TRANSATLANTIC WORKING GROUP

FOR IMMEDIATE RELEASE
June 17, 2019
Contact: Michael Rozansky | michael.rozansky@appc.upenn.edu | 215.746.0202

How governments and platforms have fallen short in trying to moderate content online

Efforts to curb hate speech, terrorism, and deception while protecting free speech have been flawed

PHILADELPHIA and AMSTERDAM – The Transatlantic Working Group, created to identify best practices in content moderation on both sides of the Atlantic while protecting online freedom of expression, has released its first working papers that explain why some of the most prominent efforts to date have failed to achieve these goals.

In a series of papers, members of the group find that current efforts by governments and platforms are flawed and have fallen short of the goals of adequately addressing the problems of hate speech, viral deception, and terrorist extremism online while protecting free speech rights.

The Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, or TWG, also released a report of the co-chairs with interim recommendations for government and platforms coming out of its initial meeting earlier this year in Ditchley Park, U.K. The co-chairs are former Federal Communications Commission member **Susan Ness**, a distinguished fellow of the Annenberg Public Policy Center of the University of Pennsylvania, and **Nico van Eijk**, professor of law and director of the Institute for Information Law (IViR) at the University of Amsterdam.

In their first report (download here), the TWG co-chairs recommended that:

- Specific harms to be addressed by content moderation must be clearly defined and based on evidence and not conjecture;
- Transparency must be built in – both by governments and platforms – so that the public and other stakeholders can more accurately judge the impact of content moderation;
- Due process safeguards must be provided so that authors of material taken down have clear and timely recourse for appeal;
- Policy makers and platforms alike must understand the risks of overreliance on artificial intelligence, especially for context-specific issues like hate speech or disinformation, and should include an adequate number of human reviewers to correct for machine error.

IViR

THE ANNENBERG PUBLIC POLICY CENTER OF THE UNIVERSITY OF PENNSYLVANIA

The set of working papers includes:

- **Freedom of Expression**: A pillar of liberal society and an essential component of a healthy democracy, freedom of expression is established in U.S., European, and international law. This paper looks at the sources of law, similarities and differences between the U.S. and Europe, and how laws on both sides of the Atlantic treat hate speech, violent extremism, and disinformation. Download

  - **Brittan Heller**, The Carr Center for Human Rights Policy, Harvard University
  - **Joris van Hoboken**, Vrije Universiteit Brussels and University of Amsterdam

- **An Analysis of Germany's NetzDG Law**: Arguably the most ambitious attempt by a Western state to hold social media platforms responsible for online speech that is deemed illegal under a domestic law, German's Network Enforcement Act took effect January 1, 2018. While instituting some accountability and transparency by large social media platforms, its provisions are likely to have a chilling effect on freedom of expression. Download

  - **Heidi Tworek**, University of British Columbia
  - **Paddy Leerssen**, Institute for Information Law (IViR), University of Amsterdam

- **The Proposed EU Terrorism Content Regulation:** The European Commission proposed the Terrorism Content Regulation (TERREG) in September 2018, and it is currently in the final stages of negotiation between the European Parliament and the Council. As drafted, the TERREG proposal raises issues of censorship by proxy and presents a clear threat to freedom of expression. Download

  - **Joris van Hoboken**, Vrije Universiteit Brussels and University of Amsterdam

- **Combating Terrorist-Related Content Through AI and Information Sharing:** Through the Global Internet Forum to Counter Terrorism, the tech industry uses machine learning and a private hash-sharing database to flag and take down extremist information. The analysis of this private sector initiative raises transparency and due process issues, and offers insight into why AI failed to promptly take down the Christchurch, New Zealand, shooting videos. Download

  - **Brittan Heller**, The Carr Center for Human Rights Policy, Harvard University

- **The European Commission's Code of Conduct for Countering Illegal Hate Speech Online**: The Code, developed by the European Commission in collaboration with major tech companies, was introduced in 2016. As a voluntary code, it was seen as less intrusive than statutory regulation. But the code is problematic: It delegates enforcement actions from the state to platforms, lacks due process guarantees, and risks excessive interference with the right to freedom of expression. Download

  - **Barbora Bukovská**, ARTICLE 19

The full set of papers and the co-chairs report may be downloaded as a single PDF.

The Transatlantic Working Group consists of more than two dozen representatives of government, legislatures, the tech industry, academia, journalism, and civil society organizations in search of common ground and best practices to reduce online hate speech, terrorist extremism and viral deception without harming freedom of expression. The group is a project of the Annenberg Public Policy Center (APPC) of the University of Pennsylvania in partnership with the Institute for Information Law (IViR) at the University of Amsterdam and The Annenberg Foundation Trust at Sunnylands. Additional support has been provided by the Embassy of the Kingdom of the Netherlands.

For a list of TWG members, click here.

Additional reading about these issues and the TWG:

- How (Not) to Regulate the Internet (Peter Pomerantsev, The American Interest, June 10, 2019)
- Regulating the Net is Regulating Us (Jeff Jarvis, Medium, May 31, 2019)
- A Lesson From 1930s Germany: Beware State Control of Social Media (Heidi Tworek, The Atlantic, May 26, 2019)
- What Should Policymakers Do To Encourage Better Platform Content Moderation? (Mark MacCarthy, Forbes, May 14, 2019)
- Proposals for Reasonable Technology Regulation and an Internet Court (Jeff Jarvis, April 1, 2019)
- Protect Freedom of Speech When Addressing Online Disinformation, Transatlantic Group Says (APPC, March 6, 2019)
- Transatlantic Working Group Seeks To Address Harmful Content Online (APPC, Feb. 26, 2019)
- Wake Up Call (Susan Ness, Medium, Oct. 12, 2018)

The Annenberg Public Policy Center (APPC) was established in 1993 to educate the public and policy makers about the media's role in advancing public understanding of political, health, and science issues at the local, state, and federal levels. Follow us on Facebook and Twitter @APPCPenn.

For information on the Privacy Policy of the University of Pennsylvania, please click here.

If you would like to stop receiving news about the Transatlantic Working Group from the Annenberg Public Policy Center of the University of Pennsylvania, click here to send an email to be removed from this list.

TRANSATLANTIC WORKING GROUP

The Ditchley Park Session

First session of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, February 27 – March 3, 2019, at Ditchley Park, U.K.

Co-Chairs Report No. 1

Key findings and recommendations
—*Susan Ness and Nico van Eijk*

Working Papers

Freedom of Expression: A Comparative Summary of United States and European Law
—*Brittan Heller and Joris van Hoboken*

An Analysis of Germany's NetzDG Law
—*Heidi Tworek and Paddy Leerssen*

The Proposed EU Terrorism Content Regulation: Analysis and Recommendations with Respect to Freedom of Expression Implications
—*Joris van Hoboken*

Combating Terrorist-Related Content Through AI and Information Sharing
—*Brittan Heller*

The European Commission's Code of Contact for Countering Illegal Hate Speech Online: An analysis of freedom of expression implications
—*Barbora Bukovská*

TRANSATLANTIC WORKING GROUP

Co-Chairs Report No. 1: The Ditchley Park Session

Susan Ness, Annenberg Public Policy Center
Nico van Eijk, Institute for Information Law, University of Amsterdam

May 2, 2019

Introduction & mission statement

The Transatlantic High Level Working Group on Content Moderation and Freedom of Expression (TWG) held its inaugural meeting at Ditchley Park in the United Kingdom from February 27 to March 3, 2019. Comprised of leading academics, policy makers, and industry representatives, the Working Group convened to discuss the future of freedom of expression in the digital age. This report offers an overview of key outcomes.

Freedom of expression is one of the cornerstones of democracy and international human rights law. Yet this right has never been absolute: democratic societies have deemed certain types of speech so harmful that they are unacceptable. Historically, hate speech and incitement to violence frequently have been subject to restrictions. Deception and false representation have also been found unworthy of protection under certain circumstances.

These types of harmful speech are as old as history, but the internet allows them to propagate at unprecedented speed and scale. Politicians, policy makers, the tech community, and citizens on both sides of the Atlantic are grappling with these new phenomena, considering and often adopting initiatives to restrict "unwanted" content online. Despite best intentions, such efforts have the potential to restrict rightful freedom of expression. The momentum to regulate the (perceived) threats of hate speech and viral deception therefore risks undermining the very democratic systems governments and politicians seek to protect. And despite the internet's transnational reach most measures are considered in national contexts, notwithstanding potential global effects.

The Transatlantic Working Group was formed in response to these trends to develop concrete tools, guidelines, and recommendations to help policy makers navigate the challenges of governing content in the digital age.

Our discussion took into account the many platforms that foster this global conversation – not just the large social media companies and search engines, such as Facebook and Google, which are often the focus of initiatives to address unwanted content, but also smaller European, American and, indeed, global platforms as well as nonprofit, crowd-sourced informational services.

This session of the Transatlantic Working Group explored in depth hate speech and violent extremist content. To this end, we examined four different initiatives designed to address hate speech and incitement to terrorism:

- Germany's Network Enforcement Law (NetzDG)
- The European Union's proposed Terrorism Content Regulation
- The European Union's (voluntary) Code of Conduct for Countering Illegal Hate Speech Online
- The Global Internet Forum to Counter Terrorism's Hash-Sharing Database

Briefing papers from the TWG's examination of these measures are posted on the Institute for Information Law (IViR) website. In addition, our discussion also generated cross-cutting themes and insights, which we discuss below.

The members of the Transatlantic Working Group participating in the Ditchley Park Session may not necessarily agree with or endorse every observation noted below, and undoubtedly have other important ones to add to this summary. But they accept that this report reflects the main points we discussed and agreed in principle during our meeting, with the understanding that additional details and views will be reflected in subsequent publications of the Working Group.

Key findings and recommendations

We encourage policy makers, the tech industry, and other stakeholders to consider these points as they seek ways to address harmful content online without chilling free speech:

Clearly define the problems being addressed, using an evidence-based approach. Policy measures directed at vaguely defined concepts such as "extremism" or "misinformation" will capture a wide range of expression.

- Before taking any steps to restrict speech, regulators should explain clearly and specifically the harms they intend to address, and also why speech regulation is necessary for this purpose.
- The rationale should be supported by concrete evidence, not just theoretical or speculative concerns.
- Any government action should also be subjected to timely review, in order to assess whether it continues to serve its intended purpose. To this end, "sunset clauses" can be an effective tool to encourage a thorough impact review post-implementation.

Build in transparency by government and industry alike so that the public and other stakeholders can assess more accurately the impact of content moderation.

- The industry's Hash-Sharing Database, in particular, was criticized for a lack of transparency into its workings. Germany's Network Enforcement Law (NetzDG), despite other criticism, does include some transparency reporting requirements, but they need to be tightened.
- Generally, government action to direct the content moderation practices of platforms should be documented and available for academic research as well as the public.

- Platforms should also share detailed information about their content moderation practices with the public, working with the public and academics to design such disclosures or databases while respecting the privacy of the people who use their services.

Ensure due process safeguards for online speech.

- When user-generated content is removed, the authors often have limited or no redress. This practice may facilitate unwarranted censorship and abuse or perceptions of arbitrariness. The uploading user should be offered a clear and timely recourse mechanism for considering reinstatement.

- When governments direct action to restrict online speech, their measures should comply with rule of law principles so that they are subject to judicial review; governments should *not* use informal agreements with private platforms to obscure the role of the state and deprive their targets of civil redress.

- Platforms should consider notifying content providers when they receive a formal notice from government to remove that content, so that content generators can appeal the decision with the appropriate authorities. For example, the NetzDG law does not provide for appeal mechanisms, nor are users notified of official complaints levied against their content.

Reimagine the design of both public and private adjudication regimes for speech claims.

- Many online platforms already offer internal, private appeal mechanisms. However, given the democratic values at stake, the lack of judicial oversight and the resulting "privatization" of speech regulation raises concerns. Accordingly, there may be a need to create independent, external oversight from public, peer, or multistakeholder sources.
- The Transatlantic Working Group will continue to explore designs for such external review. Some options include an increased role for independent regulators, specialized judicial "online review systems," and private or multistakeholder "standards council" solutions.
- Courts should continue to play their historical role in developing a body of law through well-reasoned decisions that would provide guidance to platforms, users, and governments.

Craft appropriately tailored policies: one size need not fit all.

- Policy discussions often refer in general terms to "platforms" and/or "online intermediaries," but these concepts are too broad. They cover a wide range of services and operate at different layers of the internet stack, with entirely different abilities (and responsibilities) to moderate online speech. Policy makers should consider the different roles and capacities of these players.

  For example, content restrictions imposed at lower levels of the "stack" (such as Internet Service Providers, CDNs and the Domain Name System) have a greater impact on freedom of expression than at higher levels (such as web forums, social media, chatrooms).
- Size is another important factor: the cost of regulatory compliance disproportionately burdens smaller and nonprofit services, and should be considered when imposing requirements or penalties.

- But, a caveat: regulatory scrutiny of larger platforms has led some bad actors to migrate to smaller platforms or encrypted services, such as 8chan or Gab, where they are less likely to be removed.

Understand the risk of overreliance on automated solutions such as AI, especially for context-specific issues like hate speech or disinformation.

- Automated approaches have had some success, such as in blocking child sexual abuse content and copyrighted material. However, identifying hate speech and disinformation often requires a nuanced assessment of context and intent. While improving, automated systems still generate a significant number of false positives.
- Automated removal can act as a prior restraint, which prevents content from ever being published. Therefore, automated systems should include an adequate number of human reviewers to correct for machine error.
- AI solutions may reinforce biases, since they are trained on historical datasets that reflect broader social contexts. This can lead to unfair and biased outcomes in content moderation, and the further marginalization of certain groups. Online services should probe for and eliminate such biases. Our second and third Working Group Sessions will do a deep dive into artificial intelligence solutions.
- Given the quantity of user-generated content, automated systems necessarily are an important part of the solution. However, policy makers and industry should avoid overstating the power to solve speech problems through technical means, and should incorporate wherever possible qualitative human oversight.

Next steps

Our second Transatlantic Working Group Session in May will examine initiatives to address viral deception (disinformation), especially in the context of elections; self-regulatory models, including the European Commission's "Code of Practice"; emerging regulatory frameworks, including the British Government White Paper; practices surrounding "takedowns"; algorithms and accountability; and will introduce a discussion of intermediary liability.

In the fall, our third and final session will further examine the earlier topics and focus in depth on artificial intelligence and on intermediary liability.

Between these sessions, we will continue to reach out to diverse stakeholders and the public in roundtables and forums for their feedback and engagement.

TRANSATLANTIC WORKING GROUP

Freedom of Expression: A Comparative Summary of United States and European Law[†]

Brittan Heller, The Carr Center for Human Rights Policy, Harvard University
Joris van Hoboken, Vrije Universiteit Brussels and University of Amsterdam[1]

May 3, 2019

Contents

Introduction

This document lays out the legal basics and normative underpinnings of freedom of expression in the transatlantic context. It provides a common starting point for our understanding of freedom of expression for the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression. It will look at the legal sources of freedom of expression in United States,

[†] 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/.

European, and international law, and detail the shared normative foundations on both sides of the Atlantic, while explaining relevant exceptions and legal differences. On this basis, it will help to ground conversations on measures related to viral deception, extremism, and hate speech and their impact on freedom of expression.

Freedom of expression, or freedom of speech in the United States, is a pillar of liberal society and an essential component – if not the central piece – of a healthy democracy. The right to freedom of expression is a global standard, protected under regional and international human rights instruments, treaties, and frameworks. At the outset it must be stressed that it does not exist in isolation. Freedom of expression is inexorably linked with the right to peaceful assembly and association, freedom of thought, conscience, and religion, and the right to privacy, as well as other rights. More generally, freedom of expression is dependent on effective enforcement of the rule of law.

Freedom of expression has a long history, shaped by political, economic, and cultural developments, and is deeply affected by technological change, ranging from the discovery of the printing press to radio broadcasting to digital technologies. With each wave of change in the communication landscape and services, new questions have emerged about how best to articulate the value of freedom of expression and protect people against new forms of government interference as well as undue limitations by private parties.[2] The internet has offered unprecedented opportunities for freedom of expression, while also giving rise to new forms of censorship, control, and threats to participation. Challenges to freedom of expression – such as hate speech and deception/propaganda – are not new. Still, our current landscape presents new questions and opportunities to respond to these challenges in a way that protects fundamental rights.

Why care about freedom of expression?

Throughout history, many political systems have recognized freedom of expression as a central value. However, the normative theories for why it matters have differed. Here we discuss the most important arguments for freedom of expression: the protection of individual liberty and self-fulfillment, the search for truth, the functioning of democracy, and as a check on government power.

First, freedom of expression is seen as the foundation of individual liberty and self-fulfillment. This theory holds that freedom of expression is rooted in the value of human liberty, freedom of choice, and the value of and respect for diversity. Freedom of expression is presumed to be one of the most basic conditions for self-fulfillment in society. It is an end in itself – and as such, deserves society's greatest protection.

Another school of thought emphasizes that freedom of expression is vital to the attainment and advancement of knowledge, and the search for truth. At the heart of this theory is the work of John Stuart Mill, who argued that

> [T]he peculiar evil of silencing the expression of an opinion is, that it is robbing the human race; posterity as well as the existing generation; those who dissent from the opinion, still more than those who hold it. If the opinion is right, they are deprived of the opportunity of

> exchanging error for truth: if wrong, they lose, what is almost as great a benefit, the clearer perception and livelier impression of truth, produced by its collision with error. (Mill, *On Liberty*)

In other words, even unpopular, untrue, or undesirable opinions deserve protection, as their expression allows them to be tested. "The cure for bad speech is more speech" is another popular expression of this theory. Related is the concept of the "marketplace of ideas," which holds that "the best test of truth is the power of the thought to get itself accepted in the competition of the market" (Holmes in *Abrams v. United States*).

Other theories see freedom of expression as essential to democracy, and tend to emphasize the freedom to express and receive information and ideas of societal and political relevance. Freedom of expression is important as it supports political participation, and allows citizens to inform themselves about matters of public concern. The press and the media receive particular attention in this theory, because they serve as a forum for deliberation and a way for the public to inform itself. Habermas's theory of the public sphere also connects freedom of expression to the interest of free and public deliberation on matters of public concern.

Finally, freedom of expression is seen as a check against government overreach and abuse. These theories emphasize the risk of government involvement in matters of speech as well as the importance of creating space for speech that is critical of government actors. In Europe and international law, it is considered the role of the State to create a positive and enabling environment for freedom of expression, pluralism and diversity in light of restrictions by others than public authorities as well.

Sources of Law

a. Freedom of expression at the international level

During its first convening, the UN General Assembly declared that "Freedom of Information is a fundamental human right and … the touchstone of all the freedoms to which the United Nations is consecrated."[3] The **United Nations Universal Declaration of Human Rights**, subsequently adopted in 1948 by the United Nations General Assembly, protects freedom of expression in **Article 19**, which states that:

> Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.[4]

The UN has several specific institutions that aim to promote and enforce human rights (incl. freedom of expression), including the UN Human Rights Council and the Office of the High Commissioner for Human Rights (OHCHR), which coordinates human rights activities in the UN system. The UN General Assembly has adopted numerous resolutions on freedom of expression. Since 1993, the UN system includes a Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (the current Special Rapporteur is David Kaye, who is a member of the

Transatlantic Working Group). The Special Rapporteur conducts country studies as well as annual and thematic reports, including one recently on online content moderation, the role of internet access providers, the role of states and private sector and encryption.[5]

Article 19 of the International Covenant on Civil and Political Rights (ICCPR) provides for more detailed protection of freedom of expression at the international level.[6] It states:

(1) Everyone shall have the right to hold opinions without interference.

(2) Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice.

(3) The exercise of the rights provided for in paragraph 2 of this article carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary:

(a) For respect of the rights or reputations of others;

(b) For the protection of national security or of public order (*ordre public*), or of public health or morals.

The ICCPR imposes a positive obligation on its signatories to "take the necessary steps" to ensure its protection, including adopting "laws or other measures as may be necessary" and providing "an effective remedy" to those whose freedom of expression has been violated. The ICCPR has its own quasi-judicial oversight body, i.e., the United Nations Human Rights Committee. The United States ratified the ICCPR in 1992, but with a number of reservations, including to Article 20 of ICCPR, as discussed below. Generally, through its courts and political institutions, the United States legal system tends to be relatively non-receptive to the influence of international law.

b. Freedom of expression in the United States

In the United States, protection for freedom of expression is codified in the **First Amendment of the U.S. Constitution**, part of the 1791 Bill of Rights:

> Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the government for a redress of grievances.[7]

The Due Process clause of the 14th Amendment carries these federal protections over to individual states. The Supreme Court is the ultimate judicial arbiter of what is or is not protected speech under U.S. law and has been responsible for a body of case law on the First Amendment spanning more than 200 years.

c. Freedom of expression in Europe

In Europe, freedom of expression is protected in foundational instruments from the Council of Europe, and at the European Union level. Additionally, it receives protection at the national level through freedom of expression provisions in national constitutions.

The cornerstone of freedom of expression protection in Europe is **Article 10 of the 1950 European Convention on Human Rights (ECHR).**[8] This guarantees freedom of expression as part of the regional human rights treaty for the Council of Europe region (including non-EU countries, such as Turkey and Russia). Article 10 of the ECHR provides:

(1) Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers. This Article shall not prevent States from requiring the licensing of broadcasting, television or cinema enterprises.

(2) The exercise of these freedoms, since it carries with it duties and responsibilities, may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary.

Compliance with the ECHR is overseen by the European Court of Human Rights (ECtHR) in Strasbourg, which can issue binding rulings for the member states and has through its case law had a major impact on freedom of expression in Europe. The Council of Europe also adopts soft law recommendations. The ECHR is a "living instrument," meaning that the ECtHR takes account of new conditions affecting the exercise of freedom of expression. According to ECtHR doctrine, all rights guaranteed by the ECHR must be "practical and effective" and not merely "theoretical or illusory." All countries in the European Union are signatories of the ECHR.

Under the European Union, Article 11 of the **Charter of Fundamental Rights of the European Union** (2000) (EU Charter) protects freedom of expression in the context of EU law.[9] It provides:

(1) Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers.

(2) The freedom and pluralism of the media shall be respected.

The Charter is a relatively new instrument. It applies exclusively to matters of EU law, as opposed to purely national law in EU Member States. EU law includes important areas for freedom of expression such as intermediary liability and illegal content online. Notably, the EU Charter states that its

safeguards are, at a minimum, equivalent to those of the ECHR. As a result, the case law of the ECtHR is also relevant for interpreting the EU Charter.

National constitutions also include provisions on freedom of expression and may add to the protection afforded by the ECHR or the EU Charter. For example, the Human Rights Act of 1998 embeds the ideals of the ECHR into UK law. The Dutch Constitution has specific provisions on prior restraint (censorship). In Germany, Article 5 in the Basic Law for the Federal Republic of Germany has been applied to relations between private entities, a doctrine called *Drittwirkung*.

Key similarities and differences between U.S. and Europe

a. What is Protected, and the Possibility of Limitations

Freedom of expression encompasses much more than the ability for people to speak without constraints. According to general UN human rights standards, which apply as a baseline in both the U.S. and Europe, freedom of expression has the following key features:

- It applies to everyone equally without distinction or discrimination;
- Its material scope (while not unlimited) is broad and encompasses information and ideas of all kinds, including political, cultural, and also commercial speech;
- It includes protection for information and ideas that may be considered harmful by some, extending protection to information and ideas that "offend, shock or disturb";
- It includes the rights to receive as well as impart information and ideas, thus protecting the rights of both listeners and speakers;
- Everyone is free to impart their ideas using any form of media;
- Its geographical scope is unlimited, as it applies regardless of frontiers.

Freedom of expression is not absolute and can be restricted, in limited circumstances. International, European, and U.S. law all share this common principle of limited state interference for individual expression, constrained to specific circumstances. When restrictions on this right are properly followed, limitations on freedom of expression can be permissible. Historically, restrictions on speech have included that of national security, intellectual property, obscenity, crime, contempt of court, and the protection of official secrets.

Article 19(3) of the ICCPR provides the international-level framework to evaluate restrictions on the freedom of expression. Interferences need to:

- be provided for by law, which needs to be clear and accessible to everyone;
- pursue a legitimate aim: respect for the rights or reputations of others, or the protection of national security, public order, public health or morals;

- be necessary and proportionate, i.e., interferences should deploy the least restrictive means required to achieve the purported aim.

Taken together, these requirements aim to ensure that governmental restrictions on free speech are limited in scope and purpose, narrowly tailored, and include adequate safeguards against abuse. Additionally, Article 20(2) of the ICCPR requires states to prohibit by law "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence," an obligation which is unpacked in the Rabat Plan of Action.[10] The United States has made a reservation with respect to Article 20 ICCPR.

A different set of limitations is included in the International Convention on the Elimination of Racial Discrimination (ICERD), which in Article 4(a) requires States to *inter alia* "condemn all propaganda and all organizations which are based on ideas or theories of superiority of one race or group of persons of one colour or ethnic origin, or which attempt to justify or promote racial hatred and discrimination in any form."[11] ICERD contains much broader positive obligations on member States to prohibit certain types of speech than the ICCPR. Interpretation of the ICCPR and ICERD standards by international and regional bodies has been too inconsistent to offer clear guidance to states on possible limitations on freedom of expression.

European doctrine follows a similar approach to the ICCPR. Under the ECHR, interferences with freedom of expression must follow a three-step test that requires that they are (1) provided by law, (2) pursue a legitimate purpose, and (3) are necessary in a democratic society. The ECHR does not contain any obligation on States to prohibit any form of expression, as under Article 20(2) of the ICCPR. However, the European Court has recognised that certain forms of harmful expression must necessarily be restricted to uphold the objectives of the ECHR as a whole. Under the EU Charter there is an additional requirement for interferences to respect the essence of the right to freedom of expression.

The United States, by contrast, operates under different doctrines. First Amendment analysis tends to take account of the type of speech as well as the type of restriction. First, it considers several narrow categories of expression as "unprotected speech," including incitement, defamation, fraud, obscenity, child pornography, fighting words, threats, defamatory lies (libel or slander), and lying under oath. For protected categories of speech, First Amendment analysis focuses on the type of restriction. Here an important distinction is made between "content-based" restrictions, which regulate speech based on its substance, and "content-neutral" restrictions. Content-based restrictions are subject to strict scrutiny, which renders them presumptively invalid. Content-neutral restrictions are subject to a balancing test, known as "intermediate scrutiny," in order to determine whether the restriction is legally permissible.

Though these approaches in Europe and the United States may appear different, they often reach similar outcomes. Most of the categories of "unprotected speech" identified by the U.S. Supreme Court, such as child pornography and violent threats, also receive no or relatively low protection in Europe. And in deciding whether a regulation is "necessary in a democratic society," the ECtHR is

more likely to accept a time, place, or manner restriction than a content-based restriction. Still, key differences remain. The most relevant for our project are discussed below: the status of hate speech and extremism, the status of disinformation/factually incorrect information, and the status of private entities, including intermediaries.

b. Status of Hate Speech and Violent Extremism

While "hate speech" has no definition under international human rights law, the expression of hatred toward an individual or group on the basis of a protected characteristic can be divided into three categories, distinguished by the response that international human rights law requires from States:

- Severe forms of "hate speech" that international law requires States to prohibit, including through criminal, civil, and administrative measures, under both international criminal law and Article 20(2) of the ICCPR;

- Other forms of "hate speech" that States may prohibit to protect the rights of others under Article 19(3) of the ICCPR, such as discriminatory or bias-motivated threats or harassment; or

- "Hate speech" that is lawful and should therefore be protected from restriction under Article 19(3) of the ICCPR, but which nevertheless raises concerns in terms of intolerance and discrimination, meriting a critical response by the State.

Treatment of hate speech differs across the Atlantic. In the United States, hate speech is fully protected by the First Amendment unless it falls under an exception, most commonly true threats, incitement to violence, or defamation.[12] These are relatively narrow exceptions. For instance, incitement to violence may only be restricted in cases where "directed to inciting or producing imminent lawless action and is likely to incite or produce such action." Accordingly, hate speech enjoys a relatively high level of protection in the United States. Famously, the U.S. Supreme Court has prohibited restrictions on public rallies by the Ku Klux Klan (*Brandenburg v. Ohio*) and by neo-Nazis (*National Socialist Party of America v. Village of Skokie*) on the basis of the First Amendment.

In Europe, countries differ significantly in how they approach and define the topic of hate speech, and in how they apply the above concepts. These variations generate significant inconsistencies in the law and its application across the region and even within countries. The European Court of Human Rights has followed a case-by-case approach. The Court has excluded from the scope of freedom of expression certain extreme forms of expression, including Holocaust denial, as an "abuse of rights." In other situations, it has dealt with these cases through the regular test of Article 10 that interferences need to be necessary in a democratic society. The case law suggests that if there is doubt about certain expression qualifying as hate speech, the ECtHR follows the framework of Article 10 ECHR and its test for interferences with freedom of expression.

Many European countries have distinct laws to combat hate speech.[13] These laws mostly emerged after World War II with the aim to quell hatred pertaining to religion and ethnicity. In France, Section

24 of the Press Law of 1881 criminalizes "racial discrimination hatred, or violence on the basis of one's origin or membership (or non-membership) in an ethic, national, racial, or religious group." The German Penal Code criminalizes a range of expression including hate speech, Holocaust denial, membership in or support of banned political parties, dissemination of means of propaganda of unconstitutional organizations, use of symbols of unconstitutional organizations, and insulting of faiths. The new German NetzDG law creates mechanisms for dealing with those speech offenses in the online environment.

Violent extremism is a term used in governmental programs to counter violence and incitement to violence. It has potential overlap with the concept of hate speech, but from the perspective of human rights and freedom of expression it is different. As stated in the Joint Declaration on Freedom of Expression and Countering Violent Extremism, if the concept is used as a basis for restricting freedom of expression, the concept should be clearly and narrowly defined and restrictions should be "demonstrably necessary and proportionate to protect, in particular, the rights of others, national security or public order."[14]

c. Status of Disinformation

Disinformation (also: fake news, false news, misinformation, viral deception) is not a well-developed concept in freedom of expression law or theory.[15] As emphasized recently at the international level,[16] freedom of expression is "not limited to 'correct' statements" and only in specific circumstances does disinformation map to a category of speech that can be legally restricted. Examples of speech that can be restricted on the basis of factual incorrectness under applicable standards are defamation and false or misleading advertising. Disinformation that amounts to hate speech can be restricted under applicable freedom of expression standards discussed above.

Generally speaking, a distinction has been made between the statement of opinions on the one hand (which cannot be false and may generally not be restricted) and statements of facts, which can be false and can be restricted in narrow circumstances. In *Gertz v. Robert Welch, Inc.*, the U.S. Supreme Court ruled that "there is no such thing as a false idea. However pernicious an opinion may seem, we depend for its correction not on the conscience of judges and juries, but on the competition of other ideas." Similarly, the ECtHR has judged that "the existence of facts can be demonstrated, whereas the truth of value judgments is not susceptible of proof. … As regards value judgments this requirement is impossible of fulfilment and it infringes freedom of opinion itself, which is a fundamental part of the right secured by Article 10."

False statements of fact are treated differently but have generally been accepted as a necessary part of free debate. Strict legal requirements for proving the truth of publications (e.g., by journalists) are considered in violation of freedom of expression as they would keep the media from fulfilling their societal function to inform the public. Under Article 10 of the ECHR, journalists are expected to be "acting in good faith and on an accurate factual basis and provide 'reliable and precise' information in accordance with the ethics of journalism."

Propaganda and disinformation originating from or disseminated by State actors have been of special concern from the perspective of freedom of expression. In Europe and at the international level, this concern has translated into the stipulation of positive obligations on the State to create a favorable environment for expression and abstention from these practices.

d. Status of Private Entities

The status of private parties under freedom of expression is an important area in which the United States doctrine differs from European and international standards. This is of particular relevance to the TWG project, considering the role of platforms in facilitating and (potentially) restricting online expression. Private intermediaries play a central role in most (proposed) regulations and policies in the area of hate speech, extremism and disinformation.

A first question is whether corporations, in contrast to individuals, have a right to freedom of expression. In Europe, although perhaps not widely known or accepted, this is clearly the case, both for Article 10 ECHR and Article 11 EU Charter. In the United States, the protection of corporate speech is particularly broad as evidenced by judgments such as *Citizens United v. FEC* in which certain restrictions on corporate contributions to political campaigns were struck down on the basis of the First Amendment. Under international law, the issue remains contentious.

A second question is whether private entities can have any (legal) obligations under freedom of expression. Under international human rights law, States are the bearers of legal responsibilities. The key guidance on how to apply the international human rights framework to private entities appears in the **Guiding Principles on Business and Human Rights**, also known as the Ruggie Principles.[17] Under this framework, States have a duty to protect while corporate actors have a duty to respect human rights, including freedom of expression. States have a duty to ensure that private entities under their jurisdiction fulfill this obligation. The UN Special Rapporteur on freedom of expression has recently adopted a number of reports that specifically deal with the question of the role and responsibilities of the private sector in view of freedom of expression in the digital age.[18]

In the United States, the First Amendment requires so-called State action. In brief, this means that if there is no relevant exercise of government power, the First Amendment does not apply and no legal case can be brought to court. The First Amendment functions as a *negative right*, protecting private entities from undue interferences by public authorities.[19] The Supreme Court has been very reluctant to create any exceptions to this principle.[20]

In the European context, the situation is more complex. In the first instance, freedom of expression under Article 10 ECHR is similarly focused around State interferences. But in addition, the ECtHR recognizes so-called positive obligations and the possibility of indirect horizontal effect. These doctrines emphasize the need of the state to act to support freedom of expression and provide safeguards against particular forms of private abuse of power that restricts freedom of expression.

Positive obligations are duties for the state to proactively foster the freedom of expression (as opposed to merely refraining from interference). Among the main positive obligations for the State under

Article 10 ECHR are the obligation to guarantee pluralism, to promote the free exercise of the right to freedom of expression and provide for the societal conditions in which free exercise can prosper. These positive obligations do not tend to translate into clear legal rights, but tend to be used as (important) arguments in Article 10 ECHR case law.

Indirect horizontal effect entails the interpretation of private law (such as contract law or property law) in light of the value and need to respect the effective enjoyment of fundamental rights. At the national level, this indirect effect is often effectuated through the interpretation of open norms in private law, such as general duties of care, fault requirements, equity and fairness, or of the interpretation of other norms in light of constitutional guarantees. In the context of the ECHR, indirect horizontal effect is typically effectuated through recognizing positive obligations on the State to protect the enjoyment of fundamental rights in the sphere of relations between individuals or in cases of a complaint relating to a conflict between private parties and competing fundamental rights.

Finally, both in the United States and in the European Union, the law provides for statutory safe harbors that protect online intermediary services (like social media platforms, web hosting services, cloud infrastructure, etc.) from liability for their user's content. In the U.S., the main safe harbors are the **Communications Decency Act** (CDA, 1996) and the **Digital Millennium Copyright Act** (DMCA, 1998).[21] In the Europe Union, the **e-Commerce Directive** harmonizes intermediary liability in the Member States.[22] These laws were created in the 1990s and are seen as important conditions for online innovation and free speech to flourish. The U.S. CDA, Section 230 is by far the strongest safe harbor provision internationally, since it immunizes online intermediaries unconditionally for the speech of others (outside of the area of intellectual property and federal criminal offenses). It also immunizes online intermediaries for decisions to filter and remove speech from their services. In Europe, Article 12-15 ECD provide similar safe harbors, but they have important exceptions: Once an online intermediary obtains knowledge about illegal content, it risks becoming liable. This provides the basis for notice and takedown procedures. Article 15 ECD prohibits states from imposing general duties on intermediaries to monitor their services for illegal content. CDA 230 and the ECD safe harbors have come under growing pressure and the EU is widely expected to revise the ECD in the next Commission period. Both in the United States and in Europe, courts have indicated that some safe harbor protection may be mandatory based on freedom of expression. The Council of Europe has adopted a series of soft-law instruments related to intermediary liability, including a recommendation from 2017.[23]

Appendix: Major Provisions on Freedom of Expression

<u>International</u>

- At the first meeting of the United Nations General Assembly in January 1946, States passed a resolution that recognized freedom of information as a fundamental human right and "the touchstone of all the freedoms to which the United Nations is consecrated."

- It provided an early definition of freedom of expression: "Freedom of information implies the right to gather, transmit and publish news anywhere and everywhere without fetters. As such it is an essential factor in any serious effort to promote the peace and progress of the world."

- In 1948, the **Universal Declaration of Human Rights** was enacted. Article 19 of the Declaration stipulates that: "Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers."

- Freedom of expression is ensured under international law, as codified in Article 19 of the **International Covenant on Civil and Political Rights** (ICCPR), which was enacted by the United Nations in 1966. Article 19 of the ICCPR stipulates that:

    (1) Everyone shall have the right to hold opinions without interference.

    (2) Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice.

    (3) The exercise of the rights provided for in paragraph 2 of this article carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary:

        (a) For respect of the rights or reputations of others;

        (b) For the protection of national security or of public order (*ordre public*), or of public health or morals.

- The **International Covenant on Economic, Social and Cultural Rights** (ICESCR) recognizes the right to freedom of expression under Article 15(3): "The States Parties to the present Covenant undertake to respect the freedom indispensable for scientific research and creative activity."

- There are more treaties that demonstrate an international consensus on freedom of expression, such as the International Convention on the Elimination of All Forms of Racial Discrimination; the Convention for the Rights of the Child; and the International Convention on the Protection of All Migrant Workers and Members of their Families.

<u>European: the Council of Europe</u>

- The **European Convention of Human Rights (ECHR)** was signed in 1950. By now, the ECHR counts 47 signatories, including EU member states but also states as far west as Iceland and as far east as Azerbaijan. Article 10 ECHR protects the right to freedom of expression:

    (1) Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference

by public authority and regardless of frontiers. This article shall not prevent States from requiring the licensing of broadcasting, television or cinema enterprises.

(2) The exercise of these freedoms, since it carries with it duties and responsibilities, may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary.

- In addition, Article 17 ECHR prohibits the abuse of fundamental rights granted under this instrument:

> Nothing in this Convention may be interpreted as implying for any State, group or person any right to engage in any activity or perform any act aimed at the destruction of any of the rights and freedoms set forth herein or at their limitation to a greater extent than is provided for in the Convention.

- The Council of Europe offers additional guidance for the interpretation of these rights in the case law of the European Court of Human Rights (ECtHR) and the 'soft law' instruments of its Parliamentary Assembly and Committee of Ministers.

<u>Europe: the European Union</u>

- The European Union is also subject to the **Charter of Fundamental Rights of the European Union** (also known as the 'European Charter' or 'EU Charter'), created in 2000. This document is binding for all EU Member States, as well as its supranational institutions such as the European Commission and European Parliament.

- Article 11 of the EU Charter protects the freedom of expression and information:

> (1) Everyone has the right to freedom of expression. This right shall include the freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers.
>
> (2) The freedom and pluralism of the media shall be respected.

- Article 52(1) governs restrictions on fundamental rights, including freedom of expression and information:

> Any limitation on the exercise of the rights and freedoms recognised by this Charter must be provided for by law and respect the essence of those rights and freedoms. Subject to the principle of proportionality, limitations may be made only if they are necessary and genuinely meet objectives of general interest recognised by the Union or the need to protect the rights and freedoms of others.

- Article 52(3) requires that the rights laid down in the Charter offer, at a minimum, the same level of protection as the corresponding rights in the European Convention on Human Rights.
- Article 54 of the Charter prohibits abuse of rights:

> Nothing in this Charter shall be interpreted as implying any right to engage in any activity or to perform any act aimed at the destruction of any of the rights and freedoms recognised in this Charter or at their limitation to a greater extent than is provided for therein.

- Additional guidance on the interpretation and enforcement of the EU charter is provided in the opinions of the Court of Justice of the European Union (CJEU) and in the publications of the EU's Fundamental Rights Agency (FRA).

<u>Europe: National Constitutions</u>

- Most if not all European states also have free speech rights codified in their national constitutional systems. What follows are some key examples:
- In France, freedom of expression is protected under Article 11 of the Declaration of the Rights of Man of 1789:

> The free communication of thoughts and of opinions is one of the most precious rights of man: any citizen thus may speak, write, print freely, except to respond to the abuse of this liberty, in the cases determined by the law.

- The French Constitution of 1958 incorporates this declaration as a document of constitutional status.
- In Germany, the Federal Constitution of 1949 (or "Basic Law") protects freedom of expression under Article 5:

> Every person shall have the right freely to express and disseminate his opinions in speech, writing and pictures and to inform himself without hindrance from generally accessible sources. Freedom of the press and freedom of reporting by means of broadcasts and films shall be guaranteed. There shall be no censorship.

- These rights shall find their limits in the provisions of general laws, in provisions for the protection of young persons and in the right to personal honour.
- Arts and sciences, research and teaching shall be free. The freedom of teaching shall not release any person from allegiance to the constitution.
- In the United Kingdom, freedom of expression is guaranteed under the Human Rights Act of 1998. This instrument was enacted as a means to implement the UK's duties under the ECHR.

As such it does not offer its own definition of freedom of expression, but refers to the right defined in the ECHR.

United States

- In the United States, freedom of expression is protected under the First Amendment to the United States Constitution, alongside the freedom of establishment of religion:

> Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances.

- These federal protections also apply to individual states, by virtue of the Due Process clause of the Fourteenth Amendment,
- The United States Supreme Court is the ultimate judicial arbiter of what is or is not protected speech under U.S. law.
- In addition, freedom of expression is also protected under most state constitutions. For instance, the New York Constitution protects freedom of expression under Article 1, Section 8:

> Every citizen may freely speak, write and publish his or her sentiments on all subjects, being responsible for the abuse of that right; and no law shall be passed to restrain or abridge the liberty of speech or of the press. In all criminal prosecutions or indictments for libels, the truth may be given in evidence to the jury; and if it shall appear to the jury that the matter charged as libelous is true, and was published with good motives and for justifiable ends, the party shall be acquitted; and the jury shall have the right to determine the law and the fact.

Notes

[1] Brittan Heller (https://carrcenter.hks.harvard.edu/people/brittan-heller) is a technology and human rights fellow at the Carr Center for Human Rights Policy at the Harvard Kennedy School. She works at the intersection of technology, human rights, and the law, and is an expert on hate speech and the movement from online conduct to offline violence. She also is a senior associate at the CSIS Business and Human Rights Initiative. Joris V. J. van Hoboken (http://www.jorisvanhoboken.nl/) is Professor of Law at the Vrije Universiteit Brussels (VUB) and a senior researcher at the Institute for Information Law (IViR), University of Amsterdam. At VUB, he is appointed to the "Fundamental Rights and the Digital Transformation" Chair, which is established at the Interdisciplinary Research Group on Law Science Technology & Society (LSTS), with the support of Microsoft. Research assistance on this paper was contributed by Paddy Leerssen, a doctoral candidate at IViR and a non-resident fellow at the Stanford Center for Internet & Society. We received helpful comments and input from Barbora Bukovská, Heidi Tworek, Peter Chase and Susan Ness, in addition to discussions on a previous version of this document at the Amsterdam meeting in October 2018.

[2] The question of the status of private parties under freedom of expression is addressed in the end of Section 4.

[3] The concepts of freedom of information and freedom of communication are directly related to freedom of expression. These concepts were partly used to stress the importance of the free use of new information and communication

infrastructures, such as telecommunications, and were prominent in the debates that informed the UNDHR in the 1940s.

[4] United Nations Universal Declaration of Human Rights (1948). www.un.org/en/udhrbook/pdf/udhr_booklet_en_web.pdf

[5] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (2018). http://daccess-ods.un.org/access.nsf/Get?Open&DS=A/HRC/38/35&Lang=E
Research Paper 1/2018, of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression Encryption and Anonymity follow-up report.
https://www.ohchr.org/Documents/Issues/Opinion/EncryptionAnonymityFollowUpReport.pdf

[6] International Covenant on Civil and Political Rights (1966),
https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx

[7] US Constitution Bill of Rights (1791).
https://en.wikisource.org/wiki/Constitution_of_the_United_States_of_America

[8] European Convention on Human Rights (1953). https://www.echr.coe.int/Documents/Convention_ENG.pdf

[9] Charter of Fundamental Rights of the European Union (2000). www.europarl.europa.eu/charter/pdf/text_en.pdf

[10] Rabat Plan of Action, Annex to UNHCR Report on the experts workshop on the prohibition of incitement to national, racial, or religious hatred A/HRC/22/17/Add.4 (2013).
https://www.ohchr.org/Documents/Issues/Opinion/SeminarRabat/Rabat_draft_outcome.pdf

[11] International Convention on the Elimination of All Forms of Racial Discrimination (1965).
https://www.ohchr.org/en/professionalinterest/pages/cerd.aspx

[12] As most hate speech will be qualified as opinion, defamation will generally not be relevant.

[13] For a recent overview of approaches to hate speech in Europe, see Article19, "Responding to 'hate speech': Comparative overview of six EU countries," 2018. https://www.article19.org/wp-content/uploads/2018/03/ECA-hate-speech-compilation-report_March-2018.pdf

[14] The United Nations (UN) Special Rapporteur on Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe (OSCE) Representative on Freedom of the Media, the Organization of American States (OAS) Special Rapporteur on Freedom of Expression and the African Commission on Human and Peoples' Rights (ACHPR) Special Rapporteur on Freedom of Expression and Access to Information, Joint Declaration on Freedom of Expression and Countering Violent Extremism (2016).
https://www.article19.org/data/files/medialibrary/38355/Joint-Declaration-on-Freedom-of-Expression-and-Countering-Violent-Extremism-2016.pdf

[15] For a brief discussion, see e.g. T. McGonagle, "'Fake news': False fears or real concerns?," Netherlands Quarterly of Human Rights 2017, Vol. 35(4), p. 203-209.

[16] The United Nations (UN) Special Rapporteur on Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe (OSCE) Representative on Freedom of the Media, the Organization of American States (OAS) Special Rapporteur on Freedom of Expression and the African Commission on Human and Peoples' Rights (ACHPR) Special Rapporteur on Freedom of Expression and Access to Information, Joint declaration on freedom of expression and "fake news," disinformation and propaganda (2017).
https://www.osce.org/fom/302796?download=true

[17] Guiding Principles on Business and Human Rights (2011).
https://www.ohchr.org/documents/publications/GuidingprinciplesBusinesshr_eN.pdf

[18] Report of the Special Rapporteur to the Human Rights Council on Freedom of expression, states and the private sector in the digital age A/HRC/32/38 (2016). Report of the Special Rapporteur to the human rights council on the role of digital access providers A/HRC/35/22 (2017). Report of the Special Rapporteur to the Human Rights Council on online content regulation A/HRC/38/35.
Each report available online here: https://www.ohchr.org/en/issues/freedomopinion/pages/annual.aspx

[19] "Negative right" indicates a right against government interference, as opposed to a positive right to enlist government protection or support (a "positive right").

[20] One could note here that expressive torts (intentional infliction of emotional distress, privacy violations and defamation) have been constitutionalized.

[21] US Communications Decency Act (1996). https://fas.org/sgp/crs/misc/LSB10082.pdf

[22] e-Commerce Directive (2000), https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32000L0031

[23] Council of Europe Recommendation CM/Rec(2018)2of the Committee of Ministers to Member States on the roles and responsibilities of internet intermediaries
https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680790e14

TRANSATLANTIC WORKING GROUP

An Analysis of Germany's NetzDG Law†

Heidi Tworek, University of British Columbia
Paddy Leerssen, Institute for Information Law, University of Amsterdam[1]

April 15, 2019

Contents

Introduction: What is the NetzDG?

Germany's Network Enforcement Act (*Netzwerkdurchsetzungsgesetz* or NetzDG) entered into full force on January 1, 2018. Known colloquially as a "hate speech law," it is arguably the most ambitious attempt by a Western state to hold social media platforms responsible for combating online speech deemed illegal under the domestic law. Other countries, like France, are using NetzDG as a basis for proposed legislation, making analysis of NetzDG urgent and important.[2]

While NetzDG has encouraged accountability and transparency from large social media platforms, it also raises critical questions about freedom of expression and the potential chilling effects of legislation. During a first meeting at Ditchley Park, UK (February 28-March 3, 2019), the Transatlantic Working Group (TWG) analyzed NetzDG based on an earlier draft of this document. The current version was revised to incorporate crucial insights from those discussions. This updated document introduces NetzDG's content and political context and discusses its implications for freedom of

† 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/.

expression. It closes with next steps and recommendations for more research and transparency, while suggesting how to mitigate the troubling elements of the law.

At the outset, it is important to clarify that NetzDG does not actually create new categories of illegal content. Its purpose is to enforce 22 statutes in the online space that already existed in the German criminal code and to hold large social media platforms responsible for their enforcement. The 22 statutes include categories such as "incitement to hatred," "dissemination of depictions of violence," "forming terrorist organizations," and "the use of symbols of unconstitutional organizations." NetzDG also applies to other categories, such as "distribution of child pornography," "insult," "defamation," "defamation of religions, religious and ideological associations in a manner that is capable of disturbing the public peace," "violation of intimate privacy by making photographs," "threatening to the commission of a felony" and "forgery of data intended to provide proof."

NetzDG targets large social network platforms, with more than 2 million users located in Germany. It requires these platforms to provide a mechanism for users to submit complaints about illegal content. Once they receive a complaint, platforms must investigate whether the content is illegal. If the content is "manifestly unlawful," platforms must remove it within 24 hours. Other illegal content must be taken down within 7 days. Platforms that fail to comply risk fines of up to €50 million.

NetzDG also imposes transparency requirements. If a platform receives more than 100 complaints per year, it is required to publish semi-annual reports detailing its content moderation practices. The act stipulates in some detail what types of information must be included.[3] The first round of reports was published in June 2018; the second round appeared in early 2019. Their results are discussed in further detail below.

Freedom of Expression Implications

Keywords:

- Over-removal
- Privatized enforcement
- Definition of "unlawful content"
- The Streisand effect (counterproductive outcomes of censorship)
- Inspiration for authoritarian regimes around the world to restrict speech

During its development and implementation, NetzDG triggered fierce debate and widespread concern about its implications for freedom of expression. Then Justice Minister (and current Foreign Minister) Heiko Maas from the SPD presented it as a means to tackle online hate speech and viral deception: "The freedom of expression also protects offensive and hateful statements. But it is not an excuse to commit crimes."[4] The measure seems popular with German voters; one poll in March 2018 showed an approval rate of 87%, including 67% who "strongly approved" of the law, and only 5% disapproval.[5] NetzDG's proponents included Stefan Heumann, co-director of the digital policy think tank Stiftung Neue Verantwortung, who emphasized Germany's constitutional tradition of minority protection, the rule of law online, and German public support.[6]

Criticism from the tech industry, activists, and academics seemed to outweigh support.[7] Although the law excludes journalistic platforms, the German Journalists Association joined civil rights activists, academics, and lawyers in signing a joint statement warning that the law "jeopardizes the core principles of free expression."[8] The Global Network Initiative (GNI), a multi-stakeholder self-regulatory body funded by social media companies, e.g., Google and Facebook, asserted that the law "poses a threat to open and democratic discourse."[9] Other critics included Wikimedia Deutschland, the Internet Society, and the German Startups Association. Despite their widespread objections, the law was drawn up swiftly and passed before the German elections in October 2017 with little time to consult civil society organizations or experts.

The first concern surrounding freedom of expression was that NetzDG would encourage the removal of *legal* content, also known as "over-removal." Online platforms, it was argued, would not have the expertise or time to assess every complaint in detail. Making such legal assessments typically requires significant expertise in German language and jurisprudence, as well as complex case-by-case analysis and investigation. Given these costs as well as NetzDG's tight deadlines and heavy fines, platforms would have a strong incentive simply to comply with most complaints, regardless of their actual merits. This would lead to over-removal.

Relatedly, critics objected to NetzDG as an instance of "privatized enforcement" because, rather than courts or other democratically legitimated institutions, platforms assess the legality of content.[10] The NetzDG process does not require a court order prior to content takedowns nor does it provide a clear appeals mechanism for victims to seek independent redress. Wenzel Michalski, Germany Director at Human Rights Watch, argued that the law "turns private companies into overzealous censors to avoid steep fines, leaving users with no judicial oversight or right to appeal."[11]

Many critics also disagree with the substance of Germany's speech prohibitions as too broadly defined or simply wrong on principle. As Article 19, a free speech advocacy group, put it, several provisions "should not be criminal offences in the first place," including blasphemy, broad definitions of "hate speech," criminal defamation and insult.[12] There were broader worries about the law's potential effects on German democracy: prominent cases of deletion might fuel anti-government sentiment or publicize the deleted material far more widely, a phenomenon often known as the Streisand effect.

Critics of NetzDG were typically positive about NetzDG's transparency requirements. Although Article 19 requested the repeal of the law itself, it has asked that transparency requirements be maintained in a separate act.[13]

The law symbolized a deeper disagreement on the role of free speech in democracy. Many West German politicians — conservatives and Social Democrats alike — believed in a "militant democracy" (*wehrhafte Demokratie*), where free speech could be constrained to protect democratic norms. As historian Udi Greenberg has put it, "curbing rights became synonymous with the democratic order."[14] West Germany was the only country in postwar Europe to ban both nationalist parties and the Communist Party in the early 1950s. During the creation of NetzDG, Heiko Maas drew explicitly on Germany's Nazi past and the tradition of militant democracy to assert that "freedom of speech has boundaries."[15]

Beyond Germany, there were concerns that NetzDG would serve as a blueprint or precedent for authoritarian regimes to repress online speech. The Global Network Initiative claimed that NetzDG "posed unintended but potentially grave consequence for free expression in Germany, across the EU, and worldwide" and expressed concerns that it may "empower authoritarian leaders."[16] For example, Russia copied passages from NetzDG in mid-2017 for an anti-terror law that required internet and telecoms providers to save the contents of all communications for six months.[17]

Finally, a striking aspect of the NetzDG debate is that the law's implementation and subsequent transparency reports do not seem to have changed many minds. Proponents see the law as a vital measure to rein in platforms and implement German law online, often with an implied broader goal of preserving German democracy. Critics worry about the freedom of expression implications. The battle lines drawn before the law's implementation remain entrenched.

The NetzDG in Practice: What Does the Evidence Show?

A few days after NetzDG came into force in January 2018, prominent AfD (Alternative für Deutschland) politician Beatrix von Storch saw one of her social media posts removed from Twitter and Facebook under the law. Widespread media coverage of this incident, including the post's content and its potential illegality, seemed to confirm fears of the Streisand effect, or what one journalist dubbed the Storch effect.[18] The AfD has marshalled NetzDG as part of a broader argument that its voice and opinions are being silenced.

Germany's liberal party, the FDP, has claimed that NetzDG's restrictions on freedom of expression violated the German constitution. Senior FDP politicians stated that they refrain from posting on social media because of NetzDG. After their parliamentary initiative to revoke the law failed, the FDP sued the government in November 2018 seeking repeal of NetzDG. The case is currently pending before the Cologne Administrative Court (*Verwaltungsgericht*), though some experts give it little chance of success because the FDP has not produced specific examples.[19]

The best evidence to date about the specific effects of NetzDG comes from the law's transparency requirements. Four major online platforms released their first transparency reports in June 2018: Google (i.e., YouTube), Facebook, Twitter, and Change.org. This provoked another round of debate about the law's impact and efficacy. Perhaps unsurprisingly, opinion remains divided.[20]

Before delving into the transparency reports, it is important to note that these data only cover removal decisions arising from NetzDG complaints, and do not account for other removals based on other types of complaints, referrals, or injunctions.[21] Furthermore, the metric of takedowns does not reveal whether NetzDG has achieved its purpose of combating hate speech and other online excesses. The differences between complaint mechanisms and the reports themselves make certain types of comparison difficult. It is also hard to know how the volume of content removal compares to the overall volume of illegal speech online. Some key results are summarized below:

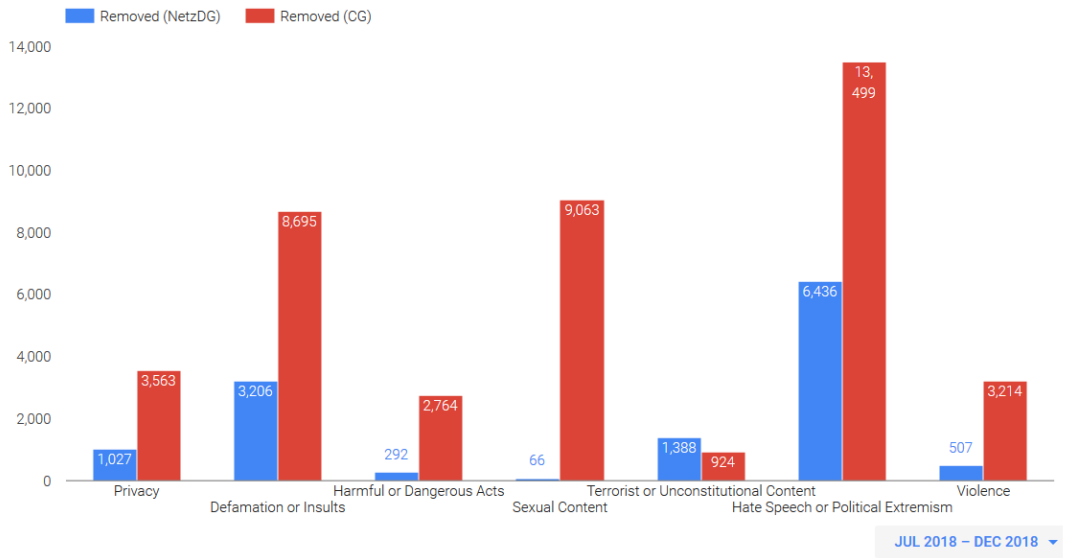
Table 1. Overview of reported numbers by platform

Platform	Total items reported	Total Removal Rate	Removal within 24 hrs
Facebook	1,704	362 (21.2%)	76.4%
Google (YouTube)	241,827	58,297 (27.1%)	93.0%
Twitter	264, 818	28,645 (10.8%)	93.8%
Change.org	1,257	332 (26.4%)	92.7%

Source: Echikson & Knodt 2018[22]

Despite Facebook's size, it received significantly fewer reports than YouTube and Twitter. More than 100 times fewer, in fact, because Facebook's NetzDG complaint form was relatively hard to access. YouTube and Twitter integrated NetzDG complaints into their regular "flagging" interface, which can be accessed through direct links next to every piece of content. Facebook, by contrast, placed their complaint form on a separate, less prominent page, requiring multiple clicks to access. The report data suggest that this design choice had massive impacts on the actual uptake of NetzDG complaints.

It is important to note that most of the takedowns resulting from NetzDG complaints removals appear to have occurred under the companies' community guidelines (or "terms of service"), rather than the German speech laws which NetzDG is intended to enforce. Google, Facebook, and Twitter all prioritize compliance checks with their community guidelines; with each complaint, they first consider whether it violates their community standards. Any content that fails this check is removed. Only the content that passes is then considered for removal under one of the 22 statutes of the German criminal code enforced by NetzDG. Accordingly, as Google's transparency report shows, a majority of removal decisions are based on the platform's private standards, and not on German speech laws. Facebook and Twitter do not specify this data in their reports, but they do review complaints in the same order, prioritizing community guidelines.



Community Guideline enforcement versus NetzDG statutes (Source: Google)

In this light, it may be that NetzDG's most important effect was to ensure swifter and more consistent removal of content within Germany under the companies' community guidelines.

The transparency reports also break down complaints by content type. The data (see Appendix) show that cases involving hate speech and defamation/insult were the most common. For Google, complaints related to hate speech and political extremism were most common (75,892 items), followed by defamation or insult (45,190 items), and sexual content (27,308 items).[23] Facebook and Twitter break down their data differently than Google, making direct comparison difficult. Google aggregates data for comparable offences, such as "insult" and "defamation" or "incitement to hatred" and "propaganda for unconstitutional organizations," whereas Facebook and Twitter list the data for each specific criminal offence separately. Furthermore, Facebook and Twitter's data focus on the number of *complaints*, whereas Google's focus on the number of *content items* (these do not correspond since one complaint may refer to multiple content items). Still, it is clear that insult-related offences were the most common on Facebook; with 460 complaints, insult is cited in over half of all complaints. Twitter provides the most detailed overview, since it also breaks down compliance rates for individual content types. Its data also show incitement to hatred (*Volksverhetzung*) leading with 82,095 complaints, followed by insult (*Beleidigung*), with 75,925.[24]

The major platforms consulted outside counsel with comparable frequency (for fairly few cases): 40 cases for Google, 54 for Facebook, and 107 for Twitter. Each of these platforms was also a member of the self-regulatory advisory body on protecting youth online, Freiwillige Selbstkontrolle Multimedia-Diensteanbieter e.V. (FSM), though the platforms do not seem to have consulted FSM up to this point. FSM and eco, another advisory body, have hotlines for consumer complaints, which they then forward to the companies for review. Google provides feedback about its decisions back to the hotline.[25]

The transparency reports also outline the companies' procedures for notifying uploaders and submitters. They follow the same pattern: submitters receive a confirmation once the complaint is received, and another update once a decision has been made. Uploaders are notified only if and when a content removal decision is taken; they are not notified when a complaint is received about them, or when such complaints are dismissed.

Google's report provides another interesting data point: interactions with uploaders and submitters. In some cases, the platform might require additional information from the uploader to assess whether an alleged falsehood is indeed untrue (e.g., for defamation cases). Google received 2,380 incomplete requests, for which it needed more information from the requester. Remarkably, Google seems to have reached out to uploaders zero times, ostensibly because "the majority of legal NetzDG complaints are unsubstantiated (even after YouTube explicitly asks for further information)."[26] Twitter and Facebook do not address this issue in their reports.

A second round of reports was released in late January 2019. They do not reveal any major changes. The three major platforms all registered decreases in their total number of complaints received. Google & Twitter posted relatively minor changes (15% or less), whereas Facebook, already an outlier, declined more significantly. Removal rates also appear to have dropped for most types of complaints. Perhaps the most significant changes occurred at Twitter, where the number of complaints dropped by over 50% for cases involving child pornography and certain forms of hate speech (*Volksverhetzung*).

Overall, the inconsistencies among companies' complaint forms make comparison of these numbers difficult. A problematic side effect of the reports could be that the number of takedowns or number of complaints become a metric to measure the law's efficacy; these takedowns might be ineffective or even counterproductive in combating the overall prevalence of hate speech. The law's actual impacts on hate speech may be difficult to prove empirically, since this complex phenomenon is influenced by countless other factors as well, including political, cultural, demographic, and economic shifts. The Federal Office for the Protection of the Constitution claims that NetzDG is hindering recruitment efforts by far-right extremists, since it has led to the shutdown of several influential Facebook pages and forced their organizers to move to other less visible platforms.[27] However, the office has not published data to support its claim. Furthermore, these same censorship measures may well have other unintended consequences for counter-extremism, such as radicalization of the censored parties and interference with counter-speech and surveillance efforts. In short, it will require much more research — and greater access to data — to determine whether NetzDG is achieving its aim, and whether any benefits outweigh the harms to free speech.

Outlooks and Next Steps

There is little evidence that the data published in NetzDG's transparency reports has changed anyone's mind about the law. Opinion remains divided, along the lines of deeply held beliefs about the constitutional merits of criminal prohibitions in areas like hate speech and the role of nation-states in content regulation. Beyond repeal or narrowing the scope of the law to remove some of the categories deemed unlawful speech, NetzDG suggests some general lessons for policymakers about the challenges of regulating social media. The recommendations below focus specifically on NetzDG, though the High Level Working Group will release further papers over the next few months that examine some of these suggestions in a broader, transatlantic context.

Transparency and Research

A major challenge is that we need further research on the impact of NetzDG, such as potential chilling effects, but researchers do not have the data to conduct that research. Takedown metrics are problematic as a measure of success because they may simply encourage over-removal. There are multiple fundamental empirical questions that require further investigation. First, how do moderation practices affect the underlying problems and harms? Second, how does a law like NetzDG affect moderation practices and what metrics would enable us to measure that effect? Third, what are the broader societal effects or unintended consequences of such laws?

NetzDG's transparency requirements offer a starting point to enable more robust research. Transparency is the one part of the law that has received almost universal support. Some of the most useful aspects of the transparency reports were about companies' procedures, rather than the raw numbers of complaints or takedowns. The transparency reports showed that NetzDG is more a "community guidelines enforcement law" than anything else. Future transparency reports might contain more details on how companies train moderation staff, operate appeals mechanisms, and conduct quality control.

Still, NetzDG's transparency mechanisms show room for improvement. One common complaint is the lack of standardized reporting formats; this hinders direct comparisons between different reports.[28] Another area of improvement would be to offer more granular data. For instance, the reports specify compliance rates and the number of takedowns by content type, but they do not always break down compliance rates by content type. In other words, it is difficult to assess whether platforms were more likely to comply with hate speech claims than defamation claims.

A more fundamental complaint is that the reports offer little insight into individual cases, and make it hard for third parties to evaluate the merits of platform's decisions. For this reason, the Green Party has proposed to develop a "Clearing House" (*Clearingstelle*) which would publish complaints from users who disagreed with platforms' handling of their NetzDG cases.[29]

A more ambitious approach would be for the German government to encourage and support the creation of research repositories that would combine data from multiple platforms. The repository could start by collecting *all* deleted content or all complaints under NetzDG. Both platforms and enforcement authorities within governments could be more transparent in this regard. German ministries could lead the way by making transparent their complaints bodies' submissions. The repository could serve multiple purposes simultaneously. First, it could facilitate appeals processes. Second, it could provide data to researchers and more broadly enable regulators, researchers, or other watchdogs to review platform decisions. These types of disclosure could reveal important information about whether and how the law harms freedom of expression. Eventually, the repository might also expand beyond NetzDG.

There are already some suggestions about how to create such a repository. U.S. Senator Mark Warner, for instance, suggested a Public Interest Data Access Bill in August 2018.[30] Any such repository would have to address social media companies' legitimate legal and business concerns about sharing proprietary information or users' private data. This could follow similar access procedures to researching confidential medical or statistical records, something that many governments already facilitate.

Due Process and Design Structure

Within the constraints of the current NetzDG, there are several ways to improve users' rights and ensure due process both on platforms and more generally. First, NetzDG highlights the importance of design thinking for user-facing interventions, such as complaint mechanisms. YouTube and Twitter's user-friendly implementations of NetzDG complaint forms led to significantly higher uptake than Facebook's. The German Green Party, the Counter Extremism Project, and the Center for European Policy Studies are currently advocating for additional rules about accessibility in NetzDG complaint forms.[31] If implemented, such rules could be an interesting case study in creating tools that are not just accessible, but visible and easy to use.

Second, NetzDG offers little recourse to uploaders who believe their content has been wrongfully deleted. The Green Party proposes a reinstatement procedure (*Wiedereinstellungsverfahren*), where uploaders could appeal to the platform to have their content reinstated.[32] Others have suggested an appeals procedure before an independent oversight body or a judge.

A third way to safeguard the freedom of expression would be to improve the notification rights of uploaders; NetzDG's transparency reports show that none of the major platforms notifies uploaders about possible complaints submitted about their content. Only the complaint issuer is informed throughout the process, and the uploader who is the subject of the complaint is only informed if and when the platform decides to remove their content. The current system only allows for uploaders to appeal after their content is taken down. It may be worth considering whether to notify uploaders immediately and not to wait until a decision is reached to enable those people to defend their post as lawful. Improving notice rights could enhance procedural fairness, and adjust the playing field in favor of free speech.

Multi-Stakeholder Relationships

More broadly, the question remains how and whether outside stakeholders can and should be involved in platform content moderation processes or their regulation. Like the European Commission's Code of Conduct on Hate Speech, the NetzDG law itself was drawn up in 2017 with little to no input from civil society organizations. A law is also a rather static instrument for addressing such a swiftly changing ecosystem.

Beyond the specifics of NetzDG, the law also suggests wider lessons about regulation and platforms. Platforms may consider how better to contribute to a positive framework of cooperation among social media companies, governments, and civil society organizations. Germany's approach has seemed to illustrate that, currently, the only way countries outside the U.S. receive sustained attention from social media companies is if they are a massive market (like China or the European Union) or journalists uncover significant human rights violations or they threaten companies with significant fines. Real financial liability commanded platform companies' attention. Germany had tried a voluntary compliance system with the companies since 2015 but found it ineffective. The German government chose the path of law only after it deemed the companies insufficiently compliant. A government study in early 2017 found that YouTube had deleted 90% of criminal content, Facebook 39%, and Twitter only 1% within the requested 24 hours (though there are broader questions from a freedom of expression perspective about whether takedown compliance is an appropriate metric).[33] Since the introduction of the law, transparency reports indicate that compliance rates are far higher.

There are many other potential approaches to redefine the relationships between platforms, governments, civil society, and users. One is to design robust transparency mechanisms that enable research on the pressing questions about social media and their broader societal effects before governments undertake any regulation. Another approach is the creation of independent appeals bodies for takedown decisions. A third is to consider rapid-response judicial mechanisms to adjudicate complaints. A fourth is the creation of Social Media Councils that regularly convene platforms, government, and civil society organizations to share information and debate possible new approaches.

Overall, it is hard to predict and measure the full effects of legal policies. The narrow focus on the number of complaints and narrow, problematic categories of illegal speech tell us little about any potential larger effects on Germany's information ecosystem or political discourse. German politicians drew lessons from history to try to protect democracy by curtailing free speech. In the long run, however, they must be careful not to undermine the freedoms embedded in the political system that they seek to protect.

Appendix: Links to transparency reports

Google:

- Transparency Report: Removals under the Network Enforcement Law

Facebook:

- Facebook NetzDG Transparency Report, July 2018
- Facebook NetzDG Transparency Report, January 2019

Twitter (NB: Only available in German):

- Twitter Netzwerkdurchsetzungsgesetzbericht: Januar-Juni 2018
- Twitter Netzwerkdurchsetzungsgesetzbericht: Juli-Dezember 2018

Notes

[1] Dr. Heidi Tworek (heidi.tworek@ubc.ca) is assistant professor at the University of British Columbia as well as a non-resident fellow at the German Marshall Fund of the United States and the Canadian Global Affairs Institute. Her book, *News from Germany: The Competition to Control World Communications, 1900-1945*, was published in March 2019 by Harvard University Press. Paddy Leerssen is a doctoral candidate at the Institute for Information Law (IViR) at the University of Amsterdam and a non-resident fellow at the Stanford Center for Internet & Society.

[2] https://www.buzzfeednews.com/article/lesterfeder/france-facebook-anti-semitism-hate-speech

[3] e.g. the platform's content removal procedures; the number of complaints received and their source (users or authorities); the number of content removal decisions based on these complaints as well as the reason for removal; and their personnel and other resources dedicated to content moderation.

[4] http://www.spiegel.de/netzwelt/netzpolitik/netzdg-heiko-maas-verteidigt-netzwerkdurchsetzungsgesetz-gegen-kritik-a-1186118.html

[5] https://daliaresearch.com/blog-germans-approve-of-social-media-regulation-law/

[6] https://www.csmonitor.com/World/Europe/2018/0408/Is-Germany-s-bold-new-law-a-way-to-clean-up-the-internet-or-is-it-stifling-free-expression

[7] Critics include local digital rights organizations such as Netzpolitik and Digitale Gesellschaft, global international groups such as CDT, Access Now, Article 19, and European Digital Rights (EDRi), Reporters Without Borders, and Human Rights Watch. One prominent German academic critic is Wolfgang Schulz, Director of the Hans Bredow Institute for Media Research. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3216572

[8] http://deklaration-fuer-meinungsfreiheit.de/en/

[9] http://globalnetworkinitiative.org/proposed-german-legislation-threatens-free-expression-around-the-world/

[10] http://deklaration-fuer-meinungsfreiheit.de/en/

[11] https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law

[12] https://www.article19.org/wp-content/uploads/2017/09/170901-Legal-Analysis-German-NetzDG-Act.pdf, p. 2.

[13] https://www.article19.org/wp-content/uploads/2017/09/170901-Legal-Analysis-German-NetzDG-Act.pdf

[14] Udi Greenberg, The Weimar Century: German Émigrés and the Ideological Foundations of the Cold War (Princeton: Princeton University Press, 2014), 204.

[15] "Die Meinungsfreiheit hat auch Grenzen," Medienpolitik.net, 09.01.17.

[16] https://globalnetworkinitiative.org/wp-content/uploads/2018/06/GNIAnnualReport2017.pdf , p. 14.

[17] https://www.heise.de/tp/features/Russland-kopiert-deutsches-Netzwerkdurchsetzungsgesetz-3773642.html

[18] https://www.zeit.de/digital/2018-01/netzdg-meinungsfreiheit-internet-soziale-medien-debatte

[19] https://www.handelsblatt.com/politik/deutschland/netzdg-reine-pr-nummer-fdp-erntet-fuer-netzdg-klage-scharfe-kritik/22672176.html?ticket=ST-4067079-IRYak1DUBdWxocughGHI-ap5

[20] Opposing: https://www.handelsblatt.com/politik/deutschland/gesetz-gegen-hass-im-netz-it-verband-bitkom-zweifelt-an-erfolg-des-netzdg/23793674.html?ticket=ST-111540-bYmBGrN4QHZ0lgiOqv4U-ap5
https://www.handelsblatt.com/politik/deutschland/digitalisierung-internetverband-warnt-regierung-vor-falschen-regulierungsansaetzen-in-der-digitalpolitik/23785968.html
https://www.djv.de/startseite/profil/der-djv/pressebereich-download/pressemitteilungen/detail/article/djv-fordert-abschaffung.html
https://www.reporter-ohne-grenzen.de/pressemitteilungen/meldung/netzdg-fuehrt-offenbar-zu-overblocking/
Supporting:
https://www.n-tv.de/politik/NetzDG-erschwert-Rechten-Rekrutierung-article20597308.html
https://www.handelsblatt.com/politik/deutschland/gesetz-gegen-hass-im-netz-nicht-alles-perfekt-aber-vieles-gut-justizministerium-zeigt-sich-zufrieden-mit-dem-netzdg/23752306.html?ticket=ST-473131-RXLg0ra77rrsPHDkzL7f-ap1
https://www.cducsu.de/themen/innen-recht-sport-und-ehrenamt/netzwerkdurchsetzungsgesetz-wirkt
https://www.spdfraktion.de/presse/pressemitteilungen/netzdg-wirkt

[21] For takedowns unrelated to NetzDG, many platforms offer additional transparency reports on a voluntary basis.

[22] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3300636

[23] See Appendix for complete data. Source: https://transparencyreport.google.com/netzdg/overview?hl=en

[24] See Appendix. (Note: these figures for Twitter are the sum of "Beschwerden von Nutzern" + "Beschwerden von Beschwerdestellen").

[25] https://transparencyreport.google.com/netzdg/youtube

[26] https://transparencyreport.google.com/netzdg/overview?hl=en

[27] https://www.n-tv.de/politik/NetzDG-erschwert-Rechten-Rekrutierung-article20597308.html

[28] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3300636

[29] http://dip21.bundestag.de/dip21/btd/19/059/1905950.pdf

[30] https://www.scribd.com/document/385137394/MRW-Social-Media-Regulation-Proposals-Developed

[31] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3300636
http://dip21.bundestag.de/dip21/btd/19/059/1905950.pdf

[32] http://dip21.bundestag.de/dip21/btd/19/059/1905950.pdf

[33] Federal Ministry of Family Affairs and Federal Ministry of Justice and Consumer Protection, "Löschung von strafbaren Hasskommentaren durch soziale Netzwerke weiterhin nicht ausreichend" (2017).

The Proposed EU Terrorism Content Regulation:

Analysis and Recommendations with Respect to Freedom of Expression Implications†

Joris van Hoboken, Vrije Universiteit Brussels and University of Amsterdam[1]

May 3, 2019

Contents

Introduction and recommendations

In the last two decades, the use of internet communications and related services for terrorism[2] (the live streaming of the Christchurch mosque shooting and subsequent viral distribution through white supremacist networks being the latest high-profile example[3]) has been a major area of concern for government regulation of the internet. A series of European Union legislative and policy initiatives has defined new terrorism-related crimes at the EU level, including policies for law enforcement and the responsibility of online service providers. Most recent is the new proposal for a so-called Terrorism Content Regulation (TERREG). The European Commission proposed this measure in September 2018 and it is currently under debate in the European Parliament and the Council.[4]

---

† 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/.

The Transatlantic Working Group (TWG) used part of its first meeting at Ditchley, UK, to discuss the strengths and weaknesses of the TERREG proposal on the basis of an earlier version of this document.[5] This document has been updated to reflect crucial insights from these discussions as well as recommendations in light of the ongoing debate about the proposal at the EU level. Taking into account these discussions in the TWG, the central conclusions and recommendations are as follows:

- It is essential, in particular given the difficulty of defining terrorism in the first place, that legislative **definitions of "terrorism content"** strictly follow established rule of law and freedom of expression requirements. The original proposal's definitions are too wide in this respect and, if adopted, can be expected to become a significant source of abuse.
- Content removal laws like the TERREG proposal risk concentrating resources into an area with **limited tangible benefits**. The proposal is not sufficiently integrated in and connected to the broader legal and policy framework with respect to violent extremism and terrorism.
- The proposed content takedown order procedure should offer **independent judicial oversight** over public interferences with freedom of expression. In principle, such prior independent review should be a requirement for content takedown orders to be issued. For emergency situations, which should be adequately, explicitly and strictly defined, such review should still take place as a rule (and not be made dependent on an appeal), but could be started immediately after an emergency removal order is issued.
- The **one-hour removal** time frame is too rigid and simplistic. A more flexible requirement (promptly, without undue delay) would signal similar urgency, while better respecting established freedom of expression and due process values.
- The proposal's **referral procedure**, which requires platforms to handle law enforcement notifications under their terms of service standards, undermines due process as well as public legitimacy and accountability for limitations on freedom of expression.
- Although a full **separation of public and private regulation** may not be feasible, new rules on public enforcement actions with respect to online expression should follow and further develop established legal safeguards.
- The proposal enlists **platforms as de facto regulators** of online speech. This regulation through proxy challenges the legitimacy of subsequent restrictions on freedom of expression, complicates legal action on behalf of users' freedom of expression, and poses a central challenge to protecting free expression online.
- **Public accountability** and reporting on the use of proposed measures and procedures by competent authorities should be significantly enhanced.
- The **proactive monitoring provisions** violate the ban on preventive monitoring from Article 15 of the e-Commerce Directive (ECD), and lack clarity and supporting evidence for the effective and proportionate use of automation in tackling relevant content. They would create significant legal uncertainty and can be expected to cause large-scale removal of legal speech by relevant platforms.
- Automation in content moderation can be helpful in tackling issues at scale, but there are inherent risks and limitations as a result of the current state of the art of **artificial intelligence for content moderation** purposes and the lack of appropriate and functioning safeguards for users.

- The application of automation in online content moderation should not result in a shift from a presumption of legality of online information and ideas to a presumption of illegality. The principle of **freedom of expression by default** should be developed and implemented.
- **New institutions are needed** to support rule of law and fundamental rights safeguards in the development, application and regulation of new forms of AI in online content moderation.

The TERREG proposal and its freedom of expression implications

The core aim of the TERREG proposal is to tackle the availability of "terrorism content" online, thereby preventing potential radicalization and support for terrorism caused by the dissemination of such content.[6] The proposal does so by (1) providing a general definition of terrorism content at the EU level, (2) establishing two mechanisms for public authorities to obtain removal of relevant content by a broad class of service providers (orders and referrals) and (3) imposing new duties of care on relevant service providers to combat the availability of similar content through their services, including through proactive automated means. A key part of the proposal is that content removal orders would require an effective response in as little as one hour.

The proposal is the first legislative text in Europe, together with the new copyright proposal, to require proactive filtering of illegal content, breaking with the e-Commerce Directive approach to intermediary liability. The proposal builds on earlier policy documents released by the European Commission in the broader area of tackling illegal content online, and the co-regulatory initiatives of the EU Hate Speech Code of Conduct and the EU Internet Forum. As this document was being finalized, the European Parliament adopted the LIBE Committee report coming out of the EP committees. This sets the stage for the "trialogue negotiations" between the Council, the Parliament and the Commission to be initiated.[7] The LIBE Committee report and the Council position diverge significantly on many crucial aspects of the proposal from a freedom of expression perspective.

The implications for freedom of expression of the proposal are varied, significant and widely acknowledged.[8] The proposal itself contains a number of safeguards to address freedom of expression concerns. Most significantly, the proposal puts forward an obligation on service providers to allow users to complain if they believe their content has been removed unjustifiably. And the proposal requires human oversight and verification of automated tools for the removal of terrorism content to prevent unjustified removals.

These proposed safeguards notwithstanding, the draft regulation presents a clear threat to freedom of expression. First, the definitions of terrorism content lack the legal detail and precision that should be required for restrictions on freedom of expression. Second, the proposal targets a broad heterogeneous set of intermediary and online service providers and further enlists them into a project of privatized enforcement without proper human rights accountability. Third, the safeguards in the proposal fall short of European and international freedom of expression standards and best practices.

Definitions of targeted speech and communications

Key findings:

- Weak evidence for targeting speech defined as terrorism content in support of counter-radicalization;
- Definition of terrorism content does not include an intent requirement;
- Proposed definitions are broader than the criminal offences currently defined in EU law;
- Definition of "terrorism content" can easily include protected speech, while being subject to takedown orders and referrals;
- Definitions fail to meet the (freedom of expression) prescribed by law standard.

A first concern is that the proposal provides insufficient evidence demonstrating a causal connection between terrorist actions and "terrorism content" as defined in the proposal. While the covered content will generally be shocking and disturbing,[9] there is no clear evidence linking these particular kinds of content and terrorist radicalization or offenses. There is evidence that the internet allows terrorists to effectively disseminate their motivations for committing their crimes,[10] but evidence shows that radicalization may as well be caused by consumption of daily news (including coverage of terrorist acts). Available evidence also shows that radicalization tends to occur primarily as a result of offline rather than online dynamics.[11] This puts the proposal on a weak footing, including from a freedom of expression perspective.

If one accepts that new legal procedures are needed to tackle certain terrorism-related information and communications on the internet, the regulatory challenge is to define precisely which information should be allowed to be targeted by public authorities,[12] thereby satisfying European and international freedom of expression standards. The weak evidence for a causal link between terrorism offenses and terrorism content should have informed a narrow definition of which material could be targeted, namely material that causes an actual risk and/or imminent harm. Under the broad and seemingly simple notion of "terrorism content" in this proposal, however, lurks a wide variety of targeted terrorism-related offences and activities. Notably, the definitions in the proposal are broader than the speech-related terrorist offences currently provided for under EU law. The new definitions generally lack the precision required by tests under freedom of expression law and do not include an intent requirement.

The European Commission's proposed definition for terrorism content in Article 2(5) is the following:

> 'terrorist content' means meets one or more of the following information:
>
> (a) inciting or advocating, including by glorifying, the commission of terrorist offences, thereby causing a danger that such acts be committed;
>
> (b) encouraging the contribution to terrorist offences;
>
> (c) promoting the activities of a terrorist group, in particular by encouraging the participation in or support to a terrorist group within the meaning of Article 2(3) of Directive (EU) 2017/541;
>
> (d) instructing on methods or techniques for the purpose of committing terrorist offences.

The current Council version amends as follows (edits underscored):

(5) 'terrorist content' means material which may contribute to the commission of the intentional acts, as listed in Article 3(1)(a) to (i) of the Directive 2017/541, by:

(aa) threatening to commit a terrorist offence;

(a) inciting or advocating, such as the glorification of terrorist acts, the commission of terrorist offences, thereby causing a danger that such acts be committed;

(b) soliciting persons or a group of persons to commit or contribute to terrorist offences;

(c) promoting the activities of a terrorist group, in particular by soliciting persons or a group of persons to participate in or support the criminal activities of a terrorist group within the meaning of Article 2(3) of Directive (EU) 2017/541;

(d) instructing on methods or techniques for the purpose of committing terrorist offences

The lead European Parliament Committee's report (LIBE) offers the following definition (edits underscored):

(5) 'Terrorist content' means one or more of the following material:

(a) inciting the commission of one of the intentional offences listed in points (a) to (i) of Article 3(1) of Directive (EU) 2017/541, where such conduct, directly or indirectly, such as by the glorification of terrorist acts, advocates the commission of terrorist offences, thereby causing a danger that one or more such offences may be committed intentionally,

(b) soliciting another person or group of persons to commit or contribute to the commission of one of the offences listed in points (a) to (i) of Article 3(1), of Directive (EU) 2017/541, thereby causing a danger that one of more such offences may be committed intentionally;

(c) soliciting another person or group of persons to participate in the activities of a terrorist group**,** including by supplying information or material resources, or by funding its activities in any way within the meaning of Article 4 of Directive (EU) 2017/541, thereby causing a danger that one of more such offences may be committed intentionally;

(d) providing instruction on the making or use of explosives, firearms or other weapons or noxious or hazardous substances, or on other specific methods or techniques for the purpose of committing or contributing to the commission of one of the terrorist offences listed in points (a) to (i) of Article 3(1) of Directive (EU) 2017/541;

(e) depicting the commission of one or more of the offences listed in points (a) to (i) of Article 3 (1) of Directive (EU) 2017/541, and thereby causing a danger that one or more such offences may be committed intentionally.

For terrorist offences, EU law typically includes strict requirements of intent and likelihood of speech resulting in criminal action. But these are notably absent from the above definitions of "terrorism content" and "terrorism content dissemination" in the proposal.[13] By using relatively weak legal language, such as "causing a danger that" (EC, proposal) or "may contribute to" (Council and EP version), the definitions open up more space for restrictions than is necessary.

In other words, online media posts could be considered "terrorism content" even though they are clearly not intended to support or incite terrorism. To give an example: TERREG might encourage platforms to remove content by (citizen) journalists and opinion leaders who are responding to and quoting from terrorist propaganda. These parties may actually be rebutting terrorist causes, not supporting them, but since the Directive's definitions of "terrorist content" do not include intent, platforms may still find reason to remove this speech. It is worth noting, finally, that the amended definitions of the European Parliament's LIBE report and the Council do include some references to intent, although these are predominantly tautological references to the intent requirements in the underlying terrorist offenses instead of intent requirements related to the posting of content and the intended results thereof.

The proposal's recitals do clarify the scope of the definitions. For instance, recital 9 stipulates that "content disseminated for educational, journalistic, counter-narrative or research purposes should be adequately protected," but these safeguards are lacking in the actual legal text of the proposal.

In addition, the definition of terrorism content is not connected to existing speech-related offences in the area of terrorism, such as recruitment, training, and financing, which were already defined previously in Article 5-12 of Directive 2017/541. The proposal seems to have prioritized a broad and sweeping definition over a precise, legally sound set of definitions. More precise definitions would help to ensure that the Directive's most far-reaching provisions, such as the framework for proactive measures, are only imposed when absolutely necessary.

With the current definitions, the TERREG proposal provides for government-sanctioned removal mechanisms for speech that does not present an actual or imminent risk for terrorist offenses, including communications that are themselves not necessarily criminal under EU law. This raises pertinent questions about the precise legal basis for government-mandated removal of such material, and whether these mechanisms can be considered necessary and proportionate in the first place.

Regulation by proxy and privatized enforcement

Key findings:

- Proposal targets very broad range of online service providers;
- Proposal undermines the existing safe harbor regime in e-Commerce Directive;
- Proposal violates ban on general monitoring;
- Proposal codifies problematic practice of informal referrals by law enforcement and extralegal removal of information online.

The proposal targets a broad range of online service providers ("hosting service providers"), ranging from cloud infrastructure companies to online marketplaces, file storage services, social media and

search engines. The proposal uses the definition of hosting service providers from the e-Commerce Directive (2000/31/EC), which was introduced to limit the responsibility that could be imposed on intermediary services. It now connects to this definition to introduce new obligations to police and remove online speech.[14] Thus, the proposal narrowly focuses on the ability of online service providers to act as control points and censors of expression online, without taking account of the precise role different services play in the online environment and their relationship with expressive activities they help to facilitate. The scope of the proposal is one of the key areas of debate and amendment. Infrastructural service providers, like Amazon Web Services or Microsoft Azure, and other cloud infrastructure or service companies with more remote relations to the actual content are excluded from the regulation in the LIBE report.

Furthermore, the proposal undermines the existing safe harbor regime in the e-Commerce Directive, by creating a proactive duty of care for hosting service providers and moving beyond the reactive notice and takedown obligations that follow from the ECD framework. The ECD, adopted in 2000, is currently under pressure from different sides and risks being eroded completely through a combination of this proposal and others (audiovisual regulation, copyright enforcement). The most striking departure from the ECD is the introduction of legal obligations to prevent known "terrorism content" from becoming available (upload filtering) and more general preventive duties to remove terrorism content through automated content recognition tools. These provisions violate the ban on general monitoring in Article 15 ECD, which the Court of Justice of the European Union has found to support the freedom of expression rights of internet users (e.g., the *Scarlet Extended SA v. Sabam* case).

The proposal codifies the already existing problematic practice of informal referrals by law enforcement and the subsequent extralegal removal of information online on the basis of a company's terms of service.[15] The proposed referral mechanism for online content does not entail a determination by an appropriate authority that the content falls within the definition of terrorism content and whether the content is actually illegal. Instead, when receiving referrals, a company must decide upon content removal on the basis of its terms of service, which tend to be much broader and more flexibly enforced than requirements under criminal procedural law. As a result, the proposal undermines existing legal procedures and due process safeguards for internet users. The proposal obliges services to operate a complaint procedure for internet users whose content is removed, but does not create effective avenues to appeal referrals at the source, i.e., the public authority that has made the referral. Finally, the broader law enforcement referral practices anticipated by the proposal could be amplified through existing industry coordination in the GIFCT hash-sharing database initiative, if law enforcement referrals become a significant source for this industry database.[16]

Deficient freedom of expression safeguards

Key findings:

- Inflexible one-hour response deadline for content takedown orders;
- Takedown orders lack independent judicial review;
- Safeguards for affected speakers/audiences;

- Risks related to deficiencies and bias in automated content recognition tools.

The proposal imposes an inflexible one-hour response deadline for content takedown orders. The stated reason for this short response window is that the most intense dissemination of terrorism content tends to take place in the first hours after its posting. This raises a number of concerns. First, if this is the case, how would a one-hour response window help to address this? Typically, it will take time for content that is posted online to be identified as "terrorism content" by relevant authorities.[17] On top of the time that it will take to process the takedown order, it seems unlikely that dissemination in the first hours can be effectively addressed.[18] Second, smaller service providers will likely lack the resources to provide for effective 24-hour staffing to be able to comply with this obligation. Third, the short window to process orders will incentivize service providers to minimize review of such orders. Considering the lack of judicial review on content takedown orders before they are sent to service providers, this presents a big risk for freedom of expression. Overall, a more flexible obligation to act expeditiously, without undue delay, would better support the necessary and proportionate requirement for interferences with online speech.

Another safeguard that is lacking is judicial review on content takedown orders before they are sent to service providers, or as soon as they are sent in the case of emergency situations in which a prior review would cause undue delay.[19] The proposal refers to appeal mechanisms for service providers that it expects to be in place in the Member State but the proposal does not set minimum standards for these appeal mechanisms.[20] This creates a significant risk of abuse. First, it creates the possibility for non-judicial authorities to use the procedure to censor content without due process. This is particularly problematic for European countries with broader rule of law issues.

Existing appeal mechanisms in the Member States are likely to take far longer than the stipulated one-hour response window.[21] The proposal's broad scope in terms of service providers, which would include collaborative journalism platforms and others for public debate, poses real dangers for robust debate on terrorism-related matters of public concern.[22] The lack of judicial review and effective appeal mechanisms for takedown orders is one of the aspects of the proposal that most clearly violates established freedom of expression case law.

The proposal adopts a narrow view of whose freedom of expression rights require protection. It does not recognize that service providers can invoke freedom of expression when confronted with takedown orders and referrals. It also doesn't recognize that there are others besides the users posting the content whose freedom of expression can be curtailed by the takedowns. First, those other internet users who would have wanted to access the content are now prevented from seeing it. Second, the authors of content that is taken down are not always the same as the individuals uploading the content and may see their expression taken down without an effective remedy. For this reason, it could be worth broadening standing in relevant appeal procedures beyond the user (re-)posting particular content to others unduly impacted in their freedom of expression.

Finally, the proposal clearly relies upon the efficacy of automation to identify and take down illegal content, and does so without weighing the evidence and research on these tools. The evidence on automation shows significant problems of false positives (and negatives), and bias with respect to the expression of different viewpoints and groups. The proposal stipulates that any proactive measures

"shall provide effective and appropriate safeguards to ensure that decisions taken concerning that content, in particular decisions to remove or disable content considered to be terrorist content, are accurate and well-founded," but doesn't say what that means. It merely states that safeguards should entail "human oversight and verifications where appropriate and, in any event, where a detailed assessment of the relevant context is required in order to determine whether or not the content is to be considered terrorist content." The proposals for automation appear to connect to the existing industry initiative of a shared database of hashes for violent extremist content that is used to proactively remove such content from their services.

Conclusion

The use of the internet for recruitment and the dissemination of violent extremist materials raises significant policy challenges for public authorities and internet services alike. Freedom of expression has an important role to play in shaping regulation and industry policies. It is clear from the above that the TERREG proposal creates substantial risks with respect to freedom of expression that should be addressed before its adoption.

Official documents

- European Commission proposal
- European Union's Legislative Observatory link with relevant links to the proposal and official texts adopted in the Parliament
- Latest public draft text of the Council, December 2018

Notes

[1] Prof. dr. Joris V. J. van Hoboken, Professor of Law at the Vrije Universiteit Brussels (VUB) and a Senior Researcher at the Institute for Information Law (IViR), University of Amsterdam. At VUB, I am appointed to the Chair "Fundamental Rights and the Digital Transformation," which is established at the Interdisciplinary Research Group on Law Science Technology & Society (LSTS), with the support of Microsoft.

[2] It's worth noting at the outset that there is no uniform definition of the term terrorism, and what is understood as terrorism has also changed significantly over time. This lack of clarity can be a cause for too broad application and misuse of relevant legislation adopted at the national level.

[3] For a detailed overview see Wikipedia, Christchurch mosque shootings: https://en.wikipedia.org/wiki/Christchurch_mosque_shootings#Video_distribution.

[4] The European Parliament's committee rapporteurs for TERREG are: Daniel Dalton, UK (LIBE Committee, lead); Julia Reda, DE (IMCO Committee); Julie Ward, UK (CULT Rapporteur). The Council's presidency is currently held by Romania.

[5] In addition, the hash-sharing database industry initiative (GIFCT) was discussed.

[6] For an overview of counter radicalization strategies in media and communications, see Ferguson, Countering violent extremism through media and communication strategies: A review of the evidence, 2016 available at http://www.paccsresearch.org.uk/wp-content/uploads/2016/03/Countering-Violent-Extremism-Through-Media-and-Communication-Strategies-.pdf. See also Daphne Keller, 'Internet Platforms: Observations on Speech, Danger and Money, A Hoover Institution Essay, 2018. Available at https://www.hoover.org/sites/default/files/research/docs/keller_webreadypdf_final.pdf. Keller stresses how little is

known about the impact of content removal practices on people at risk of radicalization and the dangers of well-intentioned campaigns against violent extremist content backfiring.

[7] Trialogues are expected to begin after the summer.

[8] See, e.g., Aleksandra Kuczerawy, "The Proposed Regulation on Preventing the Dissemination of Terrorist Content Online: Safeguards and Risks for Freedom of Expression," 2018. Available at http://dx.doi.org/10.2139/ssrn.3296864; Letter of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, the Special Rapporteur on the right to privacy and the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, 7 December 2018, available at https://spcommreports.ohchr.org/TMResultsBase/DownLoadPublicCommunicationFile?gId=24234; Faiza Patel, "EU 'Terrorist Content' Proposal Sets Dire Example for Free Speech Online," Just Security, 5 March, 2019, available at https://www.justsecurity.org/62857/eu-terrorist-content-proposal-sets-dire-free-speech-online/. EDRi, FRA and EDPS: Terrorist Content Regulation requires improvement for fundamental rights, 20 February 2019, https://edri.org/fra-edps-terrorist-content-regulation-fundamental-rights-terreg/.

[9] The fact that content is shocking and disturbing doesn't mean it won't be protected speech in Europe (*Handyside v. UK*).

[10] Most recently, for instance, CNN reports, that the San Diego synagogue shooter posted a letter on 8chan ("The letter writer talks about planning the attack and references other attacks on houses of worship, including the attack on the Tree of Life Synagogue in Pittsburgh and the Christchurch mosque shootings in New Zealand"). See Ray Sanchez and Artemis Moshtaghian, "Mayor says synagogue shooting in California that left 1 dead and 3 wounded was a 'hate crime,'" CNN, 28 Paril 2019, available at https://www.cnn.com/2019/04/27/us/san-diego-synagogue/index.html. Leading voices have called for the blocking of 8chan by dominant internet companies.

[11] See Ferguson 2016.

[12] The question of what is the right definition for internet companies to use when targeting terrorism and violent extremism is a different one. For a detailed discussion of relevant considerations for industry policies in this area, see Brian Fishman, "Crossroads: Counter-terrorism and the Internet," Texas National Security Review, Vol 2, Issue 2 (April 2019), available at https://tnsr.org/2019/04/crossroads-counter-terrorism-and-the-internet/ (arguing that public authorities may only see a tip of the iceberg of what a company like Facebook is doing with respect to terrorist content). Fishman leads efforts against terrorist and hate organizations at Facebook.

[13] The terrorist offences provided for at the EU level in Directive 2017/541 generally require that these acts, "given their nature or context, may seriously damage a country or an international organization." In addition, they are only defined as "terrorist offences" where committed with the aim of "seriously intimidating a population," "unduly compelling a government or an international organisation to perform or abstain from performing any act" and/or "seriously destabilising or destroying the fundamental political, constitutional, economic or social structures of a country or an international organization."

[14] The scope of the hosting service provider definition is legally contested and hotly debated. For a discussion, see Van Hoboken et al., Hosting Intermediary Services and Illegal Content Online, Study for the European Commission, (forthcoming).

[15] For a discussion of referrals from a human rights perspective, see Jason Pielemeier and Chris Sheehy, "Understanding the Human Rights Risks Associated with Internet Referral Units," The GNI Blog, 25 February 2019, available at https://medium.com/global-network-initiative-collection/understanding-the-human-rights-risks-associated-with-internet-referal-units-by-jason-pielemeier-b0b3feeb95c9.

[16] Due to the lack of transparency about the details of the GIFCT initiative, the question whether this could be the case is difficult to answer.

[17] It would have strengthened the proposal if better data was (made) available on the time it takes relevant authorities to become aware of relevant material online and in what ways that timeframe could be minimized effectively.

[18] The dynamics around the Christchurch mosque shooting show how difficult it is to effectively stop dissemination, as groups mobilize to circumvent removal mechanisms deployed by internet companies. See, e.g., Kate Klonick, "Inside the Team at Facebook That Dealt with the Christchurch Shooting," The New Yorker, 25 April 2019, available at https://www.newyorker.com/news/news-desk/inside-the-team-at-facebook-that-dealt-with-the-christchurch-shooting. See also Brian Fishman, "Crossroads: Counter-terrorism and the Internet," Texas National Security Review, Vol 2, Issue 2 (April 2019), available at https://tnsr.org/2019/04/crossroads-counter-terrorism-and-the-internet/.

[19] The proposal currently does not stipulate an emergency procedure, but seems to build on the assumption that for terrorism content as defined such an emergency always exists. Considering the issues with the definitions discussed earlier, it's not clear that this assumption is correct.

[20] The proposal does contain a mechanism for service providers to ask for clarification in case of missing information in the order or manifest errors.

[21] The proposal lacks documentation and analysis of existing laws and procedures in the Member States.

[22] The EC proposal only contains a (weak) reference to the need to protect journalistic coverage of terrorism content in the preamble (recital 9). In support of journalism and freedom of expression, the Council position adds some more detail to this recital and adds a provision that the "Regulation shall not have the effect of modifying the obligation to respect fundamental rights and fundamental legal principles as enshrined in Article 6 of the Treaty on the European Union" (Article 1(3)). This provision is purely declaratory: the Regulation has to comply with the Treaty of the EU and the Charter of Fundamental Rights regardless of this text. The EP has the clearest exception for journalism in its 1st reading of the proposal, with a new provision in Article 1, paragraph 2 a stating that the "Regulation shall not apply to content which is disseminated for educational, artistic, journalistic or research purposes, or for awareness raising purposes against terrorist activity, nor to content which represents an expression of polemic or controversial views in the course of public debate." All of these provisions leave important questions as to what will be effectively covered by these exceptions open, with a particular danger that the concept of journalism and who can claim to be engaged in journalistic activities, will be narrowly construed.

TRANSATLANTIC WORKING GROUP

Combating Terrorist-Related Content Through AI and Information Sharing†

Brittan Heller, The Carr Center for Human Rights Policy, Harvard University[1]

April 26, 2019

Contents

Introduction

In its first meeting at Ditchley Park, UK, in March, the Transatlantic Working Group (TWG) focused on the tech industry's efforts to address hate speech and extremist content while still respecting freedom of expression. One effort we examined was a mechanism to combat online extremism through private-to-private information sharing efforts, the Global Internet Forum to Counter Terrorism (GIFCT) and its industry-only hash-sharing database. The following analysis served as a basis for our discussions, and now has been revised to incorporate crucial insights from those discussions. It explains the background behind the GIFCT database, which often operates with secrecy, and analyses its implications for freedom of expression. This document also highlights important considerations for industry and policy makers about applying private industry-based information sharing as a method to address controversial, dangerous, or illegal online content.

---

† 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/.

What is the GIFCT Database?

In December 2016, Google, Facebook, Twitter and Microsoft announced an industry-led initiative to disrupt the terrorist exploitation of their services. This led to the June 2017 formation of the GIFCT. Its objective is to "substantially disrupt terrorists' ability to promote terrorism, disseminate violent extremist propaganda, and exploit or glorify real-world acts of violence."[2] Functionally, the GIFCT is a private enterprise to address a public harm; it is run by tech companies for the mutual benefit of tech companies.

Not much is publicly known about how the GIFCT operates. This paper inquires how private information sharing is structured and how it is technically implemented. The GIFCT's work is organized into three pillars: joint tech innovation, research, and knowledge sharing. Information-sharing efforts are housed under the GIFCT's joint tech innovation pillar, to focus on building shared technology for use within the tech industry to prevent and disrupt the spread of terrorist content online. These efforts have resulted in a common industry database of "hashes" — unique digital fingerprints — for violent terrorist imagery or terrorist recruitment videos or images that the member companies have removed from their services. When pro-terrorist content is identified and removed by one GIFCT member, the content's hash is shared with the other participating companies to enable them to identify and block the content on their own platforms.

The GIFCT's definition of terrorism is not fixed and is drawn from guidance given by the United Nations. Each member company defines and captures what qualifies as "terrorism related content" under its own terms of service. According to GIFCT's nonprofit partner, Tech Against Terrorism, the approach of defining terrorism is challenging:

> [Tech Against Terrorism] acknowledges that there is no universal definition of terrorism. In fact, one of our observations when engaging with tech companies is that they struggle with moderating content on their sites due to this uncertainty. Moreover, it is sometimes difficult to define whether a video is part of terrorist propaganda, or whether it is an important piece of news that sheds light on human right abuses. When tech companies fail to make this distinction they are often criticized, but the fact is that there is no regulating body providing clear guidelines to companies whose platforms and audiences span the entire world. Tech Against Terrorism advocates for more coherence on this matter, and therefore suggests a global normative approach rather than an ad hoc approach from single governments. We recommend companies to consult the Consolidated United Nations Security Council Sanctions list, as it provides the best framework to the international consensus on individuals and groups defined as a terrorist. Having said that, we note the absence of certain groups in that list and particularly far-right terror groups. Therefore, companies should also consult the proscribed groups and individuals' list in the specific region and/or country where the content is flagged.[3]

According to representatives affiliated with GIFCT who were interviewed for this research, public authorities do not directly interface with the shared industry hash database by design. Administratively,

the database is maintained and run by one of the four main GIFCT member companies. Governments or intergovernmental organizations reportedly do not have access to the roster of what content is indexed in the database. Efforts to enforce laws related to terrorism-related content would go through individual member companies' legal teams, but this would involve directly requesting information related to individual pieces of content through preexisting legal processes — without regard to whether or not the content is listed in the shared database.

While law enforcement or governments can theoretically come to companies with content they claim is terrorism-related, this would not necessarily mean it would be indexed by the GIFCT. The hash-sharing consortium member companies individually designate the particular content that should be flagged, tagged, or removed in accordance with their terms of service, and not against legal constraints.

With this background in mind, policy makers concerned with transparency may have concerns about the structure of the GIFCT and its relationship to public authorities. Industry claims that private information sharing has increased their capacity to respond to terrorism-related content quickly because they do not need to duplicate efforts to identify the "worst of the worst" type of information, like terrorist recruitment videos or images of graphic violence like beheadings. However, there is no external auditing of the database. Hash sharing is a closed effort, occurring outside public oversight. This raises concerns for the freedom-of-expression rights of individuals whose content mistakenly may have been flagged or accounts erroneously removed. Further, in the GIFCT database context, there is no right to any appeal for content removals. Much of the flagged content disappears from the platforms before it is even posted, making it challenging to even know if content has been removed in error. Without access to this information, there are concerns about accountability for tech companies who may be overzealous in enforcement.

Privatized efforts to deal with content that is likely to be illegal also implicate related concerns about interfacing with public authorities. Policy makers can look to a similar industry-based hash-sharing effort for child sexual exploitation issues, run by the National Center for Missing and Exploited Children (NCMEC), which was touted as having similar types of benefits for combating illegal child pornography. However, courts in the United States are still debating whether or not this structure – where private actors shared content marked as child pornography with public authorities – created a special relationship with law enforcement. If so, should that collaboration result in NCMEC qualifying as a "government entity or agent" and thus warrant Constitutional protections for the accused?[4] While courts have not yet decided, the issue of mixing private and public online enforcement mechanisms still raises concerns for digital rights. In the current regulatory landscape, where policy makers seek to designate types of controversial or harmful content as illegal, the GIFCT should provide a model to examine the limitations and challenges of looking to private companies to perform some functions that are traditionally the purview of law enforcement.

Is the GIFCT Database Effective?

The tech industry has treated this collaboration as a success that results in greater efficiency in online policy enforcement and decreases in online terrorist content. As of 2018, there were 13 hash-sharing

consortium members and the database contained more than 100,000 hashes.[5] Another 70 companies reportedly discussed joining in 2018.[6] Consortium member companies used the GIFCT hashes to identify and remove matching content – videos and images – that violated their respective policies or, in some cases, blocked terrorist content from being posted.[7]

Early data indicate that GIFCT's hash-sharing efforts are working, if content removal is the metric of success. Statistics presented by the GIFCT seem to reinforce its claim that the database is focused on a small but significant slice of content, accounting for the worst type of terrorist-related content. Between 2015 and 2017, Twitter reports having suspended over 1.2 million terrorist accounts.[8] In the second half of 2017, YouTube removed 150,000 videos for violent extremism, and over 10,000 in Q3 2018.[9] Nearly half of these were removed within two hours of upload.

Technical limitations and threats to free speech

But policy makers should recognize that challenges still remain as to the other implications of these efforts. First, there is a staggering amount of online content. Over five billion people are online. As of 2018, every single minute at least 510,000 comments and 136,000 photos were shared on Facebook, 350,000 tweets posted on Twitter, and 500 hours of video uploaded to YouTube.[10] This has increased exponentially and will only continue to grow. Given this immense number of postings, policy makers should understand that the hash-sharing database only affects a sliver of information available on the internet, so statements about impact should be contextualized.

Additionally, the GIFCT's standards do not address freedom of expression-related concerns, especially considering the problem of potential overreach when combined with technical limitations. Given the quantity of information, companies extensively rely on AI to manage identification and removal efforts. Facebook uses image matching to prevent users from uploading a photo or video that matches another photo or video that has previously been identified as terrorist-related content. YouTube has reported that 98% of the videos that it removes for violent extremism are flagged by machine-learning algorithms.[11]

However, it is important that policy makers understand that machine-learning algorithms cannot be expected to identify terrorist content with 100% accuracy. It is difficult to know how accurate these methods are in practice, since so little information is published about them. But even a 99.5% accuracy rate would create false positives affecting millions of people. Some content will be wrongly identified as "terrorist" and blocked or removed. One apparent victim of overreach is the Syrian Archive, a nonprofit aimed at documenting evidence of war crimes in the Syrian conflict. Reports from June 2018 show that its content was repeatedly removed from YouTube, leading to widespread and sometimes permanent losses of what might be crucial evidence of war crimes. According to Wired magazine, Google has taken down 123,229 of the 1,177,394 videos that Syrian Archive backed up in 2012-2018.[12] "I think we have already lost a lot of content since YouTube started using machine learning in 2017," said Hadi Al-Khatib, founder of the Syrian Archive. "There will be a big impact on the Syrian accountability process if we aren't able to retrieve it."[13]

Circumvention and subversion by extremist groups

Research about the scope and locations of online extremists' content should also factor into policy makers' efforts to evaluate the GIFCT. Consider the "whack-a-mole" dilemma, wherein companies participating in these efforts may not be the places where terrorists convene online. As a result of scrutiny from tech giants, many online extremists have migrated to the so-called dark web, to alternative gaming-based platforms like Discord, or to fringe platforms like 4chan, 8chan, and Gab for their use as communications channels.[14] These forums have been reported to have few if any restrictions on hate speech, disinformation, and other types of conduct that have led to offline violence. Online extremists have also found enforcement on major platforms to be irregular, with some platforms being more permissive than others. Also, as mentioned in the GIFCT's statement on the definition of terrorism, the database does not always capture more contextual, country-specific threat patterns and risks. In other words, there may be a distortion in the categorization of identified content – mostly ISIS-related, and less domestic extremist or white supremacist-related content – which does not match the risk profile when policy makers consider online radicalization in their countries.

In addition, policy makers need to acknowledge the technological sophistication of some extremists. Groups leveraging online content to commit offline harms are frequently early adopters of tactics to circumvent technologically oriented limitations. In response to disruption by Twitter, supporters of ISIS have tried to circumvent content blocking technology by "outlinking," spreading content through using links to other platforms. Sites often outlinked include justpaste.it (a new member of the GIFCT), sendvid.com, and archive.org. This appears to be a deliberate strategy to exploit the limited resources and expertise of smaller companies.

Case study: Christchurch shooting incident

The technical limitations of hash-sharing technology were clearly demonstrated during recent extremist violence, accompanied by a media proliferation strategy. In the wake of the March 2019 televised attack on a mosque in Christchurch, New Zealand, the gunman's live video of the shooting circulated around the world. According to YouTube, "The volume of related videos uploaded to YouTube in the first 24 hours was unprecedented both in scale and speed, at times as fast as a new upload every second."[15] Facebook summarized the scope and scale of the attack and its online coverage:

> The video was viewed fewer than 200 times during the live broadcast. No users reported the video during the live broadcast. Including the views during the live broadcast, the video was viewed about 4,000 times in total before being removed from Facebook. Before we were alerted to the video, a user on 8chan posted a link to a copy of the video on a file-sharing site. The first user report on the original video came in 29 minutes after the video started, and 12 minutes after the live broadcast ended. In the first 24 hours, we removed more than 1.2 million videos of the attack at upload, which were therefore prevented from being seen

> on our services. Approximately 300,000 additional copies were removed after they were posted.[16]

Hash-sharing efforts failed in this instance for several reasons. Initial images did not match closely enough to any images already in the database. The shooter's first-person perspective captured clean shots to his victims. These images would not be bloody, and gore is what AI filters are often trained to identify in looking for the worst type of content. There was not enough similar preexisting content in the database to allow the machine learning to match mass shooting-related content.

Additionally, the platforms presume cooperative users who will proactively flag the worst extremist content, which is then queued to be hashed and thus prevent subsequent downloads. With a sympathetic audience waiting to spread the content, the Christchurch video was not reported until almost a half-hour after the live video began.

The Christchurch video showed the limitations of hash-sharing efforts, given the problem of virality. Mass coordination by a group of bad actors aimed to distribute copies of the video to as many people as possible through social networks, video-sharing sites, and file-sharing sites. These individuals collaborated to continually edit, upload, and create new versions of the video. The multiple versions were designed to thwart hash-sharing efforts and stymie filters looking for original versions of the content. A day after the shooting, Facebook had over 800 slightly modified duplicates of the video in its hash-sharing database.[17]

Adding to the problem was a wider population who distributed the video and unwittingly made it harder to match copies. Facebook described how "[s]ome people may have seen the video on a computer or TV, filmed that with a phone and sent it to a friend. Still others may have watched the video on their computer, recorded their screen and passed that on. Websites and pages, eager to get attention from people seeking out the video, re-cut and re-recorded the video into various formats."[18] Legitimate news outlets shared the video as well, both online and in broadcasts.[19] The variety of formats undermined existing hashing technology, and did so in a way that may have exemplified the tension inherent in examining freedom of expression and hash-sharing technology.

Conclusion

From our review, it is clear that private information sharing plays a constructive role as one tool in the toolbox for combating violent terrorist content online. It is not a panacea. But in its present form, private information sharing could be improved in order to provide better protections for freedom of expression.

For private information-sharing efforts like the GIFCT to be grounded in freedom of expression, the tech industry should adopt the following safeguards:

- Prioritize transparency;
- Develop mechanisms for increased accountability for its work, including civil-society oversight for any information-sharing models that implicate digital rights;

- Implement a right to appeal for errant content removal.

Policy makers, for their part, should take the following into consideration:

- Consider the scope and scale of internet content and extremist online activity to evaluate the full impact of the GIFCT database;
- Be aware that interactions of public authorities with private information-sharing efforts may implicate not only freedom of expression, but also other fundamental rights;
- Be aware of the technical limitations of this technology – as exemplified by deletions of online evidence by YouTube and deliberate exploitation by extremist groups of social media's hash-sharing systems;
- In particular, understand vulnerabilities emerging from hash sharing's reliance on artificial intelligence and assumptions embedded in its user design-based interfaces.

It remains to be seen what part the GIFCT collaboration will play in viable solutions to online extremism, and if it adequately protects users' ability to express themselves freely and safely on online platforms. At best, policy makers should consider it to be only a part of a multifaceted solution, given the concerns related to freedom of expression, the technical limitations of hash sharing, and evolving techniques by extremist groups to subvert indexing efforts like the GIFCT. There may be promising alternative uses of the database, like sharing images identified with viral deception. Of course, these will have the same limitations as terrorism-related content. Critically evaluating the GIFCT through the lens of freedom of expression will help policy makers and governments to protect the fundamental rights of their citizens.

Notes

[1] Brittan Heller (https://carrcenter.hks.harvard.edu/people/brittan-heller) is a technology and human rights fellow at the Carr Center for Human Rights Policy at the Harvard Kennedy School. She works at the intersection of technology, human rights, and the law, and is an expert on hate speech and the movement from online conduct to offline violence. She also is a senior associate at the CSIS Business and Human Rights Initiative. Research assistance on this paper was contributed by Paddy Leerssen, a doctoral candidate at the Institute for Information Law (IViR) at the University of Amsterdam and a non-resident fellow at the Stanford Center for Internet & Society.

[2] https://www.gifct.org/about/

[3] https://www.techagainstterrorism.org/about/faq/

[4] United States v. Ackerman, No. 14-3265 (10th Cir. 2016).

[5] Ask.fm, Cloudinary, Facebook, Google, Instagram, Justpaste.it, LinkedIn, Microsoft, Oath, Reddit, Snap, Twitter and Yellow. Source: https://www.gifct.org/partners/

[6] https://www.bbc.com/news/technology-44408463

[7] The GIFCT provides more than just hash sharing to combat terrorism-related content. Under its "knowledge sharing" initiatives, small- and medium-sized tech companies are trained on strategies to address online extremism via partnerships with an affiliated non-profit, Tech Against Terrorism. It has worked with more than 100 tech companies on four continents. The GIFCT provides toolkits for startups to prevent terrorist exploitation of their platforms and services. It has also convened forums in Europe, the Asia Pacific region, and Silicon Valley for companies, civil society groups, and governments to share experiences and get suggestions for further efforts. These efforts are seen as complementary, but separate from the hash-sharing database.

[8] https://techcrunch.com/2018/04/05/twitter-transparency-report-12/
[9] https://transparencyreport.google.com/youtube-policy/featured-policies/violent-extremism?hl=en_GB
[10] https://www.dsayce.com/social-media/tweets-day/
https://www.omnicoreagency.com/youtube-statistics/
https://zephoria.com/top-15-valuable-facebook-statistics/
[11] https://www.gifct.org/about/
[12] https://www.wired.co.uk/article/chemical-weapons-in-syria-youtube-algorithm-delete-video
[13] *Id.*
[14]https://slate.com/technology/2018/10/discord-safe-space-white-supremacists.html
https://mashable.com/2017/08/17/alt-right-free-speech-online-network-gab/
[15] https://twitter.com/YouTubeInsider/status/1107645354361741312
[16] https://newsroom.fb.com/news/2019/03/update-on-new-zealand/
[17] https://newsroom.fb.com/news/2019/03/technical-update-on-new-zealand/
[18] *Id.*
[19] https://www.theatlantic.com/technology/archive/2019/03/facebook-youtube-new-zealand-tragedy-video/585418/

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 that "online hate speech should be addressed as part of the general reform of hate crime law."[27]

- 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.