hofmann, 'Delegation, Discretion and the Duty of Care in the Case Law of the Court of Justice of the European Union' [2018] SSRN Electronic Journal <https://www.ssrn.com/abstract=3169744> accessed 28 August 2018.

defined public interest criteria. In the case of social media platforms, these public interests would be defined by the fundamental rights that these public fora are meant to protect. Duty of care would mean that companies would be obliged to work proactively to ensure that these rights are being respected on their platforms. This regulatory approach is not new to the EU. Public interest criteria are the basis for formulating essential requirements in product regulation, minimum standards in health and safety laws and environmental protection and many more areas. Importantly, a system based on the duty of care would not be restricted to reactive, *ex-post* content removal. Reducing harm for users would entail that platforms engage in holistic risk management, which includes preventive activities. Mandatory proactive infringement prevention is often decried for its potential effects on human rights, such as over-blocking and intrusion into privacy. However, the approach pursued here differs from the Copyright Directive and the AVMSD proposals currently put on the table by the Commission. Regulators would now have oversight and be able to see whether the preventive activities of platforms in combatting illegal content are proportional, i.e. adequate and necessary. Arriving at a proportional solution entails the balancing exercise between the various fundamental rights at play in these public fora. The mechanisms behind these balancing exercises would be under public scrutiny. The complexities of establishing the adequate level of duty of care with regard to preventing allegedly unlawful content have been witnessed in the *Delfi*[118] and *MTE*[119] judgments of the ECtHR, which were discussed elsewhere in this book. While courts are still ideally placed to perform this exercise, it would be more effective for all stakeholders that the bulk of these decisions would be operationalised in a transparent and accountable way.

By setting a standard for the duty of care, the outdated and simplistic distinction between general and specific monitoring for illegal information could be overcome. It means that the principles behind content moderation change from a rules-based enforcement of community standards or terms and conditions to a risk-based approach.[120] A regulator would require that a social media network has identified and classified the risks related to illegal activity on its platforms (e.g. hate postings, defamatory content, data theft, etc.) and has adjusted its preventive activity accordingly. For example, the platform would assess the risks of illegal hate postings by correlating the likelihood of it occurring in certain user and network contexts with the degree of harm caused (i.e. the impact). Certain keywords, their combinations and frequency could thus be subject to a different risk score depending on the conversational context and user characteristics, user provenance, previous sanctions, group characteristics, involvement of paid/sponsored content etc.). The platform could be required to focus its prevention efforts on high risk activities. This could mean, for example, deploying enhanced content filtering and manual review in these areas. The risk assessment would need to be documented for and may be reviewed by the regulator. Regulatory risk assessment is standardised and has been adopted by the EU in various other regulatory areas.[121]

Most platforms are likely to be familiar with this methodology. Risk assessment activities, such as fraud detection and IT security threat analysis, are part and parcel of any of the large internet platforms operating in the marketplace today. This can already be seen from looking at the Google Transparency report[122] mentioned above.

[118] *Delfi AS v Estonia, no 64569/09* (ECtHR (Grand Chamber)).
[119] *Magyar Tartalomszolgáltatók Egyesülete and Index.hu zrt v Hungary, no 22947/13* (ECtHR (Fourth Section)).
[120] For an in-depth discussion on the development of content moderation systems on social media platforms see: Kate Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech' 131 HARVARD LAW REVIEW 73, 1631.
[121] See for example: EU Commission, 'EU General Risk Assessment Methodology (Action 5 of Multi-Annual Action Plan for the Surveillance of Products in the EU (COM(2013)76))' 12–15. EU Commission, 'Risk Assessment and Mapping Guidelines for Disaster Management - Staff Working Document, SEC(2010) 1626 F' (21 December 2010) 15–19 <https://ec.europa.eu/echo/files/about/COMM_PDF_SEC_2010_1626_F_staff_working_document_en.pdf> accessed 29 August 2018.
[122] Google (n 95).

Following the risk assessment, the platform would need to show that its risk responses are adequate with regard to the risk identified. This methodology is in line with modern risk regulation in which purely outcome-based provisions are combined with more individualised, internal risk management processes.[123]

The EU Commission's September 2018 proposal for a regulation on preventing terrorist content online[124] explores this kind of approach for the first time. It defines and imposes minimum duties of care on hosting providers, not only for removing terrorist content notified by Member States and users, but also for preventing its dissemination by the use of proactive measures.[125] The proposal builds on the Recommendation on measures to effectively tackle illegal content online of March 2018.[126] It aims at prescribing enhanced measures with regard to tackling terrorist speech.[127] Legislative action was endorsed by the European Council and the European Parliament in 2017.[128] The reactive duties provide a procedural framework aimed at fair and efficient decision-making and the removal of notified content. They impose time limits for removals, procedural safeguards, such as complaints mechanisms and information requirements, as well as transparency reporting obligations on the information hosts.[129]

More remarkably, however, the proposal mandates, for the first time, the use of proactive measures aimed at preventing certain content, using a risk-based approach.[130] Other than the proposed Copyright Directive and AVMSD recast, this proposal requires platforms to perform a risk assessment, taking into account exposure to terrorist content, fundamental rights of users and freedom of expression and information. Online hosts that were subject to final content removal orders by national authorities will have to be able to demonstrate that they have taken specific proactive measures, including automated tools, to prevent the re-upload of content, and to detect, identify and remove this kind of content.[131] Failure to put such measures in place may result in Member States ordering the platform to do so. The Commission even relativizes its hitherto sacrosanct application of Article 15 of the e-Commerce Directive, which prohibits Member States from imposing general monitoring obligations on information hosts.[132] While industry may complain about a further privatisation of law enforcement,[133] the proposal actually tries to remedy this tendency by obliging platforms to report on their proactive measures to competent authorities. The idea behind this is to ensure public oversight and a fair balance of all public interest principles at stake. Such a public review assessment would, for example, look at the human oversight and verification procedures employed for automated tools, the economic capacities of hosting services, fundamental rights at stake, etc.[134] The proposal does, therefore, exactly the opposite: it will drag opaque automated proactive content identification and removal procedures into the open by making them subject to public review and transparency. The proposal carries the hallmarks of risk regulation mentioned above in this section.

[123] Kenneth A Bamberger, 'Technologies of Compliance: Risk and Regulation in a Digital Age' (2009) 88 Tex. L. Rev. 669, 673.
[124] Proposal for a regulation on preventing terrorist content online, COM(2018) 640 final 2018.
[125] Ibid Recital 12 and Article 3
[126] EU Commission, 'Recommendation on Measures to Effectively Tackle Illegal Content Online' (n 114).
[127] COM(2018) 640 final (n 127). 1-4, Recital 4.
[128] European Parliament, 'Online Platforms and the Digital Single Market - P8_TA(2017)0272' para 33 <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT%2BTA%2BP8-TA-2017-0272%2B0%2BDOC%2BXML%2BV0//EN> accessed 2 October 2018; European Council, 'European Council Meeting (22 and 23 June 2017) – Conclusions - EUCO 8/17' para 2 <https://www.consilium.europa.eu/media/23985/22-23-euco-final-conclusions.pdf> accessed 2 October 2018.
[129] COM(2018) 640 final (n 150 Articles 4, 5, 10, 11.
[130] Ibid Recitals 16. 19 and Article 6 (1).
[131] ibid. Article 6 (2).
[132] ibid. Recital 19.
[133] 'Terrorist Content Online: EuroISPA Concerned as Commission Privatises Law Enforcement' (*Europe internet services providers association*, 12 September 2018) <http://www.euroispa.org/terrorist-content-inline-euroispa-concerned-commission-privatises-law-enforcement/> accessed 1 October 2018.
[134] COM(2018) 640 final (n 127). Recital 13.

It remains open, however, how competent authorities would be able to judge the proportionality and effectiveness of the proactive and automated measures of online platforms. They will need to set common technical criteria and indicators across the EU. In addition, they will need to have extensive technical expertise in order to assess and audit complex algorithmic decision-making.

A solution could be that the legislator mandates one of the European standards bodies to create a technical standard for duty of care or harm reduction[135] on social media platforms regarding the various types of illegal and infringing content. The regulator could borrow from existing models used in IT Security (ISO 27000), Occupational Health and Safety (ISO 45001), product standards, or even transaction risk monitoring in anti-money laundering.[136] The achievement of such technical standards would provide proof of conformity with an acceptable level of duty of care. The standards lay down the technical requirements for ensuring that online hosts prevent and remove illegal content in line with the public interest. These public interest principles would be set out in sector-specific legislation.

The EU could make use of its experience gained in the *New Approach*[137] legislation in which industry-led standardisation is a key component. The New Approach has been considered one of the success stories of European integration[138] and the EU has continuously reformed it.[139] Through the Joint Initiative on Standardisation,[140] the EU is expanding standardisation across a wide variety of industries, with one of the focus areas being the digital single market.[141]

The advantage of developing standards is that they can be adopted to different types of platforms and content and could eventually cover the entire ISP sector. It might require an overhaul of a platforms' risk management activities. Legal compliance would need to take a clearly defined place within the commercial risk management. The regulator would have the mandate to review the content (risk) management choices and processes of these platforms and test whether public interest criteria are being respected.

However, the process of standard creation is essentially managed by the industry. Platforms would have the freedom to design solutions that correspond to the technical capabilities of their systems. The EU regulators would be involved and consulted in this process and would control whether the public interest criteria are being met during the design and implementation phases of these standards. This would entail the review of and involvement in major decisions, from algorithm design of infringement detection and removal systems to procedural arrangements for notice and takedown or statutory reporting.

Duty of care is, therefore, not only focussed on preventive actions. A holistic system would also ensure that procedural rights are being observed. It would prescribe formal notice-and-takedown and

[135] Perrin and Woods (n 109).
[136] Ullrich (n 13).
[137] The New Approach goes back to 1985 and was developed in response to the CJEU's Cassis de Dijon judgment (Case 120/78 [1979] ECLI:EU:C:1979:42). As a result, entire products sectors were regulated by spelling out EU wide, mandatory essential requirements, which corresponded to public interests of consumer safety. In this co-regulatory system the EU mandated the industry to develop technical standards that would provide proof of conformity with the essential requirements for companies. More detail can be found in: EU Commission, 'COMMISSION NOTICE The "Blue Guide" on the Implementation of EU Products Rules 2016, Offical Journal of the EU (2016/C 272/01)' (2016) 59 Offical Journal of the European Union.
[138] Rob Van Gestel and Hans-W Micklitz, 'European Integration through Standardization: How Judicial Review Is Breaking down the Club House of Private Standardization Bodies' (2013) 50 Common Market L. Rev. 145, 156–157.
[139] Regulation (EU) No 1025/2012 of the European Parliament and of the Council of 25 October 2012 on European standardisation, amending Council Directives 89/686/EEC and 93/15/EEC and Directives 94/9/EC, 94/25/EC, 95/16/EC, 97/23/EC, 98/34/EC, 2004/22/EC, 2007/23/EC, 2009/23/EC and 2009/105/EC of the European Parliament and of the Council and repealing Council Decision 87/95/EEC and Decision No 1673/2006/EC of the European Parliament and of the Council 2010.
[140] 'Joint Initiative on Standardisation: Responding to a Changing Marketplace - Growth - European Commission' (*Growth*) </growth/content/joint-initiative-standardisation-responding-changing-marketplace-0_en> accessed 29 August 2018.
[141] EU Commission, 'Communication: ICT Standardisation Priorities for the Digital Single Market COM(2016) 176 Final' <https://ec.europa.eu/digital-single-market/en/news/communication-ict-standardisation-priorities-digital-single-market> accessed 29 August 2018.

automated takedown procedural requirements, such as the content of notifications filed to platforms, turnaround times, information requirements to users and uploaders following a notice action or counter claim procedures. In addition, the standard would prescribe regular and standardised statutory reporting by platforms to the public and to the regulatory authorities. Some information may only be accessible to the regulatory authorities. In complex technical environments, reporting is a transparent way to demonstrate compliance to the public, regulators or political representatives.[142] The *NetzDG* actually provides a useful starting point in this regard. Meanwhile, the EU Communication on tackling illegal content online also provides a useful starting point in the regard as it recognises the need for platforms to create and publish standardised transparency reports on notice and takedown.[143]

It should be underlined that the solution proposed rests on concerted efforts of all sides involved and that content responsibilities are not being allocated solely to platforms. Liability will remain shared between the activities of the content uploader, the platforms and, where adequate, the content user.

## VI. Conclusion

The Commission Code of Conduct on Hate Speech is an attempt by the EU Commission to stem the flood of illegal speech on social platforms by means of self-regulation. From the evidence presented so far on this soft law, it is unlikely that the private actors involved will indeed come together and present effective solutions in order to achieve this. Their content management decisions are deeply embedded in their internal systems and, so far, no regulator has been able to get enough insight into the criteria employed when removing content. It is questionable whether outsourcing this decision-making process on the legality of speech to private entities, driven by commercial interest, will lead to the desired outcomes of removing illegal speech with due respect for fundamental rights. The same is true for the other self-regulatory non-binding instruments analysed in the first part of this chapter. Social media platforms have become important enablers of speech, arguably with quasi-public and gatekeeping powers. Given the fundamental rights at stake, and the often delicate and complex balancing exercises required to make a decision on content removal, self-regulation is an inappropriate tool. Stronger regulatory oversight is needed when it comes to protecting fundamental rights for the entire gambit of participants on social media platforms.

Some EU Member States have moved forward in a bid to regulate the content management activities of these new internet behemoths. The German *NetzDG* is an attempt at formalising certain *ex-post* requirements of the notice and takedown process and introducing statutory reporting requirements. It is a useful attempt at making social networks more responsible actors. However, the *NetzDG* does not seek to respond to the threat of private algorithmic decision-making on content nor does it seek to regulate activities aimed at proactive identification and removal of illegal content. It seems that, as of now, the e-Commerce Directive provides a barrier for any attempt to instil more responsibility through preventive obligations on platforms.

However, additional degrees of mandated responsibility may be needed if the fight against illegal content is not to slip out of the hands of the regulator. Preventive duties along the Copyright Directive or the AVMSD recast proposals may still not be an answer. Safeguarding and balancing the fundamental rights of free speech, dignity, security, and privacy behind the closed doors of global private corporations is hardly what the EU regulator may want. Instead, the public functions of the social media companies, as enablers of speech and gatekeepers of information, should be adequately recognised. In such a regulatory system, platforms would need to apply duties of care with regards to the protection of their users. True to the platforms' *quasi* public function, that duty of care should be structured along the public interest principles of the fundamental rights involved. It would mean that these companies have transparent and accountable risk management mechanisms in place with regard to the activities on their

[142] Cohen (n 13) 406.
[143] COM(2017) 555 Final 16.

platforms. Such a risk management system would entail proportional and transparent measures of preventing and detecting illegal activity, as well as open and accessible notice-and-takedown mechanisms. The EU has by now considerable experience in developing risk regulation, starting from the *New Approach* to the more recent GDPR. It could use this expertise to mandate the development of technical standards of the duty of care which would respond to essential requirements set by fundamental rights.