How Platforms Respond to Human Rights Conflicts Online: Best Practices in Weighing Rights and Obligations in Hybrid Online Orders

Kettemann, Matthias C. (Ed.)

Erstveröffentlichung / Primary Publication
Sammelwerk / collection

MATTHIAS C. KETTEMANN (ED.)

How Platforms Respond to Human Rights Conflicts Online

Best Practices in Weighing Rights and Obligations in Hybrid Online Orders

EU COST Action – CA19143 – Global Digital Human Rights Network 2022

„All human beings are born free and equal in dignity and rights.“

Art. 1, sentence 1, Universal Declaration of Human Rights (1948)

EU COST Action – CA19143 – Global Digital Human Rights Network

How Platforms Respond to Human Rights Conflicts Online

Best Practices in Weighing Rights and Obligations in Hybrid Online Orders

edited by Matthias C. Kettemann
LEIBNIZ INSTITUTE FOR MEDIA RESEARCH | HANS-BREDOW-INSTITUT, HAMBURG, GERMANY
DEPARTMENT OF THEORY AND FUTURE OF LAW | UNIVERSITY OF INNSBRUCK, AUSTRIA

Cite as: Matthias C. Kettemann (ed.), How Platforms Respond to Human Rights Conflicts Online. Best Practices in Weighing Rights and Obligations in Hybrid Online Orders (Hamburg: Verlag Hans-Bredow-Institut, 2022). DOI: https://doi.org/10.21241/ssoar.81872

This is a publiation in the framework of the Global Digital Human Rights Network (GDHRNet). GDHRNet is funded as EU COST Action – CA19143 – by the European Union.

All outputs and working papers can be downloaded from leibniz-hbi.de/GDHRNet and GDHRNet.eu

Publisher: Leibniz Institut für Medienforschung | Hans-Bredow-Institut (HBI)
Rothenbaumchaussee 36, 20148 Hamburg
Tel. (+49 40) 45 02 17-0, info@leibniz-hbi.de, leibniz-hbi.de

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

TALLINN UNIVERSITY

LEIBNIZ INSTITUTE FOR MEDIA RESEARCH HANS-BREDOW-INSTITUT

universität innsbruck
Department of Legal Theory and Future of Law

Contributors

Name	Affiliation
Barthelmes, Mara	Student Assistant, Leibniz Institute for Media Research | Hans-Bredow-Institut, Hamburg
Barrat Esteve, Jordi	Professor of Constitutional Law, Rovira i Virgili University (Catalonia); Training Coordinator, EODS (Election Observation and Democracy Support)
Bubalo, Lana	Associate Professor of Law, University of Stavanger, Head of the Department of Accounting and Law
Böke, Julius	Student Assistant, Leibniz Institute for Media Research | Hans-Bredow-Institut, Hamburg
Costantini, Federico	Researcher and Lecturer of Legal informatics, Department of Law, University of Udine
Dinar, Christina	Junior Researcher, Leibniz-Institute for Media Research | Hans-Bredow-Institut
Fernández Aller, Celia	Professor of Law and Ethics. Technical University of Madrid; Advisory Board Fundación Alternativas. One of the experts in charge of Digital Rights Charter in Spain.
Fertmann, Martin	Junior Researcher, Leibniz Institute for Media Research | Hans-Bredow-Institut, Hamburg; PhD Fellow, Centre for Law in Digital Transformation, University of Hamburg
Fischer-Lessiak, Gregor	Researcher, lecturer and project manager at the European Training and Research Centre for Human Rights and Democracy, University of Graz, Austria
Gradulewski, Max	Student Assistant, Leibniz Institute for Media Research | Hans-Bredow-Institut, Hamburg
Hinrichs, Lena Marie	Student Assistant, Leibniz Institute for Media Research | Hans-Bredow-Institut, Hamburg
Hofmann, Vincent	Junior Researcher, Leibniz Institute for Media Research | Hans-Bredow-Institut, Hamburg,; Junior Researcher, Humboldt Institute for Internet and Society, Berlin
Kalaja, Laurena	Legal Representative, Lecturer in Law, POLIS University, Tirana
Kettemann, Matthias C.	Professor of Innovation, Theory and Philosophy of Law; Head of the Department of Theory and Future of Law, University of Innsbruck; Research Program Head, Leibniz Institute for Media Research | Hans-Bredow-Institut, Hamubrg
Koerrenz, Nicolas	Student Assistant, Leibniz Institute for Media Research | Hans-Bredow-Institut, Hamburg
Millner, Clara	Jurist, Antidiscrimination Office Styria (ADS), Graz
Neuvonen, Riku	Senior Lecturer in Public Law, University of Helsinki; Senior Researcher, Tampere University, Communications Rights in the Age of Digital Disruption (CORDI) research consortium funded by the Academy of Finland
Onesti, Alan	Ph.D. Student, Department of Law, University of Udine
Peña-Acuña, Beatriz	Vice-Dean of Quality, Practices, Students and Employment, Associate Professor, University of Huelva; Member International Academy of Social Sciences
Sackl-Sharif, Susanne	Postdoc Researcher, University of Music and Performing Arts Graz; Lecturer in empirical research methods, University of Graz

Schleif, Linda	Student Assistant, Leibniz Institute for Media Research | Hans-Bredow-Institut, Hamburg
Sekwenz, Marie-Therese	Researcher, Sustainable Computing Lab and Vienna University of Economics and Business
Simic, Jelena	Associate Professor, Union University School of Law, Belgrade
Topidi, Kyriaki	Senior Researcher, Head of Cluster on Culture and Diversity, European Centre for Minority Issues, Flensburg

Editorial support

Felicitas Rachinger, Department of Legal Theory and Future of Law, University of Innsbruck (team lead)

Johanna Erler, Department of Legal Theory and Future of Law, University of Innsbruck

Anna Schwärzler, Department of Legal Theory and Future of Law, University of Innsbruck

Linus Wörle, Department of Legal Theory and Future of Law, University of Innsbruck

Design

Larissa Wunderlich

Table of Contents

Preface

Platforms have power. But this power is not unchecked. Governments have an important role to play in protecting their citizens' rights vis-à-vis third parties and ensuring a communication order in which rights are not violated. (And in addition, of course, they need to respect human rights themselves and not arbitrarily shut down sites or use their power to make the Internet less free and open). As leader of working group 2 it is my distinct privilege to present this collection which unites studies by researchers within the Global Digital Human Rights Networks on issues connected to the overarching question of how platforms deal with human rights and their human rights obligations. This study is a key deliverable of our working group in the second year of the Global Digital Human Rights Network's activities. We will follow-up with Guidelines for platforms and an Assessment Model for states and other stakeholders in 2024. We developed this study under Corona conditions but were able to meet in the Tyrolean Alps in Obergurgl, Austria, in July 2022 to finalize this study.

Matthias C. Kettemann
LEIBNIZ INSTITUTE FOR MEDIA RESEARCH | HANS-BREDOW-INSTITUT, HAMBURG, GERMANY
HUMBOLDT INSTITUTE FOR INTERNET AND SOCIETY, BERLIN, GERMANY
DEPARTMENT OF THEORY AND FUTURE OF LAW | UNIVERSITY OF INNSBRUCK, AUSTRIA

Foreword

The Global Digital Human Rights Network is proud to submit this important study into the practices of online platforms when faced with human rights challenges. The overarching concern whether human rights can be safeguarded online as efficiently as offline is reflected in the topics of platforms power, hate speech and discrimination, limitations of private governance mechanisms, the governance of election-related information and the institutional responses to making private platform rules better. The broader philosophical assertion of the sameness of human rights online and offline is made manifest in specific issues related to the transposeability of offline rules and principles to the online environment.

The Network wishes to recognize the efforts of Professor Matthias C. Kettemann and his team, as well as all contributors for having undertaking this timely research and produceed valuable insights and conclusions. This comparative study aims to demystify how platforms deal with rights – and clarify whether they take rights seriously. It contributes to the mission of academia to engage with civil society, and political and corporate stakeholders in conceptualizing the challenges of human rights protection online. We expect that the study will not only provide a valuable contribution to human rights scholarship, but will influence more widely human rights discourse at various levels and in different regions.

Mart Susi
PROFESSOR OF HUMAN RIGHTS LAW, TALLINN UNIVERSITY
CHAIR OF GLOBAL DIGITAL HUMAN RIGHTS NETWORK

Executive Summary

Part I: Tools and Vectors of Platform Power

The Power of App Stores and Their Normative Orders

VINCENT HOFMANN, CHRISTINA DINAR, MATTHIAS C. KETTEMANN, LENA HINRICHS, JULIUS BÖKE AND MAX GRADULEWSKI

App stores govern content one institutional level above social networks. They exercize power through largely vague and ambiguous terms and conditions, which leads to inconsistent app moderation practices. Regulation to date has not addressed this problem sufficiently. However, the rules in the upcoming Digital Services and Digital Markets will increase the obligations app stores have.

(Niche) Platforms as Experimental Spaces for Content Moderation - Mapping Small, Medium and Niche Platforms Online

CHRISTINA DINAR AND LENA HINRICHS

We map and define the term of 'niche platforms' and look into different aspects of the definition of small and medium platforms in law as well as sociologically and ethnographically. We question the current regulation by size of platforms and show, which other factors have a role in establishing a niche platform (such as thematic orientation of a platform). We plead for a more nuanced regulation of platforms.

Facebook and Artificial Intelligence. Good Practices Review

BEATRIZ PEÑA-ACUÑA AND CELIA FERNÁNDEZ ALLER

Humans are usually good at distinguishing which content can hurt sensibilities, but machines still have a lot of trouble differentiating between hate speech, race, sex, politics, etc. This is one of the great challenges of artificial intelligence. The problem of detecting such content and comments has not been solved adequately by the AI system Facebook has in use.

Part II: Hate Speech and Discrimination

Discrimination on Online Platforms: Legal Framework, Liability Regime and Best Practices

LAURENA KALAJA AND LANA BUBALO

Online discrimination may resemble traditional discrimination, but it can have more serious consequences, as the internet plays an essential role in our lives, shaping our view of the world, our opinions and our values. But who is responsible? Even though the safe harbor principle still applies in Europe and the USA, platforms are less than. They have the ability to shape the information published on the platform, and they profit financially form the interaction that users have with information present on their platforms. In addition, the design of platform can shape the form and substance of their users' content. By analysing existing regulation and community standards, we show which measures are best suited for preventing and redressing online discrimination.

provide sufficient guarantees for protection against hate speech. Rather, the application of existing legal frameworks needs to re refined.

Digital Rights of Platform Workers in Italian Jurisprudence

FEDERICO COSTANTINI AND ALAN ONESTI

The social relevance of so-called "platform workers" has surged dramatically during the pandemic. The contribution explores how the issues concerning "digital rights" have been addressed in the Italian legal system and suggests possible remedies to reduce the vulnerability of members of this new workforce.

Part IV: Platforms and Elections

The Legal Framework of Online Parliamentary Election Campaigning - An Overview of the Legal Obligations of Parties and Platforms

MATTHIAS C. KETTEMANN, VINCENT HOFMANN, MARA BARTHELMES, NICOLAS KOERRENZ, LENA MARIE HINRICHS AND LINDA SCHLEIF

The German legal system provides for an interplay of rights and obligations for both political parties, especially if they are members of the government, and platforms. For political parties, these are, in particular, constitutional principles and their formulation in simple law, while platforms have so far been primarily regulated by media law. The EU's regulatory projects, especially the DSA and DMA, supplement this catalogue with far-reaching obligations for platforms.

Part V: Improving Platform Rules

Platform-proofing Democracy - Social Media Councils as Tools to increase the Public Accountability of Online Platforms

MARTIN FERTMANN AND MATTHIAS C. KETTEMANN

New institutional configurations represent a good opportunity to increase the legitimacy of the power platforms wield over internet users and to advance the protection of individual rights against platform overreach; such institutional configurations can be conceived as expert-based or participatory "Social Media Councils", but more research needs to be done on different practical avenues of their implementation.

Internet and Self-Regulation: Media Councils as Models for Social Media Councils?

RIKU NEUVONEN

Social Media Councils are at the moment mostly theoretical innovations and waiting for pilot projects. Media Councils are established part of media regulation and therefore could provide role models as well as best practices to the social media councils. It is especially important to build trust between different stakeholders when new institutions are formed or otherwise institutions are in crisis mode at the beginning.

Best Practices

The Global Digital Human Rights Network has set out to identify – through sectoral, platform-specific and state-specific studies – key best practices of platforms and platform governance. These cover five key regulatory areas.

Make better rules: Rules have power and so does infrastructure

- Regulating infrastructure behind the platforms is a powerful normative vector.
- The regulation of infrastructure providers can lead to substantial collateral damage.
- App developers are particularly interested in clear communication with the app stores and concrete information about what content of the app led to the app store's actions.
- The User/Monthly active user (MAU) figure itself, if it is a fixed component of regulation, should be regularly reviewed and revised (if necessary, by a government-independent expert opinion) also to allow exclusions, economic development and growth potential.
- Cross-platforming, especially in the case of problematic content (e.g., illegal content, election-related content, etc.), should increasingly bring platforms into exchange with each other.
- Small and medium platforms should come into low-threshold contact with larger platforms, not only in "crisis situations".
- Content moderation needs to be understood more broadly and the differences artisanal, industrial and automated content moderation models need to be better understood.

Ensure rights: rights matter and discrimination needs to be eliminated

- Online discrimination can have grave consequences for public safety and social inclusion and should be expressly addressed in international, legal and national regulations, and these sources of law should be harmonized.
- States, tech companies and NGOs should work together on raising awareness of the problem of discrimination online, so people can recognize discriminatory practices and know their rights.
- More research about online discrimination is needed so this practice can be recognized and better addressed.
- Tech companies ought to share best practices in detecting and avoiding discriminatory practices.
- Tech companies ought to cooperate on developing the automated systems of content control instead of developing parallel systems, which would be more cost efficient and result in more harmonized systems.
- Filtering algorithms would require human review to prevent human rights violations and discrimination.
- The existing mechanisms for reputational and copyright protection such as notice and take down procedures and the right to be forgotten can analogously be applied in case of online discrimination.

Respect the rule of law: Platforms have to stick to rule of law standards

- There is a need for additional transparency measures for online platforms, including on the algorithms used. Platforms that feature user-generated content should offer users a clear explanation of their approach to evaluating and resolving reports of hateful and discriminatory content, highlighting their relevant terms of service.
- Greater ease for reporting cases of online discrimination (user-friendly mechanisms and procedures).
- Platforms should enforce sanctions of their terms of service in a consistent, timely and fair manner.
- Platforms should abide by duty of care, going beyond notice-and-takedown based legal models.
- Legislative framework for handling of requests to take down discriminatory content should be put in place.
- Procedural protections should be built into platforms notice-and-takedown systems.
- Rules should incentivize intermediaries and users to detect illegality, while minimizing the risks and the costs of errors and safeguarding a balance between the different human rights at stake.
- Tech companies need to ensure algorithm transparency and neutrality.
- A balance between citizens and tech companies must be struck in designing the liability rules.
- Setting up a detailed and harmonized European notice and take down procedure would provide more legal certainty.

Make platforms more democratic: To make platforms more accountable, deliberative elements can be introduced

- Minimising associated risks: If not carefully designed, social media councils and other institutional proposals for renegotiating the relationship between societies, states and platforms, may conceal actual power structures and fail to initiate real change, providing only a perceived legitimacy while stabilising a status quo many societies seem uncontempt with.
- Design Requirements for New Institutions: Against the background of these risks, new institutional solutions have to meet the highest standards of transparency not just regarding their activities, but also regarding the systemic improvements they initiate at scale, being equipped with appropriate rights to information and data access to investigate and report on these aspects.
- Both media councils and social media councils are dependent on various stakeholders (government(s), public, companies, professionals etc). Every stakeholder must trust that the council is working on their benefits and disagreements may be solved.
- Sanctions: Effective self-regulation organs must have competence to order sanctions. These sanctions do not need to be fines or measurable in money but sanctions must be credible and reason to change bad practices.
- To limit the unchecked exertion of power by large platforms over political discourses and individual expression, major platform companies should, in their private ordering systems and terms of service, refer at least partially to external standards which cannot be arbitrarily changed by these companies, and their processes should be equipped with institutional structures that negotiate the relationship between the two sets of rules.
- Online platforms should have independent bodies consisting of legal experts evaluating the reported cases of discrimination in order to achieve better balancing of rights.

- Platforms and domestic legal frameworks need to consider tackling responsibility for the consequences of regulatory choices: for online hate speech, focusing on impact in socio-legal terms means opting for substantive equality between groups
- Institutional measures (e.g. third-party reporting centres, state-non-state partnerships) can be helpful to reverse mistrust towards law enforcement authorities felt by members of religious/racial minorities and improve accessibility to the criminal legal system.
- It is important to encourage self-regulation of Internet portals that would make clear internal rules regarding the prohibition of hate speech in user-generated content.
- It is important to systematically improve preventive measures against hate speech, primarily in terms of educating citizens about the harmfulness of hate speech and its consequences.

Make platform governance innovative: Normative sandboxes can help platforms (and regulators) innovate

- To incorporate the protection of "Digital rights" directly into the algorithm governing the platforms, in order to provide built-in operating mechanisms of trade negotiation and mediation (eg. Art. 25 of GDPR). In this sense, international guidelines and collection of best practices could help.
- Collective bargaining agreements: To include a more binding and specific regulation using collective agreements between union workers and employers' associations, where possible according to the legal framework.
- Local arrangement and code of conduct: To support at a municipal or regional level, the adoption of local provisions or of voluntary codes that could improve awareness of social and digital rights among riders.
- Regulatory sandboxes and living labs: To establish provisional legal frameworks in order to experiment with new forms of regulations and models of interaction suitable to protect the "Digital Rights" of "platform workers, according to the concept of "regulatory sandbox" included in the EU proposal called "Artificial Intelligence Act" (articles 53 and 54).
- Institutional Experiments: To achieve a balance between platforms' and states' power over the behaviour on the internet requires a multiplicity of bold institutional experiments. The Meta (Facebook) Oversight Board is, despite some shortcomings, a noteworthy and already at least partially successful experiment in this regard, but should not be elevated to an archetype of a social media council or conceptually monopolise the space for – still needed – further institutional innovation.

Part I: Tools and Vectors of Platform Power

The Power of App Stores and Their Normative Orders

Vincent Hofmann, Christina Dinar, Matthias C. Kettemann, Lena Hinrichs, Julius Böke and Max Gradulewski
LEIBNIZ-INSTITUT FÜR MEDIENFORSCHUNG | HANS BEDROW INSTIUT

Introduction

Social networks like Facebook, TikTok or Twitter moderate the content that is published on their platforms. The networks delete content, block accounts (sometimes even those of government leaders) or tag content with notices that it could be false information. This practice takes place in the area of tension between the freedom of expression of those affected and the protection of the general public from dangerous or illegal content.

If one looks at the chain of process steps on the way from the development of a social network to the deletion of certain content, it becomes apparent how many other actors are involved in the dissemination and removal of content. These actors set law through the terms of use of their respective services, which is close to state law and the terms of the social networks in its importance for the exercise of freedoms. The app stores of Google and Apple, for example, hold such a powerful position. The Play Store (Google) and App Store (Apple) have a combined market share of 94% in sales and 91% in downloads of mobile apps.[1] This makes app developers dependent on the distribution through the app stores: If an app is not available here, it will hardly be able to reach users and thus sales on mobile devices.

Most recently, a case from Russia came to light when Google and Apple removed[2] a "Navalny App" from their stores that the Russian opposition had used to facilitate "tactical voting" and the identification of opposition candidates.

In the super election year 2021, more people than ever before are forming their opinions on the internet. Therefore, it is of particular importance for democratic processes to also analyse the companies behind the social networks and their content governance.[3] This study summarises the legal framework for content governance by app stores. In doing so, it looks at the particularities in the comparison of content governance by social networks and app stores. In particular, with regard to the EU Commission's drafts of the DSA and DMA, the question is answered as to whether national or European law can provide suitable answers to the problems of content governance through app stores.

Market power of the app stores

Google and Apple seem to be well aware of their great market power, which is reflected in their market behaviour: Both stores charge a 30% commission on all sales generated in their stores and on so-called in-

[1] Figures from 2013: maclife.de, https://www.maclife.de/iphone-ipod/software/app-stores-im-vergleich-google-fuehrt-bei-anzahl-der-downloads-apple-hingegen-b

[2] https://www.reuters.com/world/europe/google-apple-remove-navalny-app-stores-russian-elections-begin-2021-09-17/

[3] See in detail: Matthias C. Kettemann, Vincent Hofmann, Julius Böke, Max Gradulewski, Jan Reschke and Leif Thorian Schmied, Superwahljahr 2021: Gesellschaftlicher, Medialer und Rechtlicher Rahmen, https://leibniz-hbi.de/de/blog/superwahljahr-2021-gesellschaftlicher-medialer-und-rechtlicher-rahmen.

app purchases (paid transactions made in the app without going directly to the app store). Developers were also prohibited from pointing out cheaper options outside the App Stores. Apple even only allowed payments via ApplePay.

Resistance is rising against this economic behaviour. The game developer Epic Games (Fortnite) is suing against the business practices in the USA. In the ruling of 10.09.2021, against which Epic Games has appealed, Apple is not considered a monopolist on the app market. Therefore, the commission was not criticised by the court.[4] However, it was decided that Apple may no longer prohibit its developers from referring to cheaper offers outside the App Stores. South Korea recently passed a law that prohibits commissions of 30% as well as the prohibition to refer to cheaper offers.[5]

Content governance through app stores

The app stores do not only make economic demands on the apps they distribute. The app stores also have an impact on the developers in terms of content. For example, Google and Apple banned Parler, a popular network among right-wing extremists and supporters of the storming of the Capitol, from their stores because it did not sufficiently moderate the content published there.[6] Parler then tried to comply with the App Stores' guidelines and is now available again in the Apple App Store. The app is not yet available on Google.[7] The app "Unjected" also disappeared from the stores because of the accusation of not taking sufficient action against the spread of false information.[8] Unvaccinated people could exchange and date each other, which is why the app was nicknamed "Tinder for vaccination opponents". On Telegram, the app stores presumably even take action against individual channels. For example, the channels of the far-right conspiracy theorist Attila Hildmann were probably blocked in the Android and iOS versions of the app under pressure from the app stores.[9]

The app stores' approach to misinformation and hate speech bears the dangers that also lurk in the moderation of content by social networks: How is it decided, especially in the case of content that is not illegal but nevertheless "harmful", what is deleted and what is not? In other words, who decides what is harmful and how? Especially in connection with false information, the question arises: What is truth and who decides about it, and how can I, as a user, defend myself against decisions?

Legal framework

The deletion of an app interferes with the fundamental rights of the app operators and users. The operators are initially affected in their freedom of occupation under Article 12 (1) of the German Basic Law. Depending on the content of the blocked app, the operators' freedom of expression (Article 5 (1) sentence 1 of the German Basic Law) may also be violated if the operators themselves use their app to disseminate or form their own opinion. The users' freedom to express and form opinions is also violated when apps are

[4] usatoday.com, Epic Games appeals Apple court ruling over App Store.

[5] arstechnica, South Korea law forces Google and Apple to open up app store payments, https://arstechnica.com/gadgets/2021/08/south-korea-law-forces-google-and-apple-to-open-up-app-store-payments/.

[6] welt.de, Parler: Internet platform popular with right-wingers no longer accessible.

[7] nbcnews.com, Apple reinstates Parler app, stands by initial ban.

[8] rnd.de, Unjected: App stores ban dating app for vaccination opponents.

[9] faz.net, Access to Telegram channels of Attila Hildmann blocked.

used in this way. However, this must be decided for each app individually and depends on the functions the app offers and the extent to which it is important for the expression or formation of the users' opinions. The size of the platform behind the app or its specialisation in a particular topic plays a role. However, app stores as private companies are not directly bound by fundamental rights. These only have an indirect effect on the relationship between users and companies. This is because the measures taken by the app stores against individual apps are also an expression of the entrepreneurial freedom of the app stores, Art. 16 EU-GRCh. As private companies, they can basically sell or block whatever they want. However, in a case concerning Facebook, the Federal Supreme Court ruled that due to their strong influence, especially on freedom of expression, they had to take the fundamental rights of users into account to a particular extent.[10]

In contrast to the moderation of social networks, the app stores can technically only remove the entire app from their stores. Targeted intervention against individual illegal content is not possible, which also affects legal content and content against which the app stores' blocking should not be directed. However, the content published by users can remain available even after an app has been excluded, be it via web versions, the store of the respective competitor or, in the case of Android, via alternative app stores. "Only" the reach of content is reduced by the measures of the app stores. However, depending on the collapse in reach, this can be tantamount to deletion.

Deleting or blocking is not the only effective means used by the app stores. Since even a temporary removal from the stores can lead to massive losses in sales, developers try to prevent such measures through compliant behaviour.[11] This was demonstrated by Telegram, where individual channels in the apps distributed via the App Store and Playstore were blocked, presumably due to pressure from the app stores.

The app stores have laid down the basis for such measures in their rules for the use of the stores. Together with the norms of state order, these form the applicable set of rules for "content governance" by the app stores.

Private orders

Apple

In its "App Store Review Guidelines", Apple has defined in which cases apps are not allowed or are blocked. According to these guidelines, apps with user-generated content must have established a content governance system and be able to take measures against users such as blocking profiles or deleting content. Apple refers to "offensive content" and "abusive users", against which the app operators must take action. An automatic system must also be set up to check users' content for "objectionable material" and prevent it from being uploaded. The app must also offer a way for users to easily contact it.

Apple thus remains vague, in particular what "offensive content" is, is not specified. The guidelines become more detailed when it comes to banning specific content. These rules apply regardless of whether the content is distributed by users or the developers of the app. "Apps should not include content that is offensive, insensitive, upsetting, intended to disgust, in exceptionally poor taste, or just plain creepy." The

[10] BGH, judgements of 29 July 2021 - III ZR 179/20 and III ZR 192/20.

[11] Previously unpublished interview by Christina Dinar (HBI) with Sven Voges (PlanetRomeo) from 10.08.2021, "Every time a block means that you are not found over the certain period of time, ... in some cases it costs 2 weeks or it took 2 weeks until we were back in and that means 2 weeks of sales that fall away".

last part of the "just plain creepy" provision in particular makes it clear how much leeway Apple gives itself in evaluating the apps. These prohibited contents are explained in the following paragraph. However, this is only done on the basis of examples in a non-exhaustive list.

The App Store Review Guidelines thus form only a little specific basis for the question of which content will not be found in the App Store or will be deleted from it again.

Google

Google's terms and conditions for the Play Store also prohibit some content for apps distributed through the Play Store.[12] The prohibited content is divided into the categories of bullying and harassment, sensitive events, violence, hate speech/ incitement of the people, as well as pornographic content and vulgar language. The various generic terms are also specified, along with some examples. There is also soft wording in Google's terms and conditions, which leave Google itself some room for manoeuvre in deciding. For example, "dangerous activities" may not be shown or favoured, and events such as natural disasters may not be treated with "insensitivity". Especially topics from the area of "sensitive content" can often be the subject of journalistic reporting or controversial debates, which could thus have a hard time in apps distributed via the Play Store.

Google's conditions on pornographic content are particularly noteworthy. This is only allowed "if it serves primarily educational, documentary, scientific or artistic purposes and is not superfluous." What content is superfluous and when is not defined and remains at Google's discretion.

Interim conclusion

The standards set by Apple and Google for the deletion of apps leave much room for interpretation. This is understandable in view of the enforcement of norms: the app stores use the space they have created themselves in enforcing the norms, whereby the broadest possible framework means more flexibility. And not only are the app stores' conditions opaque, the practice of "moderation" by the app stores also reveals little system.[13]

State order

Antitrust law

The already mentioned market share of almost 100% of Google and Apple in mobile apps also poses some risks in terms of antitrust law. The antitrust measures against the app stores relate, in the nature of antitrust law, to the economic consequences of the companies' dominance and do not directly address their "content governance". Thus, they cannot by themselves provide an answer to the content governance issues raised. Nevertheless, reducing the economic power and thus the dependence of developers on the app stores is a crucial building block for reducing the importance of moderation decisions by the app stores.

---

[12] Google Play, Developer Policy, Content Restrictions.

[13] VPN apps in Russia and China were examined: Ververis et al, Shedding Light on Mobile App Store Censorship, https://dl.acm.org/doi/10.1145/3314183.3324965.

In various countries, proceedings are underway against the app stores, in which they are accused of abusing their dominant market position.[14] The specific cause is, among other things, the commission model of the app stores. Apple and Google charge a flat rate of 30% for every paid transaction.[15] Apple also forbade its app developers to allow other payment methods apart from its own ApplePay service. The developers were also not allowed to point out that the app or services that could be purchased via in-app purchases could be purchased more cheaply outside the iOS infrastructure.[16] In addition to civil lawsuits, the competition authorities in both the USA and the EU took action against the app stores due to this market behaviour.[17] As mentioned above, in the lawsuit against Epic, a US federal court prohibited Apple from preventing developers from referring to cheaper distribution channels. And South Korea passed a law that, among other things, bans commissions of 30%.

German media law

In national media law, the Network Enforcement Act (NetzDG) and the State Media Treaty (MStV) provide the framework for content governance in internet media.

However, the NetzDG that applies to social networks does not apply to app stores. This means that the comprehensive obligations to delete illegal content do not apply to them, nor does the obligation to publish a report on the deletions and profile blocks that have taken place.[18]

The regulations for media intermediaries contained in the State Media Treaty, on the other hand, also apply to app stores. This obliges them to be transparent and non-discriminatory. What initially sounds like a solution to the problems of "content governance" is, however, hardly suitable for regulating this complex of issues. The ban on discrimination only applies to the selection of journalistic and editorial content. Journalistic-editorial content can also be found on social networks. However, alongside non-professional content, this only makes up a part of the published content. It can therefore be assumed that the apps of social networks themselves are not to be classified as journalistic-editorial content. However, even if one were to assume such a classification, this would hardly mean a difference. Discrimination is defined in Section 94 (2) MStV and is a systematic deviation from the self-given rules at the expense of journalistic-editorial content. Since the self-given norms of the networks are extremely vaguely formulated, it will hardly be possible to establish a systematic violation of these very norms. Accordingly, the MStV's prohibition of discrimination is empty in the case of app stores.

The said transparency obligation is limited to the obligation to disclose the criteria of the recommendation algorithm and the criteria when an app is blocked or removed, thus again: the self-imposed standards of the app stores. The MStV does not provide for an obligation to publish a transparency report on moderation practices, nor does it provide for an obligation to remove content.

---

[14] EU: europa.eu, Antitrust: Commission Investigates Apple's App Store Rules and Apple Pay Behaviour, USA: businessinsider.com, Google Faces Nationwide Antitrust Lawsuit Over Android App Store.

[15] cnbc.com, How the Apple-Epic court ruling could affect Google.

[16] Sources

[17] europa.eu, Antitrust: Commission investigates Apple's App Store rules and Apple Pay behaviour.

[18] A comprehensive summary of the legal situation for social networks can be found here: Hans Bredow Institute for Media Research, THE LEGAL FRAMEWORK OF THE ONLINE FEDERAL ELECTION CAMPAIGN: AN OVERVIEW OF THE LEGAL OBLIGATIONS OF PARTIES AND PLATFORMS.

Thus, German national media law does not provide an effective legal framework for regulating "content governance" by app stores.

European legislative projects

At the European level, the drafts of the Digital Services Act (DSA) and Digital Markets Act (DMA) in particular provide for fundamental changes to the legal situation in digital markets. New rules will also apply to app stores in the future.

DMA

The DMA, which is still in draft form, provides for competition law measures that are intended to limit the economic supremacy of the app stores, among others. The focus here is on the classification as a so-called gatekeeper. According to Article 3 of the DMA, this refers to very large tech companies that have a strong influence on the EU internal market as providers of central platform services, hold a strong position as intermediaries (e.g. through their user numbers) between commercial users and end users, and occupy an established and permanent position (duration of business activity). The app stores of Google and Apple belong to very large tech companies and, due to their enormous market power, are also of particular importance within the EU for the mediation of customers to the developers of mobile apps. It is therefore likely that they will be classified as gatekeepers.

The DMA provides for some new regulations for gatekeepers. According to Art. 5 lit. f, gatekeepers should be prohibited from forcing business customers and end users to register for additional services. Accordingly, it would be possible in future to use Android without a Gmail account. Users must also be able to uninstall pre-installed software and apps if they are not absolutely necessary for the functioning of the operating system. This also includes the app stores themselves.

Personal data of customers may only be linked to other data of third party providers with explicit consent (Art. 5 lit. a). Users may also not be automatically registered with other services.

According to Art. 6 lit. a, gatekeepers may no longer use such non-public data of business users and their end users in competition with business users if this data was generated through the use of the gatekeeper platform. Also, data obtained from the distribution of apps would have to be shared with developers. Business customers and end users of gatekeepers should have the right to data portability and real-time access to this data. Business customers should have free real-time access to data about their sales, customers, etc. (Art. 6 lit. h and i). This should reduce the data gap between business users and platforms.

Similarly, app stores may not prohibit app developers from offering their products at lower prices outside the app store (Art. 5 lit. b DMA).

Gatekeeper operating systems must in future also allow third-party providers to install apps on the system (Art. 6 lit. c). This regulation would force Apple to open up to alternative app stores, as is standard on Android devices or PCs.

The DMA thus primarily opposes the market power of the app stores. In particular, the dominance over the data obtained should be reduced and the associated enormous competitive advantage of being able to respond ideally to the wishes of customers should be diminished. The market for mobile apps should also be opened up to app store providers other than Apple and Google. This has an indirect effect on the quasi-

content moderation of the app stores: The less dependent the developers are on the app stores, the less incisive their moderation decisions are for the developers.

However, it is noteworthy that Google's Android operating system is already open to third-party apps. Nevertheless, the market share of the Play Store is very high. The possibility of bypassing the Play Store alone theoretically offers the chance of app distribution that is self-sufficient from Google. However, removal from the App Store still leads to major financial losses and loss of reach. Nevertheless, this measure in combination with the other measures limiting the dominant position of the App Stores should lead to a decline in the market power of Google Play Store and Apple's App Store in the long run.

DSA

Unlike the DMA, the DSA directly regulates the content governance of app stores. Explicit rules such as information and transparency obligations can be found there. The DSA provides for different obligations for different categories of service providers. These are hosting service providers, intermediary service providers, online platforms and very large online platforms.

According to recital 13 of the draft DSA, online marketplaces are to be classified as intermediary services, online platforms and hosting service providers. Due to the large number of users, the Google and Apple app stores are also to be classified as very large online platforms.[19]

First of all, according to Art. 13, 23 and 33, app stores must publish a transparency report every six months, which must disclose the measures taken against content. This includes content classified as illegal as well as content that violates the terms of use. Official orders for deletion and reports from users about unauthorised content must also be disclosed here. Article 14 obliges a procedure in which users can report content as violations of state law or the terms of use.

If content, i.e. in the case of app stores an app, is blocked or deleted, the user (app operator) must receive a justification for the decision. This should form the basis for an internal complaint. A complaint system must be set up in accordance with Art. 17 and may not be carried out exclusively by computer systems. Such a procedure is already offered by the app stores. However, this does not consistently lead to a substantive discussion of the operators' concerns.[20] In order for this to be different in the future, the procedure will be supplemented by the dispute resolution body according to Art. 18. Those affected by decisions of the app stores have the possibility to appeal to an impartial dispute resolution body, which can make binding decisions for the platforms. This leads to an impartial review of the app store's decision and thus safeguards the rights of users. The fees of such a procedure may only cover the actual costs and are only to be borne by the user in the event of a decision in favour of the app stores.

As very large online platforms, app stores must undergo a risk assessment to determine whether they pose a systemic risk to, among other things, the exercise of fundamental rights, Art. 26. In order to counteract such risks, they can take measures according to Art. 27, which provides, among other things, for cooperation with trustworthy whistleblowers or an adjustment of moderation practices. Whether the very large online

[19] Market shares of iOS and Android in Germany: statista.com, https://de.statista.com/statistik/daten/studie/251737/umfrage/marktanteil-des-apple-iphone-am-smartphone-absatz-in-deutschland/: Smartphones sold in Germany: statista.com, https://de.statista.com/statistik/daten/studie/77637/umfrage/absatzmenge-fuer-smartphones-in-deutschland-seit-2008/

[20] Previously unpublished interview by Christina Dinar (HBI) with Sven Voges (PlanetRomeo) from 10.08.2021, on Google: "you also don't have a direct contact person, you kind of have an anonymous complaint hotline that you can contact, where you then also have to deal with someone else every time".

platform adequately counteracts the risk it poses to the protected interests mentioned in Art. 26 is assessed at least annually by an independent body, which must be paid by the platforms.

The DSA regulates in great detail the requirements of content governance by online service providers. It was recognised that the big question of content governance, who is allowed to decide what is publicly accessible, does not only concern the social networks themselves. The dispute resolution procedure under Art. 18 takes away some of the interpretative sovereignty of online platforms as to what content is to be classified as "harmful". Whether the requirement that only the costs actually incurred by the user are to be charged in the event of a defeat really keeps the threshold for proceedings low is doubtful for private users. However, such a procedure is still interesting for commercial users and the costs are low compared to the losses due to an absence from the app store. However, this is another problem of dispute resolution: Every day that an app is unavailable costs money and users. It is doubtful that proceedings under Art. 18 DSA will be fast enough to protect the developers' interest in a quick decision. As for state courts, this should be taken into account through a form of summary proceedings. It is true that the dispute resolution procedure does not affect the legal process before state courts. Developers therefore continue to have recourse to civil (summary) proceedings. However, the idea of the dispute resolution procedure to reduce hurdles to disputes with the platforms should also apply to urgent decisions and be taken into account in the corresponding procedure.

In particular, the obligation to justify a decision, the obligation to transparency of content governance and also the dispute resolution procedure are sensible instruments of the DSA, which steer the area of tension of content governance (what is removed when, who decides this and what can users do about it) in a good direction for app stores as well. As with the DMA, the DSA is still no more than a draft and the final wording, especially the concrete requirements for transparency reports or content governance systems, will have a decisive influence on the effectiveness of the legal instruments. However, the further legislative process could also bring positive changes such as the introduction of an expedited procedure for dispute resolution or the introduction of fixed contact persons within the framework of the complaints procedure, which would be welcome.

Outlook for other actors

In addition to the app stores, other players are influencing the distribution of content. For example, onlyfans recently announced that the service would no longer tolerate pornographic content in the future. This was due to pressure from payment providers who threatened to stop processing payments to onlyfans if the app continued to distribute pornographic content.[21] The dispute between pornhub and payment providers Visa and Mastercard was similar. Pornhub also changed its terms of use because they were accused of distributing illegal content such as child pornography or rape videos. [22]

Similarly, technical infrastructure providers can influence networks. These services, such as cloud services or DNS services, can make networks completely inaccessible. Also with these actors acting in the background, the entrepreneurial freedom of the service provider meets the freedom of the platform

[21] politico.eu, Adult content creators in the lurch as OnlyFans bans porn.

[22] cnbc.com, OnlyFans bans sexually explicit content.

operators and, depending on the size of the network, also the fundamental rights of the users active on the platform.

According to media reports, the server service Amazon Web Services will introduce a tool for content governance that recognises and removes prohibited content automatically. This is intended to abandon the practice, which has so far been inefficient and dependent on reports from users.[23]

This issue again raises the fundamental question: Who should be allowed to interfere and how? Due to the enormous reach of platforms, they also have an enormous significance for the protection of certain fundamental rights. Even pornographic platforms shape (for better or worse) an understanding of sexuality. On the one hand, it is of course welcome if action is taken against videos of rape or child pornographic material is stopped. This is indisputable in terms of content and the removal of illegal content is in the public interest. Criminal laws in Germany were passed by democratically legitimised legislative bodies. However, when it comes to legal content, this democratic legitimacy is lacking if private actors decide independently on its dissemination. For example, the dating app Planet Romeo was banned from the app stores because of an emoji consisting of a cucumber and two tomatoes.[24]

Since social networks have become important instruments for opinion formation and expression, not only the networks themselves must be held accountable, but also their treatment by providers operating in the background must be considered in regulation. This has already been done to a welcome extent by the European legislative projects. In the final formulations of the previous drafts, however, particular attention should be paid to ensuring that such services continue to be covered by the directives. It should also be reviewed whether other influential players such as payment service providers in their function as content moderators should be covered by the new legislative proposals.

Conclusion

The current legal situation grants app stores extensive freedom. They are neither subject to the German NetzDG nor does the German MStV provide for effective measures. However, this will change with the introduction of DSA and DMA. If the guidelines are adopted in their current form, this will mean significant hurdles for the business practices of app stores. In particular, the obligation to publish transparency reports and the restriction of economic power, especially the obligation to allow alternative app stores, are crucial building blocks in strengthening the rights of users and app developers to form and express opinions. When defining a complaints process, it should be considered from a regulatory point of view that the current processes theoretically offer sufficient participation, but in practice do not give much voice to the concerns of app developers.

The attempt to influence the legislative process is already in full swing because of the drastic legal changes.[25] From the perspective of content governance that protects the rights of all stakeholders, it is to be hoped that the existing draft will not be softened further.

[23] The Verge, Amazon is planning more aggressive moderation of its hosting platform AWS.

[24] Previously unpublished interview by Christina Dinar (HBI) with Sven Voges (PlanetRomeo) from 10.08.2021, on Google: "- our app has been blocked so please take out this picture (cucumber and tomatoes)".

[25] lobbycontrol.de, DSA/DMA: How Big Tech wants to prevent new rules for digital platforms.

(Niche) Platforms as Experimental Spaces for Content Moderation - Mapping small, medium and niche platforms online

Christina Dinar and Lena Hinrichs
LEIBNIZ INSTITUTE FOR MEDIA RESEARCH | HANS-BREDOW-INSTITUT

Introduction[26]

So far, there has been a lack of debate in platform governance about platforms beyond #BigTech, i.e. Google, Facebook and Co. However, surveying content moderation on smaller platforms provides an insight into experimental fields of an area that is rarely accessible on large platforms. So far, research has had to rely on individual whistleblower reports and journalistic-investigative reporting. What exactly happens in content moderation and how it also changes often remains in a black box.

Smaller platforms offer researchers the opportunity to survey this area and thus gain more transparent insights into different forms of content moderation - and without being forced to do so by legislation. A survey of their practical, innovation-driven approaches to content moderation allows us to show how the balancing act between deleting unwanted content and preserving freedom of expression can be achieved without regulatory mandates.

How can small and medium-sized platforms (SMP) be described?

A systematic and comparative survey of small and medium-sized platforms and niche platforms that are not dominant market players such as Facebook, Twitter and Youtube does not yet exist. However, in order to capture the multi-layered content moderation and the challenge of election-related content, it is precisely these platforms that are interesting. They often have approaches in this area that are unconventional and community-based. These platforms, which are part of a larger landscape, are now categorised here for the first time in a taxonomy in order to define descriptive terms and to map the differently situated conditions that make up a smaller and medium-sized platform and/or niche platform.

Definition of small and medium-sized platforms (SMP)

Size of the platform according to the number of users

SMPs are those which, due to their size, do not fall under the German NetzDG and the Digital Services Act (DSA). The NetzDG only regulates platforms that have more than 2 million registered users in Germany. The regulations of the DSA apply in principle to all platforms. Small and micro enterprises are exempt. The classification as such a company is based on a recommendation of the EU Commission and includes

[26] More about the project "Niche Platforms" can be found in the podcast episode of the BredowCast with Christina Dinar, in which further details and background of the project are explained. More on social networks in election campaigns can also be found in the podcast of the Weizenbaum Forum. Christina Dinar talks about inclusive digital spaces in an interview in Americas Dialogues.

criteria such as turnover or the number of employees. Stricter rules, including on content moderation, apply to platforms with more than 45 million users. These are called Very Large Online Platforms (VLOPs).

Thematic design of the platform

Another legal distinction of the "major platforms" from SMP is the indirect third-party effect of fundamental rights. Some civil courts decide on the basis of the thematic "cut" of the platform. If there is a clear thematic focus (e.g. "Forum for Dachshund Friends"), platforms may remove posts that do not fit the focus at their own discretion ("In the Forum for Dachshund Friends, please only postings on dog-related topics, everything else will be deleted!) However, if the platform serves the "general exchange of information and opinions of users without thematic limitation", the decision on deletion cannot be entirely "at the discretion of the platform operator". The virtual domiciliary right cannot then simply be executed by the platform. SMP can also be described as so-called niche platforms, this describes the thematic limitation to a "niche topic" (such as "Dackelfreunde Forum"). The two terms SMP platforms and niche platforms can overlap, also in the platform's self-description, but they are two different categories.

Definition of niche platform = "extension of a platform space".

More than just along with user numbers, niche platform describes a place for interest-based exchange; a small protected area where one can stay and develop unhindered by competition. Very often, the niche is seen as a place with development potential, because target group-specific, precise marketing is possible. In ethnographic subculture research, the niche culture is a sideshow for subversive practice or even social corrective. The internet itself was also a subcultural niche for a long time with a hacker/nerd culture developing in the 90s. The structural-technical decentralisation of the internet still enables niche subculture formation today and is thus also a particularly important means of exchange and networking for marginalised groups.

Platform definition

In the context of this survey, platforms are "social media platforms" with user-generated content, so-called "UG content-hosting and interaction platforms". This refers to content that has been created by the user him/herself and with which the user can interact. The dimensions of social interaction and self-created content are particularly important here, similar to the terms social media or social networks.

User number as a value of regulation of the "internet-by-size" and as a classification category for SMP platforms

The Monthly Active User (MAU) unit as an organising element in the platform landscape

The number of users on a platform has become a value of regulation itself. This kind of "regulating internet by size" is to be critically evaluated under various aspects. Laws such as the NetzDG as well as draft laws such as the DSA and various discussions on reforming the US law Section 230 all work with limits on user units (e.g. NetzDG: 2 million domestically registered users).

These units are created differently depending on the nation and are handled very differently in the practice of the platforms themselves, especially since this number is hardly to be grasped completely. Frequently,

reference is made to Monthly Active User (MAU), yet there are uncertainties in the way this value is defined. For example, no distinction is made between a single, identified human visitor to a website ("unique visitor") or whether the website is accessed many times by the same user account. Large platforms such as Twitter and Facebook also show differences in their information about their MAUs. Thus, the specification of exact user figures remains with the self-interpretation of the platforms and their own definition of MAU.

This gives certain platforms leeway to include themselves among the regulated large platforms or not. At the same time, the market is manifesting itself and there is a dichotomy: very market-dominant platforms with large user numbers (albeit in unclear data) receive high compliance requirements; and smaller and medium-sized platforms that do not have these requirements. The latter weigh up their own growth in user numbers very carefully - also in order to avoid the cost-intensive compliance requirements for the time being.

The MAU/user number is important as a classification and differentiation category for the taxonomy of small and medium-sized platforms. This reorders the entire platform field in Germany and introduces this practical distinction. This possibly creates and manifests a structural power imbalance and consequently a changed economic and social dynamic of platform offerings in Germany.

Unit of page views as a display and generation of publicity

Another unit is the "monthly sites visits", sometimes also called "page impressions". These mainly indicate the audience that the content of the platforms can reach. The SMP may have a small number of registered users, but still provide content that has a high number of views and interactions. This indicates a transfer of a content to a broader general public, thus leaving a niche and no longer serving a niche audience (e.g. "Dachshund Friends Forum"). It is also possible that in such cases the jurisdiction of the thematic nature of the platform itself no longer applies (e.g. "The dachshund forum has contributed a critical-satirical article on dog tax to the daily political debate"). In such cases, a so-called "cross-platforming" of the content can usually be assumed (e.g. on a large platform such as Twitter, this content of the Dachshund Forum was linked and discussed).

Even a high search engine indexing on a specific issue can develop general public and information relevance in a current context (e.g. in the context of the current flood disaster in western Germany: information in a geological expert forum on severe soil erosion associated with sudden heavy rain) and can suddenly bring high page impressions.

This category indicates that even KUMPs can develop relevance and publicity - albeit usually only selectively - and also potential growth without necessarily having many MAUs.

In the taxonomy created for KUMP (see appendix), the two units are listed separately, as they want to describe different forms and relevance of KUM platforms in a differentiated way. It remains difficult to obtain user figures, such as Monthly Active User (MAU) figures for the SMPs, and to find verified sources and consistent self-reports on them.

Recommendation on regulation with User numbers

- The User/MAU figure itself, if it is an integral part of regulation, should be regularly reviewed and revised (if necessary by a non-governmental expert opinion), also to allow for exclusions, economic development and growth potentials.
- Cross-platforming, especially in the case of problematic content (e.g. illegal content, election-related content, etc.) should increasingly bring platforms into exchange with each other. SMPs should come into low-threshold contact with larger platforms, not only through "crisis situations". The exchange can be accompanied and coordinated by an independent body that develops and issues recommendations from the exchange.
- Regular evaluations of the extent to which the law inhibits economic potential and can sustainably secure public and media diversity in platforms are recommended.

Overview of small and medium-sized platforms (KUMP)

The aim of the survey was to record the small and medium-sized platforms' (SMP) practices and policies on UGC, moderation, community management, federal election-related content, their cooperation with law enforcement agencies and a general risk assessment.

The selection criteria for the platforms surveyed were:

- used and represented in the German-speaking region (DACH region);
- small and medium-sized social media platforms (SMP) or niche platforms (NP);
- are "social platforms", i.e. have UGC and interaction possibilities with user-generated content;
- all are (socially) entrepreneurial, not non-profit.

Results of the investigation

Often, the focus of the working methods of small and medium-sized platforms is on a close interlocking of people-oriented content with the community that creates the content. In some cases, active engagement with their own communities is encouraged, which often helps to enforce in-house rules on platform use or refers to them in disputes. However, the moderation systems of the platforms studied, some of which are very different, assign an important, albeit different, role to community participation.

- **The German NetzDG seems to influence the development of guidelines and policies in companies - even though this law does not affect them at all.** The law seems to have an indirect effect in that at least consideration is given to adapting it if the prescribed measures seem sensible and useful to the platforms. These are in the area of improving reporting channels, but by no means in the transfer of users' data. Restoration of deleted content in the case of objections by users was also seen as very critical.
- **Manual moderation (i.e. moderation that takes the time to look into the context of the posters) supports the content moderation team's decided examination of the posting and its environment and nuances it.** An example: so-called "agenda setting in the election campaign" is quickly prevented, since behind it there are usually accounts that are only founded for a special purpose and whose goal is to denigrate political opponents. This also ties in with an increase in negative campaigning that has been measured among conservative groups in online election campaigns. Manual moderation also helps to stop agenda-setting in the area of digital violence, such as

LoveScams and cyber-grooming, but also to stop offers or business interests that are prohibited on platforms.

- **Dealing with election-related content is always subject to the platforms' own rules. The platform rules range from allowing explicitly political groups to profile pictures with party references and tolerating the activity of candidates for the Bundestag or deliberately restricting the topic**. Some platforms also consciously and actively support the call for democratic elections.
- Community participation (e.g. as verified "trusted flaggers" or in the peer review process during the approval of uploaded images) **seems to lead to increased identification of users with the platform and its rules**. The platforms differ in the variety and method of active or passive moderation and the points at which the community is involved.
- **From an organisational sociological point of view, an initial structural neglect of the development of content moderation can be observed in the companies**. Relatively independent of when the company entered the start-up phase with the platform, the area of content and complaint management within the organisational structure was often given few priorities and resources for development at the beginning and has only received expansion, standardisation and development in recent years.

Furthermore, the platforms were generally open to science and research and were willing to cooperate and provide data for such surveys. Only a few platforms have been studied so far, but there is the prospect of being able to enrich this void with a more evidence-based survey to support a diverse, differentiated and informed debate on this in the field of regulation and future rule-making, and to be able to make nuanced regulatory proposals.

References

Corporate Finance Institute (2021). Monthly Active Users (MAU) - Definition, Uses, How To Calculate. Retrieved 29.07.2021. https://corporatefinanceinstitute.com/resources/knowledge/ecommerce-saas/monthly-active-users-mau/.

Crunchbase Discord (2021). https://www.crunchbase.com/organization/discord.

Crunchbase Gab (2021). https://www.crunchbase.com/organization/gab-online.

Crunchbase Quora (2021). https://www.crunchbase.com/organization/quora/signals_and_news.

Crunchbase Reddit (2021). https://www.crunchbase.com/organization/reddit.

Crunchbase Tellonym (2021). https://www.crunchbase.com/organization/tellonym/technology.

Crunchbase Twitch (2021). https://www.crunchbase.com/organization/twitch.

Crunchbase Wize.Life (2021). https://www.crunchbase.com/organization/seniorbook.

Crunchbase. Gutefrage.Net (2021). https://www.crunchbase.com/organization/gutefrage-net.

Crunchbase. Yodel (2021). https://www.crunchbase.com/organization/jodel.

Crunchbase. Knuddels (2021). https://www.crunchbase.com/organization/knuddels.

Crunchbase. Mydealz (2021). https://www.crunchbase.com/organization/mydealz.

Fertmann, Martin, and Keno C. Potthast (2021). Digital Time-outs for Trump: The Beginning of the End of the Privileged Treatment of Incumbents by Social Networks? In: JuWissBlog, Retrieved 30.06.2021. https://www.juwiss.de/05-2021/.

Frenzel, Nils (2018). What I experienced when I chatted on Knuddels again after 15 years. In: Vice, 13 July, https://www.vice.com/de/article/zmkv7w/bock-zu-chatten-ich-habe-mich-nach-15-jahren-wieder-bei-knuddelsde-angemeldet.

Goldman, Eric, and Jess Miers (2021). Regulating Internet Services by Size. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, May 1, 2021. https://papers.ssrn.com/abstract=3863015.

Harasim, Paul (2016). Gay Entrepreneur Found the Road to Success Full of Slurs. In: Las Vegas Review-Journal (blog), October 21, 2016. https://www.reviewjournal.com/news/news-columns/paul-harasim/gay-entrepreneur-found-the-road-to-success-full-of-slurs/.

Hasse, Michael (1994). Die Hacker: Strukturanalyse einer jugendlichen Subkultur. Master's thesis, Faculty of Philosophy, Rheinische Friedrich-Wilhelms-Universität zu Bonn.

Kaplan, Andreas, and Michael Haenlein (2010). Users of the World, Unite! The Challenges and Opportunities of Social Media. In: Business Horizons 53 (February 28, 2010): 59-68. https://doi.org/10.1016/j.bushor.2009.09.003.

Kreuter, Sara (2017). Jodel-App: Anonym, lokal und lustig kommunizieren. In: FAZ, 07.04.2017. https://www.faz.net/aktuell/stil/leib-seele/jodel-app-anonym-lokal-und-lustig-kommunizieren-15126386.html.

Lux, Torben (2019). 'How Mydealz knocked Amazon off number 1 in the app charts with just one Whatsapp message'. Retrieved 21 July 2021. https://omr.com/de/mydealz-fabian-spielberger-appcharts-marketing-black-friday/.

Nelly (2019). Maintaining Trust and Safety at Discord with Over 200 Million People." In Medium, February 21, 2019. https://blog.discord.com/maintaining-trust-and-safety-at-discord-with-over-200-million-people-f0b39adfc00c.

New Work SE (2021). Annual Report 2020 New Work SE (Xing, kununu). https://www.new-work.se/NWSE/Investor-Relations/Geschaeftsberichte/de/NEW_WORK_SE_GB_2020.pdf.

o. Verf. (2021). Grindr. In: Wikipedia, 19.05.2021. https://de.wikipedia.org/w/index.php?title=Grindr&oldid=212121365.

o. Verf. (2021). nebenan.de. In: Wikipedia, 20.06.2021. https://de.wikipedia.org/w/index.php?title=Nebenan.de&oldid=213130795.

o. Verf. (2021). Niche. In: Wiktionary, 20.08.2020. https://de.wiktionary.org/w/index.php?title=Nische&oldid=8075204.

o. Verf. (2021). PlanetRomeo. In: Wikipedia, 21.04.2021. https://de.wikipedia.org/w/index.php?title=PlanetRomeo&oldid=211169824.

o. Verf. (2021). Researchgate. In: Wikipedia, 25.06.2021. https://de.wikipedia.org/w/index.php?title=Researchgate&oldid=213270844.

o. Verf. (2021). Scruff (app). In: Wikipedia, 24.04.2021. https://en.wikipedia.org/w/index.php?title=Scruff_(app)&oldid=1019596749.

o. Verf. (2021). Tellonym. In: Wikipedia, 29.05.2021. https://de.wikipedia.org/w/index.php?title=Tellonym&oldid=212479802.

o. Verf. (2021). Twitch. In: Wikipedia, 11.06.2021. https://de.wikipedia.org/w/index.php?title=Twitch&oldid=212867902.

o. Verf. (2021). wize.life. In: Wikipedia, 06.05.2021. https://de.wikipedia.org/w/index.php?title=Wize.life&oldid=211676191.

o. Verf. (2021). gutefrage.net. In: Wikipedia, 28.06.2021. https://de.wikipedia.org/w/index.php?title=Gutefrage.net&oldid=213369959.

OLG Munich (2020). Final judgement of 07.01.2020 - 18 U 1491/19 Pre - OpenJur. Retrieved 21.07.2021. https://openjur.de/u/2294930.html.

Remmers, Ina (2020). Facts and figures on nebenan.de. In: nebenan.de, 09.04.2020. https://presse.nebenan.de/pm/zahlen-und-fakten-zu-nebenan-de.

ResearchGate (2020). ResearchGate Turns 12. In: ResearchGate, 23.05.2020. https://www.researchgate.net/blog/post/researchgate-turns-12.

Saferinternet.at (n.d.). Jugend-Internet-Monitor 2018. Retrieved 30.06.2021. https://www.saferinternet.at/presse-detail/jugend-internet-monitor-2018/.

Stark, Lars (2019). Yodel: Why it's worth taking a look at the young network. In: BASIC thinking (blog), 09.09.2019. https://www.basicthinking.de/blog/2019/09/09/jodel-portraet-marketing/.

Statista (2021). Grindr User Number 2016. Retrieved 27.01.2021. https://www.statista.com/statistics/719621/grindr-user-number/.

Statista (2021). Xing - Members of the platform in the DACH region 2021. Retrieved 09.07.2021. https://de.statista.com/statistik/daten/studie/481399/umfrage/anzahl-der-xing-nutzer-in-der-dach-region/.

Statista (2021). Xing - Unique Users in Germany 2019. Retrieved 09.07.2021. https://de.statista.com/statistik/daten/studie/418088/umfrage/online-besucherzahlen-von-xing-als-zeitreihe/.

Statista (2021). Xing.com - Visits worldwide 2021. Retrieved 25.11.2021. https://de.statista.com/statistik/daten/studie/1021448/umfrage/anzahl-der-visits-pro-monat-von-xingcom/.

t3n Magazine (2013). Interview with the founder of MyDealz: I always hunt for bargains everywhere. In: t3n - digital pioneers. Retrieved 07.07.2021. https://t3n.de/magazin/interview-mydealz-grunder-fabian-spielberger-jage-immer-232790/2/.

Facebook and Artificial Intelligence: A Review of Good Practices

Beatriz Pérez –Acuña, Celia Fernández-Aller[27]
UNIVERSITY OF HUELVA//UNIVERSIDAD POLITÉCNICA DE MADRID

Introduction. Facebook and Artificial Intelligence

This article compiles the ethical controversies reflected by journalists appearing in the media about Facebook in relation to the use of artificial intelligence (AI) in the different products of this company, with a special focus on content moderation. Different applications of AI are shown, with both positive and negative impacts.

A human rights approach (HRA) is being used to avoid the risks of "ethics washing" (Floridi, 2019) and the co-optation of AI ethics by big tech (Jobin, 2022). Many ethical recommendations are currently in place for developing algorithms (Fjel, Achten, 2020). However, recommendations that focus on human rights are still scarce. As Fjel suggests: "On its own, a set of principles is unlikely to be more than gently persuasive. Its impact is likely to depend on how it is embedded in a larger governance ecosystem, including for instance relevant policies (e.g. AI national plans), laws, regulations, but also professional practices and everyday routines".

The HRA is based on strong principles (such as universality of human rights, participation, transparency, accountability, sustainability...), and it has the element of enforcement, which allows rights holders to claim rights. Additionally, human rights are universally accepted, and this makes a great difference compared to the wide variety of AI codes of ethics. Principles alone cannot guarantee ethical AI, due to the problematic implementation (lack of proven methods) and the lack of accountability.

International human rights law is a governance regime with significant potential relevance to the impacts of AI. There is a strong connection between AI governance and human rights laws and norms. "64% of our documents contained a reference to human rights, and five documents took international human rights as a framework for their overall effort. Existing mechanisms for the interpretation and protection of human rights may well provide useful input as principles documents are brought to bear on individuals cases and decisions, which will require precise adjudication of standards like "privacy" and "fairness," as well as solutions for complex situations in which separate principles within a single document are in tension with one another" (Fjel, Achten, 2020).

Facebook and artificial Intelligence at the centre of media controversy

In September 2021, the 37-year-old Haugen, who worked as an engineer at Facebook until May 2021, in the division for civic integrity, a unit within the organisation that aims to recommend policies that protect the general public. The former employee of the company, who exposed the organisation's darkest secrets, left the company, but hours before leaving, she pulled copies of dozens of documents that she said "prove that Facebook has always been aware that its algorithms fuel division, promote hate, spread fake news, and can have a profound impact on the emotional and physical health of teenagers. And instead of correcting this,

[27] Both authors have contributed equally.

the company has looked the other way, privileging its growth and profits rather than choosing to protect its customers" (Gomez, 2021 para.3). She disclosed tens of thousands of internal Facebook documents to the US Securities and Exchange Commission and the Wall Street Journal: "last week in a room on Capitol Hill, lawmakers from both parties gathered to hear testimony from Frances Haugen" (Gómez, 2021, para.1).

Goméz (2021) reports that according to Haugen, "the main problem is that Facebook's virality depends on how the division in charge of the company's growth amplifies the algorithm to ensure that the content "sticks more and is reproduced". “The more this happens, the more time users spend on the networks and the more profit Facebook makes, because it can expose them to more ads. And numerous studies have shown that the more controversial and divisive the content, even if it is false, the more traffic it generates. Haugen argues that Facebook is aware of this situation, but prioritises its own coffers" (Gómez, 2021, para. 6).

Haugen confirms what Facebook denies, among others, that "its social networks are ideal platforms for spreading disinformation, with the documents leaked by Haugen confirming that Mark Zuckerberg and his staff also know this, although they deny it" (Bécares, 2021 para.4). Artificial Intelligence, used to filter and select information, is still many years away from being perfected. Haugen says: "that the company encourages content that angers, polarises and divides". On the other hand, he says that "internal studies show that the company is lying to the public about significant progress against hate, violence and misinformation" (Zornoza, 2021 para.3).

According to the Facebook engineer, "it doesn't matter whether the information is true or false" (Genbeta, 2021 parr.12). The algorithm does not prioritise this aspect and Haugen says that, "it is since 2018 that this aspect is so marked. In fact, he recalled how European leaders have openly questioned this fact" (Genbeta, 2021 parr.12). With this, the former employee says: "Facebook makes more money when more content is consumed. People engage more with things that provoke an emotional reaction. And the more anger they are exposed to, the more they interact and the more they consume" (Perez, 2021 para.12). These internal studies point, Haugen argues, to the fact that "Facebook has been lying about significant progress against hate, violence and misinformation. According to an internal report, after all the changes, hate on the platform would have been reduced by 3-5%" (Perez, 2021 para.11). Haugen accepted the job at Facebook in 2019, making it a condition to work against disinformation on the platform.

One of the serious problems Haugen sees is that, as users: "You have your phone. You could see 100 pieces of content if you sit down and scroll for just five minutes. Facebook has thousands of options it could show you. The algorithm chooses among these options based on the type of content you have interacted with most in the past" (Genbeta, 2021 parr.11). Facebook is showing us in our 'Feed'. This is where the famous algorithms come into play, deciding what the consumer sees and does not see. The algorithm chooses some content over others "Social networks make us sick" accuses the platform of being intentionally as addictive as tobacco (Araújo, 25 September, 2020). All these statements by Haugen have put the social network in the spotlight of an ethical debate.

An internal audit warns Facebook that its inaction in the face of hate speech is "a step backwards for civil rights" (Araújo, 10 July, 2020). According to the former directive, when we are confronted with an information space that is filled with exalted, hateful or polarising content, social peace is somehow breached: "it erodes our civic trust, it erodes our faith in others, it erodes our ability to want to care for others" (Genbeta, 2021 parr.8). As a consequence, according to Genbeta (2021) "today's version of Facebook can fragment our societies and cause ethnic violence around the world" (para.8), he said in reference to hate

crimes that take place from some people to others because of their skin colour or religion. And he recalled "the ethnic cleansing that took place in Myanmar in 2018, when the military used Facebook to present its genocide and encourage hate speech towards the Rohingya ethnic minority" (Genbeta, 2021, para.9).

The differences between nudity in pornography and a nude body in art are not at all clear; the AI's selection is bizarre. The most emblematic statues of the renaissance have already been victims of this problem. On the other hand, we know about sex trafficking through the net. It has been discussed in the past how Facebook and its social media empire cares more about people's bodies than about racism or hate speech. The Vienna Tourist Board said that "museums in Austria's capital have faced many online challenges in displaying their works. After Vienna's Natural History Museum posted images of the Venus of Willendorf, a 25,000-year-old Palaeolithic limestone statue", Facebook deleted the images and labelled them pornographic. "In 2016, Facebook removed the Pulitzer Prize-winning photograph "Napalm Girl", which shows a naked girl fleeing a napalm attack during the Vietnam War. The photo had to be restored after AI censorship, reminding us of the long list of photographs removed by Facebook in its history. Facebook, and other social networks, may even have more excuses to remove nude images in part because of the threat of SESTA-FOSTA, a US law against sex trafficking" (Bécares, 20 October 2021).

Another source of debate and dispute is Facebook, which found itself in trouble after a new bias in its algorithms was revealed that had discriminatory results. In this case, the company was the cause of controversy: "after it became known that Facebook's AI labelled a video of black people as "primates", the social network apologised and said that the error was "unacceptable"". In this case the company was controversial because its AI labelled a video with black people as primate-related content" (Erard, 6 September 2021, parr.1). The gross labelling error caused by the artificial intelligence was detected. According to the report, the video in question belongs to the British Newspaper *Daily Mail*. It was uploaded to Facebook on 27 June 2020 and, as described by the NYT, shows "black men in altercations with white civilians and the police" (Erard, 6 September 2021 parr.3).

Below the player, the social network's recommendation algorithm "asked users if they wanted to continue watching videos about primates" (Erard, 6 September 2021, para.3). The problem that is putting Facebook in trouble is not new. Several companies have come under scrutiny in recent years for algorithmic biases that favour discrimination. This has been seen mainly in facial recognition technologies that "are especially erratic when processing dark skin tones" (Erard, 6 September 2021).

Another clear example, at Facebook "hundreds of engineers are rebuilding how its ads work, as announced by Graham Mudd" (Bécares, 11 August 2021, parr.2), one of the company's top advertising executives. "Facebook claims it needs to track us to make Instagram free, and the question is who would pay €2 a month if they made it paid" (Bécares, 3 May 2021). Facebook's new rhetoric about making advertising more privacy-conscious could even be considered a defeat considering its previous statements about its use of people's information. Take its big campaign to object to Apple's ad tracking on the grounds that it was acting anti-competitively and hurting small businesses that relied on ads to reach customers: "Now Facebook says it is working on new approaches that respect privacy. It is worth remembering that it also has to do this because of legislation" (Bécares, 11 August 2021 para. 6).

The European Union is considering a ban on micro-targeted ads as part of a broad legislative proposal called the Digital Services Act, and the US government has recently signalled its "interest in monitoring "user surveillance" by "dominant internet platforms" (Bécares, 11 August 2021 para. 6).

Two years ago, "Apple threatened to remove Facebook and Instagram apps from its app shop" (Clarín, 2021). They were concerned that these platforms were being used as "a tool for trading and selling handmaids in the Middle East. In internal documents they stated that they were underestimating "confirmed abusive activity". There were domestic workers from the Philippines who shared on their social media accounts of being abused. Within Facebook, "you can see accounts with posed photographs of African and South Asian women, with ages and prices listed next to their images" (La Nación, 25 October 2021). The newspaper 20 Minutos (2021) notes that: "The Philippine government has a team of workers assigned to track" Facebook posts in an attempt to protect its citizens seeking work in Middle Eastern countries (AP, 25 October 2021).

Apple CEO Tim Cook criticises Facebook and others for not respecting privacy. According to Cook, these networks prioritise "conspiracy theories and incitement to violence" (Bécares, 29 January 2021). These are topics that engage more people in these conversations. In this way, the companies that run these social platforms have the opportunity to collect more personal information from citizens and thus have more data to sell.

Another schandal was the recent communications outage, which not only created chaos that ended up costing billions of dollars in losses, but also revealed the need for the use of this network by users. As a consequence, the CEO apologised for the fall of Facebook, WhatsApp and Instagram, somehow avoiding further criticism. In his own words: "Facebook, Instagram, WhatsApp and Messenger are back online now. Sorry for the disruption today. I know how much you rely on our services to keep you connected to the people you care about," Zuckerberg said in a terse message posted on his popular social network to alleviate the crisis cabinet (La voz de Galicia, 2021 parr.3).

The Metaverse project

Another demand that will not be easy to achieve is that of a report on the problems that may arise from the development of the metaverse, such as possible psychological or human rights harms caused by the use and abuse of the platform, and whether these can be avoided or are inherent to the evolution of the technology. Once this is done, and with all the information available, they are calling for a non-binding advisory vote by the board on whether shareholders consider it appropriate to proceed with the implementation of Zuckerberg's grand project. This CEO is gambling everything on the metaverse, and he is taking a huge risk. No tech company has ever invested so much in a single project to date. For example, this group is requesting that, "Meta produce a report analysing why the company's moderation policies have not been effective in controlling the dissemination of hateful, violent or misinformation content, and that this research be presented to the shareholder meeting. They also call for a further investigation to measure the actual and potential impact of targeted advertising on the human rights of users of the various platforms that make up Meta, and for that information to be published on the company's website by 1 June 2023" (Rodriguez, 2021 final).

Zuckerberg appears not only as CEO of the company, but as an individual defendant, exposing the social media mogul for the first time to possible financial and even criminal penalties. Racine believes that, in this case, "adding Mr Zuckerberg to the lawsuit is fully justified, and sends the message that any corporate leader, starting with the CEO himself, should be held accountable for his own actions" (Merino, 20 October, 2021 para.1) The metaverse is a hypothetical, but it is one that is likely to become a reality in a few years. "Facebook already has a gigantic list of plans to be the leading provider of virtual space for people to escape the real world" (Gonzalez, 28 October 2021 final).

Just as the company has a few years to try to do better this time, we as users also have a few years to consider its less bright and optimistic implications. "The internal business restructuring will make "Metaverse" the new Product group that will encompass all the different Augmented Reality and Virtual Reality related developments that the company holds" (Miguel, 2021 para.7). "Metaverse", the term used by Facebook to define its next steps in building a virtual universe that functions as "the next version of the internet", is so popular that it functions as a noun with a life of its own. Don't call it Facebook, call it Meta. The company changes its name and bets everything on the metaverse. The metaverse in Facebook's idea: a three-dimensional virtual zone in which humans, represented by avatars, interact in various ways. It is an evolution of the internet. Metaverse users access the environment through terminals that project them into the virtual environment.

"Mark Zuckerberg kicked off Facebook Connect 2021 by talking about the future, and for him it has only one name: the metaverse" (González, 28 October 2021). The Facebook CEO and company spent more than an hour describing a virtual world in which we will all want to live and, of course, spend money. The idea is by far a new one and virtual worlds are a thing of the past thanks to video games. However, Facebook's grand vision is to turn this into a kind of alternate reality in which to live, in the manner of the film, Ready Player One: "In fact, Zuckerberg believes the metaverse will replace the mobile web in the future and transcend the plane of screens" (Gonzalez, 28 October 2021). The problem is that no such world currently exists, and if it did, there is no way to enter it. Again, however, none of this is really accessible. Such spaces would require virtual and mixed reality equipment that simply does not yet exist, let alone be accessible to the masses, but which "Facebook clearly wants to build. Oculus is a sample of this and will be the starting point" (Gonzalez, 28 October 2021).

Facial recognition

Facebook's facial recognition had been active and at the centre of privacy debates for several years, and was intended to recognise members of the social network in any photo uploaded to it. Two years ago we even saw intentions to use mobile phone cameras to recognise faces wherever we go, something that made more than a few people panic. That doesn't mean that Meta is no longer interested in facial recognition: "the company advocates the technology to help people know when someone uploads photos of them, to unlock devices like Apple's Face ID or to prevent fraud" (Lopez, 2 November 2021 para.3). However, "at the same time, he admits that there are "growing concerns" in society and that nothing is yet clear at the regulatory level" (López, 2 November 2021 para.3).

Facial recognition systems are becoming increasingly common in everyday technologies. Mobile phone unlocking is the most obvious example, but not the only one. "We've seen for example facial recognition in security cameras, as we are sure to wonder, not so much according to new research, which has shown that this system can be bypassed with a virtual face" (Rus, 13 August 2021, para.2). Artificial intelligence is used to recreate a face that meets most of the face characteristics that facial recognition systems look for. Creating a face similar to the majority of the population is all it took for these researchers to overcome a number of existing facial recognition systems. "The researchers' moral of the story is that facial recognition systems may not be as secure as they seem. They propose using alternatives or extra checks so as not to rely entirely on a facial recognition system that, as we can see, is relatively easy to circumvent" (Rus, 13 August 2021, final para.).

TextStyleBrush

Then it's the turn of the project about handwriting. A new Facebook project manages to imitate most handwriting by reading a single word. Based on this word, the artificial intelligence learns what the typeface looks like and uses it to write any text: "TextStyleBrush is the name of Facebook AI's new project. It is an artificial intelligence that can copy the style of a text found in a photograph. What undoubtedly makes it impressive compared to other similar artificial intelligences is the small amount of data it requires: just a few letters from a word. In other words, you could theoretically write your name and the AI would know how to imitate your handwriting" (Rus, 14 June 2021 para.2). Facebook explains that while most AI systems can do this with specific tasks, in this case TextStyleBrush is flexible enough to understand all types of fonts from real environments. That is, it understands texts with all types of calligraphy, applied styles, thicknesses, colours, shading or rotations without apparent problems. They indicate that to achieve this, its artificial intelligence tries to understand the typography it analyses from a more holistic point of view. Rus comments: "They say there are still limitations and specific points to overcome, for example when the AI has to deal with metallic text and reflections" (Rus, 14 June 2021, final para.).

The actual usefulness of having an AI copy your handwriting. As is often the case with such developments in artificial intelligence, it can have both positive and potentially negative ethical consequences: "In theory such an AI could be used to forge someone's handwriting and speak on their behalf, for example" (Rus, 14 June 2021, para. 5). But also for positive applications. For example, it can be used to instantly translate real-world texts by simply photographing or pointing the camera. Be that as it may, for the time being it is only a public research by Facebook. We'll see if in the next few trials the improvement can become popular when applied to a real app or tool.

Glasses research and Immersive realities

A recent foray by Facebook into the realms of augmented reality (AR) could have a significant impact on the limits of artificial intelligence (AI) and its presence in our daily lives. "Just days after unveiling its AR glasses produced in collaboration with RayBan, Mark Zuckerberg's company announced the development of a project that could teach AI to understand the world as we see it." A quick glimpse of this process among others, Facebook keeps a record of all the websites you visit: "However, to reach this end, Facebook's experiment would have to see, hear and remember the personal information its users perceive on a daily basis" (Olaskoaga, 2021). In this vein, Facebook also announces work on virtual reality: "The company has already commercialised a popular virtual reality headset, the Oculus Quest 2, and plans to move from VR to augmented reality in the next few years" (Gilbert, 15 October 2021).

Facebook is researching AI systems to see, hear and remember user actions. Now the company's intentions with its AI developments are revealed. "If you were suspicious of their glasses in collaboration with Ray-Ban, you were right" (Márquez, 2021, parr.2). Recording your day through a pair of glasses and sharing your life with the company has consequences. What is done with all that information? Although they assure users that they will be protected. Now, through research, it aims to discover the scope of its technology: "We know that Facebook is investing a lot of resources in its smart glasses and augmented reality, but at the moment, beyond recording video and taking photos, they don't have great capabilities. That could change in the future thanks to artificial intelligence. How? Well, one of the possibilities is to give the AI the ability to understand what is seen from our eyes, listen and manage to "remember scenes" (being able to answer questions like "where did I leave my keys?") or "remember words" (being able to answer questions

like "what did John say the other day?"). A most striking project that at the same time generates multiple doubts" (Pérez, 2021, parr.1 and 2).

Facebook wants the artificial intelligence of the future to be able to "forget" irrelevant data. "Researchers at the social network are working on a method to train an AI to forget information that loses relevance. They would achieve this by setting expiry dates. Facebook's AI project is very interesting, especially in times when the volume of available information is increasing. For this reason, the researchers propose a method called Expire-Span. It is designed to enable artificial neural networks to more efficiently classify and store information related to the tasks they have to perform" (Erard, 5 September 2021).

Facebook uses AI to remember everything the user does: "Facebook has developed a system for its Artificial Intelligence (AI) mechanisms that teaches them to forget certain information when it is not important to perform their functions, saving on memory and processing costs. The new technology, called 'Expire-Span', is "the first operation of its kind" according to the company, and aims to make AI neural networks resemble the workings of the human brain, which constantly forgets data, as Facebook said in a statement" (Altmann, 14 May 2021, paras 1 and 2). The use of Expire-Span offers benefits to some common AI mechanism tasks such as character-level language modelling, and improves efficiency in long-term context tasks in language, reinforcement learning, object collision and algorithm tasks.

Facebook has been shown to profit from every piece of data given to the platform: "One point that is very much worth commenting on is Facebook's ability to make money from users who do not pay for the service. The average monthly revenue per user (globally) was $1.99 during the last quarter". "If we narrow it down to users in the US and Canada, the figure is much more impressive: $8.63 per month per user. More than what is achieved by services like Netflix or Spotify, directly paid services that, between family accounts, trial months or specific discounts have lower average revenues". "This is possible thanks to Facebook's use of our data. Theoretically, and security scandals aside, Facebook does not sell our data to third parties, but sells third parties access to us thanks to the use of our data" (Lacort, 2018 parr.7,8 and 9).

Ego4D

This may change in the future thanks to artificial intelligence. One of the possibilities is to endow AI with the ability to understand what is seen from our eyes, to listen and to "remember scenes", being able to answer questions such as: "where did I leave my keys" or "remember words", "what did John say the other day? (Erard, 2021). A most striking project that at the same time raises many doubts about the ethical ground in terms of the use applied: "Ego4D is a long-term project in which Facebook plans to investigate the ability of AI to understand and interact with the world as we do, from a first-person perspective" (Márquez, 2021 parr.4). The company's ambitions with AI projects are becoming increasingly troubling from an ethical standpoint. As certain devices will give it the power to constantly analyse people's lives using material obtained from first-person recording. In the media, privacy is still at stake. For the moment, "Zuckerberg's company says it is only a research project and not a commercial development" (Márquez, 2021 para.6). Hard to believe, when they have just released a pair of glasses that can provide results from this research. His ideas about product projection are not far-fetched, as many technology companies have a similar vision regarding the uses of Artificial Intelligence and Augmented Reality: "With the entry of Metaverso, it would seem impossible not to think about such developments" (Márquez, 2021 parr.7).

Ego4D is Facebook's project for AI to see as humans see. Its development belongs to the artificial intelligence department where, in collaboration with 13 universities around the world, have created a

database to teach AI to understand the typical images and photos recorded from first-person devices "The Ego4D project is a powerful exercise in the impact that artificial intelligence can have. Algorithms where the more information we give it and the more we let it into our daily lives, the more accurate the answers it can give us" (Pérez, 14 October 2021 final). Where to draw the line is also an important debate that will need to be addressed.

While it is common for algorithms to work with data sets of videos and photos seen from afar, Facebook wants to anticipate a situation where first-person videos become more common. The problem is that while the AI is able to identify an image of a fair, it does not have such an easy time when the image of the fair is from the perspective of the person viewing it. The same is true for all kinds of situations where the angle is not from afar: "Next-generation AI systems will need to learn from a completely different kind of data: videos that show the world from the centre of the action, rather than on the sidelines" explains Kristen Grauman, a researcher at Facebook (Perez, 14 October 2021 parr.4). According to Gilbert (2021): "Ego4D, which will use data "from 13 universities and labs in 9 countries, which have collected more than 2,200 hours of first-person video in a variety of settings, and drawn from the routines of up to 700 participants" (Gilbert, 2021, para.4). As Facebook explains, this is a twenty-fold increase in the amount of material that was available to help train the algorithms. (Perez, 14 October 2021, para.5).

If we consider how useful it might be for AI to be able to see and hear everything we do in the first person, we discover several utilities: "The Facebook team suggests five possibilities: Episodic memory: asking when something happened. By having a record of our life, the AI can answer questions that in many cases only we know. Prediction: anticipating certain routine steps. For example, in a recipe, the AI can warn us if we have skipped a step. Manipulation of objects: the AI can guide us to perform certain steps, for example, playing an instrument or giving instructions on how to position our body. Audiovisual diaries: by having a record of what we see, it would be possible to ask who said what and when. And so, for example, remember what number the teacher said or what time he/she stayed. Social interaction: by improving first-person understanding, algorithms could help improve vision or sound" (Perez, 14 October 2021, para.6 onwards). Huge potential... at the cost of teaching it our point of view. The possibilities of applying artificial intelligence from the point of view of the user's eyes opens up many possibilities, but also raises many privacy concerns. "Ten days ago, the leak of a gigantic database of Facebook users was revealed; data on 530 million people from all over the world, of which 11 million belonged to people living in Spain. Within this data are of course telephone numbers, dates of birth, places of work and, in some cases, the type of civil relationship" (Castillo, 2021, parr.1).

Ego4D, Facebook's project to make AI see as humans see, its development belongs to the artificial intelligence department where, in collaboration with 13 universities around the world, they have created a database to teach AI to understand the typical images and photos recorded from first-person devices "The Ego4D project is a powerful exercise in the impact that artificial intelligence can have. Algorithms in which the more information we give it and the more we let it into our daily lives, the more accurate the answers it can give us" (Pérez, 14 October 2021 final). Where to draw the line is also an important debate to be addressed.

While it is common for algorithms to work with datasets of videos and photos viewed from afar, Facebook wants to anticipate a situation where first-person videos become more common. The problem is that while the AI is able to identify an image of a fair, it does not have such an easy time when the image of the fair is from the perspective of the person viewing it. The same goes for all kinds of situations where the angle is not from afar: "Next-generation AI systems will have to learn from a completely different kind of data:

videos that show the world from the centre of the action, rather than from the margins" explains Kristen Grauman, a researcher at Facebook (Perez, 14 October 2021 parr.4). According to Gilbert (2021): "Ego4D, which will use data "from 13 universities and labs in 9 countries, which have collected more than 2,200 hours of first-person video in a variety of settings, and drawn from the routines of up to 700 participants" (Gilbert, 2021, para.4). As Facebook explains, this is a twenty-fold increase in the amount of material that was available to help train the algorithms. (Perez, 14 October 2021, para. 5).

Data management

Another intriguing controversy, about data privacy and the ethical consequences of AI, is "why Louis Barclay, a developer at the company who designed a tool that allows users to remove news stories that appear on Facebook, has been permanently banned from the tech giant's platform" (Asher, 2021, parr.1). Barclay had been the creator of a browser extension called Unfollow Everything: "Louis Barclay, a UK-based developer, is the creator of a browser extension called Unfollow Everything. It allowed users to automatically unfollow all their friends and pages on Facebook, leaving their newsfeed blank. Barclay uploaded the Unfollow Everything app to Google Chrome in July 2020. The scientists wanted to study the impact of not having a news feed on how satisfied Facebook users were, as well as the amount of time they spent on the platform. In July of this year, Barclay received a letter from Facebook's lawyers, ousting him from the company" (Asher, 2021, parr.2). Yet another example of how AI interferes with the way we think, as it selectively chooses the people, news and propaganda it wants, impacting the minds of users, without being able to delete and discard such content.

There is also the OnlyFans platform with exclusive sexual content: "OnlyFans is a subscription platform for exclusive and explicit sexual content. In it, there are profiles similar to those of other social networks, such as Facebook, Twitter or Instagram, but with the difference that to 'follow' or subscribe to someone's content you might have to pay". According to Avila (2021) there are free subscriptions: "There are also free subscriptions, from content creators who show their material without having to pay for it, and who generate income through other alternatives as part of their actions, or simply do not profit through the platform" (Avila, 2021 parr.2). "Some media have labelled the service as a mechanism that facilitates virtual prostitution" (Avila, 2021 para.5).

Facebook's algorithm that conditions what is seen on users' home pages has been a common controversy for many years: "People's use (or participation) in the social network is decreasing. They saw that people were hiding 50% more posts" (Bécares, 26 October 2021, para.5). "What did happen is that they browsed more content in Groups, which is where they could find more posts of interest to them. Meaningful social interactions - the comments between friends that Facebook optimises - also dropped by 20%" (Bécares, 26 October 2021, para.6). Other projects were also carried out that were able to conclude how Facebook's algorithms can direct users towards divisive content. "The project also concluded that Facebook's algorithms can lead users down the path of conspiracy theories" (Bécares, 26 October 2021, para.9).

The discrepancy between Facebook's public claims about the effectiveness of its AI and the reality of the user experience has long baffled researchers and other regular users of the platform: "AI has minimal success in removing hate speech, violent images and other problematic content, according to internal company reports" (Wells et al, 18 October 2021 para.1).

On hate speech, the documents show, Facebook employees have estimated that "the company removes only a fraction of posts that violate its rules, a low single-digit percentage, they say. When Facebook's algorithms

are not confident enough that content violates the rules to remove it, the platform shows that material to users less frequently, but the accounts that posted the material go unpunished. Employees were analysing Facebook's success in enforcing its own rules on content that it details internally and in public documents such as its community standards" (Wells et al, 18 October 2021 para.3).

The Wall Street Journal series, based on documents and interviews with current and former employees, describes how "the company's rules favour elites; how its algorithms foster discord; that it has long known that drug cartels and human traffickers openly use its services; and how anti-vaccine activists use Facebook, among other issues. An article on the effects of Instagram on the mental health of teenage girls prompted a Senate hearing in late September. Examples of content that Facebook's AI should have detected but could not be seen include close-up videos of a person shooting someone and videos of car crashes with "visible dismemberment and entrails", according to the documents. “Other violations of Facebook's policies that were leaked through AI included violent threats directed at transgender children" (Wells et al, 18 October 2021 para 22).

In March, another team of Facebook employees came to a similar conclusion "estimating that these systems were removing posts that generated between 3% and 5% of hate speech views on the platform and 0.6% of all content that violated Facebook's anti-violence and violence policies. Incitement" (Wells et al, 18 October 2021 para.9).

In 2016, pop star Selena Gomez flew to Facebook's Menlo Park headquarters to pose for photos with Zuckerberg and Facebook COO Sheryl Sandberg to celebrate her status as the most-followed account on Instagram. Not long after, she was shocked to read a comment from a user on one of her Instagram posts: "Go kill yourself," according to the star's spokesperson, (Wells et al, 18 October 2021 para.44).

Another episode is hate speech in regional elections: In March, staffers preparing for regional elections in India said hate speech was a major risk in Assam, where there is growing violence against Muslims and other ethnic groups. "Assam is of particular concern because we do not have an Assam hate speech classifier," according to a planning document (Notimundo, 2021).

AI must also be trained in foreign languages. The social network's algorithms present problems when it comes to combating violence against minorities or human exploitation. According to a December 2020 memo, "that report also makes public Facebook's inability to address hate speech and harmful content outside the United States. In fact, hate speech and misinformation is substantially worse among non-English speaking users, according to the Facebook Papers. Much of Facebook's moderation infrastructure is not sufficiently resourced to operate in languages other than English, and its software has difficulty understanding, for example, certain dialects of Arabic" (The Debate, 2021 para.7). In that sense, "Facebook's moderation algorithm was only able to identify 0.2% of the harmful material in Afghanistan. The rest of the harmful content was detected by staff, even though the company lacked moderators who spoke Pashto or Dari, the country's main languages" (The Debate, 2021 para. 8).

Globally there are concerns about the use of the platform, "Facebook's artificial intelligence content moderation is unable to read some of the languages used on the platform. Facebook employees have expressed concern about how the system has allowed the platform to be used for nefarious purposes, according to documents seen by The Journal" (Channels, 2021, para.1). A former vice president warns: "A former vice president of the company has explained to the newspaper that Facebook perceives the potential harm in foreign countries as "the cost of doing business" in those markets" (Channels, 2021, paras 7 and 8). Thus, "drug cartels and human traffickers have used Facebook to recruit victims" (Canales, 2021, para.10).

Or the use of Facebook to incite violence: "In Ethiopia, some groups have used Facebook to incite violence against the people of Tigrayan, content that went unnoticed due to the lack of moderators who speak the native language. The company had also failed to translate its community rules into the languages used in that region, according to the Wall Street Journal" (Channels, 2021 para. 13).

A team of more than 350 specialists at Facebook is focused on stopping these organisations and trying to detect risks. Facebook has a huge blacklist full of names of users with criminal and terrorist backgrounds: "Individuals and organisations from all five continents are registered on this list, which Facebook has never wanted to make public for, among other reasons, ensuring the "safety" of its employees. Facebook uses this list to veto content and conversations that can be generated on its platform and that explicitly mention names such as those on the list. Drug cartels, criminal gangs, terrorist organisations and even political parties complete a list with thousands of entries. The list dates back to 2012, when voices began to be heard warning of the risk that these social networks were being used by terrorist organisations to propagandise and recruit new recruits. The list was also key after many experts singled out Facebook for facilitating the assault on Capitol Hill earlier in the year" (Aguilar, 2021, para. 1).

This algorithm affects agencies and advertisers, they cannot rely on the reach of organic traffic because, as we have seen, it practically no longer works, they have a limited budget with which they have to maximise results, unless the Digital Marketing agency is highly specialised in a specific sector, it usually has clients from different sectors and, not knowing in depth all their audiences, they may have added difficulties when it comes to achieving results with some Facebook campaigns.

Facebook's artificial intelligence systems sift through billions of posts looking for items that might match the company's definitions of content that violate its rules. The screening algorithms, called classifiers, are the foundation of the company's content moderation system. Building these classifiers is laborious and complex, requiring an army of humans to flag a large number of posts according to a set of rules. Engineers then take these examples and train their systems to determine the likelihood of other posts violating the rules.

Facebook executives have long said that AI would address the company's chronic problems by keeping what it considers hate speech and excessive violence, as well as underage users off its platforms. However, the problem is not over: "Facebook's algorithm today has little to do with what it was a few years ago. Now the algorithm is far more complex and uses AI - specifically machine learning - to help users connect with other users rather than to help them consume content in isolation" (Nielfa, 2022, -Artificial Intelligence and the Facebook Ads algorithm-).

There are other applications with positive feedback: "Facebook's algorithm can predict up to 96 hours in advance whether a patient is going to show deterioration. Since April last year, Facebook AI has been creating and sharing disease prognostic models to help health experts determine how best to plan and allocate resources in their area. Now, they are open-sourcing all of them for governments and researchers to use" (Hernandez, 2021). Facebook points out that these models are research solutions intended to help hospitals in the coming days and months with resource planning: "A model for monitoring and forecasting covid-19 or improving learning environments for robots are some of the announcements made by the company during the Facebook AI Innovation Summit" (Hernandez, 2021).

A deepfake is a video in which a person's voice and face are changed by artificial intelligence software, making the altered video look authentic: "Facebook has developed an artificial intelligence that wants to identify deepfake images and then track down their creators" (Erard, 2021). This technique is mostly used

with public figures: "Deepfakes abound on the web and some are very difficult to identify. The technology behind these fakes allows for increasingly realistic end results, and that is a real concern. Facebook is one of the many companies working on tools to detect them, and in the last few hours it has unveiled some notable advances. The social network has teamed up with Michigan State University (MSU) to develop a method that not only aims to identify deepfakes, but also to trace their origin. What does this mean? It means that, from the analysis of an image, the technology can determine whether it was artificially generated and detect which generative model was used to produce it" (Erard, 2021 parr.1 and 2).

Good practices...or bad practices?

Positive and negative impacts of Facebook´s algorithms

As discussed in previous sections, Facebook′s use of algorithms has positive and negative impacts.

Journalists have highlighted more the negative ones: the company is an example of bad practice in relation with privacy of users, the fuel of division, the promotion of hate, the spread of fake news, and the profound impact on the emotional and physical health of teenagers.

Literature has also emphasised more the negative impacts: Facebook may seem like a social network on its surface, but its real business is trading in influence through personal data. It is more a personalized advertisement platform than a social medium (Veliz, 2021). There is a report of the british parliament that compares Facebook with a digital ganster (House of Commons, 2019).

Researchers coming from the technological field have also raised concerns. Although machine-learning used by Facebook could have a positive impact in disinformation governance, "there are several challenges that still need to be addressed. Firstly, artificial intelligence tools are not only used to counter disinformation, but they are accelerating the threat, for example by providing systems to generate fake news, images and videos. Secondly, once a detection system has been deployed, malicious actors can modify their behaviour or exploit adversarial attacks to avoid detection. Therefore it is necessary to devise detection systems that are robust to changes in the underlying distribution of the data. Finally, there is a general lack of solutions that effectively combine together multiple modalities (text, video, speech and network analysis), therefore providing more robust and accurate solutions" (Camacho, David[28]. Special Session at the International Joint Conference on Neural Networks (IJCNN) 2022. In conjunction with IEEE World Congress on Computational Intelligence (WCCI) 2022).

A good practice which can be highlighted is in relation with minors: Facebook is not going to use Instagram Kids application, which has been developed recenltly. It is due to concerns related to minor′s rights (https://about.instagram.com/blog/announcements/pausing-instagram-kids).

Facebook had an opportunity to show the world their engagement with people′s rights in the study called Experimental evidence of massive-scale emotional contagion through social networks (Kramer et al., 2014). It was a collaborative Research Ethics endeavor between Facebook and Cornell University's Departments

[28] Camacho´s research group is leading a project called "CIVIC: Intelligent characterisation of the veracity of the information related to COVID-19", which goal is to combine the knowledge of experts in communication and journalism with experts in Artificial Intelligence techniques in order to implement a tool for the general public aimed at characterising automatically information related to COVID-19.

of Communication and Information Science. In it, Facebook researchers directly manipulated Facebook users' news feeds to display differing amounts of positive and negative posts from the people they followed in order to determine whether their subsequent posts were affected by the positivity or negativity of the set of posts they were viewing. This effect, that more positive or negative posts read by a user could change their own emotional state positively or negatively, is the 'emotional contagion' referenced in the article. Facebook allowed the scientists (both internal and those from Cornell) access to the huge amount of data that was produced by manipulating what Facebook users saw according to computerized determination of positivity and negativity levels.

A good practice could have been also the commission Facebook did of a Human Rights Impact Assesment (HRIA) to evaluate its role in the genocide of the Rohingya in Myanmar. The HRIA was found to be largely ineffective at uncovering the human rights harms of Facebook's AI-enabled tools and identifying appropriate mechanisms to mitigate those harms moving forward (Latonero, 2021).

As it is highlighted by the literature (Benesh, 2020) "Facebook, Inc. is quietly running the largest system of censorship the world has ever known, governing more human communication than any government does, or ever has". There is a problem with the human rights impacts of the companies: they can not be held accountable for impacts produced outside their main establishment country. The United Nations treaty on Business and Human Rights is not legally binding yet. For this reason, rights of citizens affected by bad practices can not be enforceable yet.

There are some examples of good practices in some other social networks, as Care2.com (http://www.care2services.com/online-social-action-network) is an online social action network with over 50 million members around the world. Care2 was founded in 1998 with a simple mission: to help make the world a better place. Today, Care2 is a highly-engaged social network of over 40 million citizen activists standing together for good and making extraordinary impact - locally, nationally and internationally - by starting petitions and supporting each other's campaigns. Care2 has been a pioneer of online advocacy since its inception. They provided the first central platform for online petitions, and were the first to help nonprofit organizations tap into this passion to grow their organizations. Care2 has now helped 2,000 nonprofit organizations recruit more than 50 million donor leads.

The Care2 community realizes the power of online petitions in a world where the government can be frustratingly ineffective. Together, they have protected wolves around Denali National Park, won justice for victims of discrimination, saved dogs from euthanasia, helped get GMOs out of Hershey chocolates, stopped an old growth oak grove from destruction, shut down an abusive farm, stopped the killing of rare owls, and much more.

Another good example is Networks For Good (https://networksforgood.com/) whose key function is the user's ability to create their own groups, events and share their experiences and beliefs with others in the business. By doing so, this creates a new level of meaningful engagement and generates positive momentum in delivering a meaningful engagement strategy across an organisation that will secure lasting relationships between employers, employees, members and organisations.

The role of digital rights

Facebook′s algorithms have positive and negative impact on digital rights (positive, in the right to digital health, for example; negative, in the right to privacy and data protection, equality, freedom of expression, free information).

The European Commission unveiled recently two of the most important components of its digital agenda: The Digital Services Act (DSA) and the Digital Markets Act (DMA). The DSA proposal introduces new rules on how online marketplaces and content hosting platforms deal with illegal content, including special transparency and auditing obligations for very large platforms with more than 45 million monthly active users in the EU, a threshold surpassed by several services including Facebook, YouTube, Twitter and TikTok. The Commission did not force platforms to monitor and censor what users say or upload online.

The proposal for a Regulation (DSA) takes over the same two principles as the e-commerce Directive, namely that they are not liable for illegal content that they host or transmit, as long as they do not have actual knowledge of it, and that there is no general obligation to monitor to prevent the publication or transmission of such content. It therefore maintains the general legal framework of liability of the Directive, but nevertheless introduces new obligations for hosting service providers and online platforms: the creation of specific processes for requesting the removal of illegal content, the introduction of mechanisms for users whose content has been removed to defend themselves on the grounds of their right to freedom of expression and information, the obligation to cooperate with the competent authorities of the Member States in the process of removing illegal content and identifying certain users, the obligation to cooperate with the competent authorities of the Member States in the process of removing illegal content and identifying certain users, and the obligation to ensure that users are able to exercise their right to freedom of expression and information (Pérez de las Heras, 2022).

The DSA includes a number of due diligence obligations: some affect all intermediary service providers and include, among others: the establishment of a single point of contact for direct communication, the designation of a legal representative in the Union in the case of providers not established in the Union but providing services on European territory, information in their general terms and conditions on the restrictions imposed by the intermediary on the use of their services, including the content moderation policy and the use of algorithmic decision-making systems (Art. 10-12).

The DSA recognizes a second set of additional obligations specifically addressed to online platforms, unless they are small and medium-sized enterprises. These specific obligations include the establishment of internal complaint systems to handle the possible removal of illegal content, cooperation with out-of-court dispute resolution services, preferential treatment of takedown notices of illegal content handled by so-called "trusted reporters" (introduced in Art. 19 of the proposal), suspension of the use of the content of the platform (Art. 19 of the proposal) and the suspension of the use of the content of the platform (Art. 19 of the proposal), the temporary suspension of services to recipients with a continuous history of infringements, the reporting of activities suspected of constituting serious offences, the collection of information necessary to enable the traceability of business customers offering products or services on the platform and the proper identification of advertising displayed on its interfaces, including information on the criteria used to select advertising recipients (arts. 14-24). For large platforms, defined as those with at least 45 million active monthly users in the EU, additional obligations are added to identify potential risks and, where appropriate, mitigate them. These obligations include the preparation of a risk analysis to determine the risks related to the distribution of illegal content, the introduction of mitigating measures for the risks identified, the carrying out of independent audits once a year to monitor compliance, the description of the parameters used in their recommendation systems and the obligation to designate a compliance officer responsible for supervising compliance with the established tasks (arts. 25-33), (Pérez de las Heras, 2022).

The proposed Regulation provides also for a regime of public guardianship of these obligations. The monitoring and protection scheme is inspired by the GDPR system itself. Thus, a Digital Services Coordinator responsible for compliance with the future Regulation in each Member State and for processing user complaints is envisaged, as well as the creation of a European Digital Services Committee in which the competent authorities of each Member State will participate. Finally, a system of financial penalties is also envisaged in the event of non-compliance with the Regulation, although the Member States are given discretion to impose them within the limits established by the proposal, which can reach up to 6% of the service provider's annual revenue (Articles 40-49).

European lawmakers have the chance, with the Digital Services Act, to ensure that public interest researchers, including academia, journalists, and civil society organizations, have access to the data needed from large platforms.

Concerning impacts of Facebook algorithms on human rights, the Proposal of a regulation of the european Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) is also an important step. It is a risk-based framework and AI applications are categorised by unacceptable risk, high risk, limited risk and minimal risk. The facial recognition system would be high risk. Many other AI applications used by Facebook would be limited or minimal risk. These means less duties for the platform.

Another interesting regulation is the recent proposal of a European Directive on due diligence. The European Parliament has a Report with recommendations to the Commission on corporate due diligence and corporate accountability (2020/2129(INL)). The draft directive proposes that EU companies with more than 500 employees and a turnover of 150 million euros should be obliged to put in place specific measures to prevent human rights and environmental abuses along their supply chains, through so-called "due diligence". The draft proposes that in sectors with a higher risk of exploitation, such as agriculture and the fashion industry, it would apply to companies with more than 250 employees and a turnover of 40 million, while SMEs would be exempted. Companies from non-EU countries operating in the single market and above these thresholds would also be covered (Facebook could be an example). Under the Commission's proposal, companies could be held liable for damages committed by their subsidiaries, subcontractors and/or suppliers inside or outside their borders; and victims of these rights violations could file complaints in EU courts. This is an important step in establishing the right to remedy for victims of corporate malpractice.

Conclusions

Machine learning algorithms must put people first. Algorithms dominate our lives, and while some change our lives for the better, we are beginning to realise that letting a machine make decisions may not be such a good idea. Algorithms are beginning to dominate our reality: they do so when they recommend a song on Spotify, a series on Netflix or a purchase on Amazon, but also when they are in charge of choosing the best candidates for a job or granting a loan.

That makes for a bittersweet feeling about these algorithms that make our lives easier but also condition them: " we should be able to audit these algorithms to take back control of both those algorithms and our lives" (Pastor, 15 February 2017 paras. 1 and 2). Those who program these algorithms are humans with their own experiences and opinions that to a greater or lesser extent can end up affecting these systems: "This

means that we end up with algorithms that are biased, sexist, conditioned, subjective and as unfair as we ourselves are" (Pastor, 13 April 2018 para.25).

Humans are good at distinguishing which content can hurt sensibilities, but machines still have a lot of trouble differentiating between hate speech, race, sex, politics, etc. This is one of the great challenges of artificial intelligence. The problem of detecting such content and comments has not been solved perfectly by Facebook's artificial intelligence systems, but Zuckerberg is optimistic about the future: "Humans are good at distinguishing what content can be hurtful, but machines still have a lot of trouble. Hate speech is one of the hardest problems to tackle, but I think artificial intelligence can get us there in five to ten years" (Mollejo, 2018, para.4). Zuckerberg insists on the use of artificial intelligence to "police" Facebook content. This would make it possible to differentiate directly and in real time when such content or comments are part of a debate and are acceptable, or when there is aggression involved and should be moderated and censored.

It is not entirely clear whether this is the system they use on Facebook to try to moderate and censor inappropriate content, but what is certain is that the system is used on Instagram: "Another problem in the air, that of DeepText and other similar systems is that of false positives: blocking comments and content that were not toxic" (Pastor, 13 April 2018 parr.19). It would classify the comment in special categories such as 'prohibited behaviour', 'racism', 'sexual harassment' or 'bullying'. "Among the main risks is that of false positives: the problem is that a word that has certain meanings can be understood differently depending on the context or the passage of time" (Pastor, 13 April 2018 parr.14). Google has also been working on this problem for some time. Therefore, we find light and shadows of AI as a moderator and content cleaner: "The problem is even more worrying when we see how those who most use artificial intelligence to filter, moderate or censor do so in a dangerous way: without apparent control" (Pastor, 13 April 2018 parr.14).

The role of the human being in ensuring that such content does not become toxic is more important than ever. Mark Zuckerberg's social network has millions of users and thousands of posts per day. But there are those who break ethics and decide to spread false news, drug advertisements, harassment, racism. One of the people who fights for this "justice" on the social network is Joaquín Quiñonero, director of applied machine learning at Facebook. His 'weapon': artificial intelligence, the option for the future. Quiñonero came to Facebook after working at Microsoft" (Fiter, and Encabo, 18 November 2019, para.1). The latest data published by Facebook highlights how this technology is revolutionising its fight against this serious problem: "in recent months they have removed nearly 3.2 billion fake accounts and more than 11 million posts that incited hatred". The expert wanted to represent what kind of frauds the algorithms are up against (Fiter, and Encabo, 18 November 2019, parr.3). Artificial intelligence is also behind the identification of people appearing in photographs or videos.

This contribution concludes that platforms' efforts to combat disinformation through AI are not enough. The international regulation framework (specially the European Union framework) is an opportunity to help platforms moderating contents.

The impacts of machine learning algorithms on ethical values and on human rights are many, both positive and negative. The second ones need more effort from different stakeholders, such as states, private sector, civil organizations and intergovernmental organizations.

References

20Bits. 5 October 2021 "WhatsApp, Facebook and Instagram suffer a global downtime of more than six hours". https://www.20minutos.es/tecnologia/actualidad/caida-mundial-de-whatsapp-facebook-e-instagram-no-permite-usar-las-apps-4844020

20Bits. 7 November 2021 "A Facebook study reveals that 360 million of its users are compulsive addicts to this social network". https://www.20minutos.es/tecnologia/estudio-facebook-revela-millones-usuarios-adictos-compulsivos-esta-red-social-4882477/

Aguilar, Albert R. 14 October 2021 "Filtran la -lista negra- de nombres de Facebook: estos son los -terroristas o criminales- que están vetados en la red social". https://www.businessinsider.es/lista-negra-facebook-llena-terroristas-criminales-948005

Altmann, Gerd. 14 May 2021 "Facebook brings artificial intelligence closer to the human brain and teaches it to forget" PORTAL TIC EUROPA PRESS. https://www.europapress.es/portaltic/sector/noticia-facebook-acerca-inteligencia-artificial-cerebro-humano-le-ensena-olvidar-20210514150033.html

Scope. Technology. 19 October 2021 "What is Ego4D, Facebook's controversial new project" https://www.ambito.com/tecnologia/facebook/que-es-ego4d-el-nuevo-y-controversial-proyecto-n5301603

Antoñanzas, Diego. 2021 "Facebook uses artificial intelligence to guess love relationships" https://diegoantonanzas.com/facebook-usa-inteligencia-artificial-para-adivinar-relaciones-amorosas/

Araújo, Santi. 10 July 2020 "An internal audit warns Facebook that its inaction on hate speech is "a step backwards for civil rights"" https://www.genbeta.com/redes-sociales-y-comunidades/auditoria-interna-advierte-a-facebook-que-su-inaccion-discurso-odio-paso-atras-para-derechos-civiles

Araújo, Santi. 25 September 2020 "Social media makes us sick : former Facebook exec accuses the platform of being intentionally as addictive as tobacco" https://www.genbeta.com/redes-sociales-y-comunidades/redes-sociales-nos-enferman-ex-ejecutivo-facebook-acusa-a-plataforma-ser-intencionalmente-adictiva-como-tabaco

AP, 25 October 2021 "Apple threatened to pull Facebook app", 20Minutos Actualidad. https://www.20minutos.com/noticia/327976/0/apple-amenazo-con-retirar-la-aplicacion-de-facebook

Asher Hamilton, Isobel. 30 October 2021 " Facebook bans developer for good after he created an app that allowed users to delete their news feed" https://www.msn.com/es-es/noticias/tecnologia/facebook-expulsa-para-siempre-a-un-desarrollador-despu%C3%A9s-de-que-creara-una-aplicaci%C3%B3n-que-permit%C3%ADa-a-los-usuarios-eliminar-su-feed-de-noticias/ar-AAPqvj2?getstaticpage=true&automatedTracking=staticview&parent-title=los-jeans-con-flores-bordadas-que-te-enamorar%C3%A1n-y-puedes-hacer-t%C3%BA-misma&parent-ns=ar&parent-content-id=BBsRwzk

Avila, Lucia. 24 August 2021. "OnlyFans, a dangerous social network: What is it and how does this adult content platform work?" https://www.ondacero.es/noticias/sociedad/onlyfans-peligrosa-red-social-que-como-funciona-esta-plataforma-contenido-adulto_202108246124f3fb1827770001a3cfb0.html

BBC News World. 3 February 2022. Updated 4 February 2022 "Facebook: Meta's unprecedented plunge on the stock market after the social network's first drop in active users" https://www.bbc.com/mundo/noticias-60244251

Bécares, Barbara. 11 August 2021. "Facebook says it's looking at how to continue its big ad business, but knowing less data about you." https://www.genbeta.com/actualidad/facebook-dice-que-esta-estudiando-como-seguir-su-gran-engocio-publicidad-sabiendo-datos-ti

Bécares, Barbara. 14 April 2021. "European Union, its first regulation to regulate the use of artificial intelligence" DevelopWordNews. https://www.genbeta.com/actualidad/union-europea-sacara-su-primer-reglamento-para-regular-uso-inteligencia-artificial-adelanto-que-plantea

Bécares, Bárbara. 20 October 2021 "Vienna Museums show their nude art on OnlyFans to avoid being censored by Facebook." https://www.genbeta.com/actualidad/museos-viena-muestran-su-arte-desnundos-onlyfans-para-evitar-ser-censurados-facebook

Bécares, Barbara. 26 October 2021 "The Facebook Papers lay the company bare and reveal everything from human trafficking to why Instagram is losing young people" https://www.genbeta.com/actualidad/facebook-papers-dejan-compania-al-desnudo-desvelan-trafico-personas-que-instagram-pierde-a-jovenes

Bécares, Bárbara. 29 January 2021 "Tim Cook criticises Facebook and others for not respecting privacy, and Apple will launch a feature to prevent tracking" https://www.genbeta.com/actualidad/tim-cook-critica-que-facebook-otras-no-respeten-privacidad-apple-lanzara-funcion-para-impedir-su-rastreo

Bécares, Bárbara. 3 May 2021 "Facebook claims it needs to track us to make Instagram free, and the question is who would pay €2 a month if they made it paid" https://www.genbeta.com/actualidad/facebook-reconoce-que-necesita-rastrearnos-instagram-sea-gratis-pregunta-quien-pagaria-2eur-al-mes-hacen-pago#:~:text=This%20will%20translate%20to%202,the%20same%20as%20it%20has%20been%20until%20now

Bécares, Barbara. 4 October 2021, "The former Facebook employee who leaked its internal documents: they allow hate and disinformation, they just want to make money". https://www.genbeta.com/actualidad/exempleada-facebook-que-ha-filtrado-sus-documentos-internos-permiten-odio-desinformacion-solo-quieren-ganar-dinero

Benesch, S. 2020. "But Facebook's Not a Country: How to Interpret Human Rights Law for Social Media Companies. Yale Law School. https://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1004&context=jregonline

Camacho, David. Special Session at the International Joint Conference on Neural Networks (IJCNN) 2022. In conjunction with IEEE World Congress on Computational Intelligence (WCCI) 2022. https://wcci2022.org/

Channels, Katie. Business Insider 17 September 2021 "Facebook's AI moderation is unable to interpret all languages, leaving users in some countries more exposed to harmful content." https://www.businessinsider.es/facebook-tiene-problemas-moderar-contenidos-no-sean-ingles-933247

Castillo, Alba. 15 April 2021 "Facebook, investigated by the European Union for the leak of millions of its users' phone numbers". https://www.20minutos.es/noticia/4658218/0/facebook-investigada-por-la-union-europea-por-la-filtracion-de-millones-de-numeros-de-telefono-de-sus-usuarios/

Clarín, 19 September 2021. "Apple almost removed Facebook from the App Store: human traffickers used the social network to sell victims". The Vanguard https://www.lavanguardia.com/tecnologia/20210919/7731847/apple-estuvo-punto-eliminar-facebook-app-store-traficantes-personas-usaban-red-social-vender-victimas-pmv.html

Delgado, Monica. 10 November 21 "Graham Mudd removes from Facebook, advertising linked to sensitive issues" Consumer tick. https://www.consumotic.mx/tecnologia/elimina-facebook-publicidad-ligada-a-temas-sensibles/

The Debate. 25 October 2021 "Facebook ignored reports warning of hate speech and human trafficking" https://www.eldebate.com/tecnologia/20211025/facebook-ignoro-informes-alertaban-sobre-difusion-mensajes-odio-trata-personas.html

Erad, Gabriel. 6 September 2021 "Facebook labels black people as primates" Ipertextuai and the AI. https://hipertextual.com/2021/09/facebook-discriminacion-inteligencia-artificial

Erard, Gabriel. 5 September 2021 "Facebook wants the artificial intelligence of the future to be able to "forget" irrelevant data." https://hipertextual.com/2021/05/facebook-inteligencia-artificial-datos-irrelevantes

Erard, Gabriel. 16 June 2021 "Facebook intends to rely on reverse engineering to detect 'deepfakes' and trace their origin" https://hipertextual.com/2021/06/facebook-detecta-origen-deepfake

Erard, Gabriel. 6 September 2021 "Facebook AI labels black people as "primates"" https://hipertextual.com/2021/09/facebook-discriminacion-inteligencia-artificial

Fernández de Lara, Carlos. 4 June 2021 "Facebook to uphold Trump veto for 2 years; then revisit its case." https://www.forbes.com.mx/facebook-mantendra-veto-a-trump-por-2-anos-luego-volvera-a-revisar-su-caso/

Fiter, Miguel and Encabo, Ignacio. 18 November 2019. International Artificial Intelligence Congress: "this is how Facebook is working with artificial intelligence to identify hate" The Independent, in science and technology. https://www.elindependiente.com/futuro/2019/11/18/asi-trabaja-facebook-con-la-inteligencia-artificial-para-identificar-al-odio/

Fjeld, Jessica and Achten, Nele and Hilligoss, Hannah and Nagy, Adam and Srikumar, Madhulika, Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI (January 15, 2020). Berkman Klein Center Research Publication No. 2020-1, Available at SSRN: https://ssrn.com/abstract=3518482 or http://dx.doi.org/10.2139/ssrn.3518482

Floridi, L (2019), «Translating principles into Practices of Digital Ethics: Five Risks of Being Unethical». *Philosophy and Technology*, 32.

Gascón, Marta. 28 October 2021 "What are the 'Facebook Papers': keys to understanding the new scandal that Zuckerberg's social network is facing" https://www.20minutos.es/tecnologia/actualidad/que-son-los-facebook-papers-claves-para-entender-el-nuevo-escandalo-al-que-se-enfrenta-la-red-social-de-zuckerberg-4868034/

Gascón, Marta. 28 October 2021. "Instagram Kids: what is this social network project for children and why Facebook has now paralysed it". https://www.20minutos.es/tecnologia/ciberseguridad/instagram-kids-que-es-este-proyecto-de-red-social-para-ninos-y-por-que-ahora-facebook-lo-ha-paralizado-4835999/

Genbeta. 5 October 2021 "The ex-Facebook employee who leaked its internal documents: they enable hate and misinformation, they just want to make money."

Gilbert, Ben. 15 October 2021. "Facebook prepares an artificial intelligence that will track your every move." Business insider. https://www.businessinsider.es/facebook-prepara-ia-rastrear-todos-movimientos-948387

Gómez Maseri, Sergio. 10 October 2021, "Facebook: the story of a growing scandal". https://www.eltiempo.com/mundo/eeuu-y-canada/facebook-la-historia-de-un-escandalo-cada-vez-mayor-624207

González, Gabriela. 28 October 2021, "Facebook's Metaverse is a virtual world ideal for spending money and which there is no way to enter yet" https://www.genbeta.com/actualidad/metaverso-facebook-mundo-virtual-ideal-para-gastar-dinero-al-que-no-hay-forma-entrar-todavia

Gonzalez, Gabriela. 8 April 2016 Last updated, 11 March 2021 "85 out of every 100 adults in the world use a Facebook-related account" https://hipertextual.com/2016/04/uso-de-facebook-en-el-mundo

Hernandez, Noelia. 1 July 2021 "Facebook puts its open-source artificial intelligence at the service of science" https://www.elespanol.com/invertia/disruptores-innovadores/innovadores/tecnologicas/20210701/facebook-inteligencia-artificial-codigo-abierto-servicio-ciencia/593191230_0.html

House of Commons 2019 Digital, Culture, Media and Sport Committee, Disinformation and Fake news: final report.

The Nation. 25 October 2021 "Apple threatened to pull Facebook app" https://www.lanacion.com.ar/agencias/apple-amenazo-con-retirar-la-aplicacion-de-facebook-nid25102021/

Jobin, Anna. Principles in Ethics Guidelines & Policy Perspectives Summer School: ARTIFICIAL INTELLIGENCE, ETHICS & HUMAN RIGHTS Universidad Politécnica Madrid, 22 June 2022 Dr. Anna JOBIN

La voz de Galicia 5 October 2021 "Mark Zuckerberg apologises for the fall of Facebook, WhatsApp and Instagram" https://www.lavozdegalicia.es/noticia/sociedad/2021/10/05/mark-zuckerberg-pide-perdon-caida-facebook-whatsapp-instagram/00031633435695869723899.htm#:~:text=%C2%ABFacebook%2C%20Instagram%2C%20WhatsApp%20and,en%20his%20popular%20social%20network%20

Lacort, Javier. 27 September 2018. Updated 15 November 2018 "This is how Facebook makes money: the other advertising giant has two aces up its sleeve." https://www.xataka.com/empresas-y-economia/asi-como-gana-dinero-facebook-otro-gigante-publicitario-tiene-dos-ases-manga

Latonero, Mark and Agarwal, Aaina. Human Rights Impact Assessments for AI: Learning from Facebook's Failure in Myanmar, Carr Center for Human Rights Policy, 2021, https://carrcenter.hks.harvard.edu/files/cchr/files/210318-facebook-failure-in-myanmar.pdf

Lopez, Miguel. 2 November 2021 "Meta will delete the facial recognition data of more than a billion people used on Facebook: more knockout blows from Zuckerberg" https://www.xataka.com/privacidad/meta-borrara-datos-reconocimiento-facial-mil-millones-personas-usados-facebook-golpes-efectos-zuckerberg?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+xataka2+%28Xataka%29

McKay, Tom. 8 January 2017 "What really happened to that artificial intelligence that Facebook shut down because it had created its own language". https://es.gizmodo.com/lo-que-realmente-sucedio-con-aquella-ia-que-facebook-ap-1797424875

Merino, Marco. 20 Oct. 2021 "Mark Zuckerberg, indicted in his personal capacity in Cambridge Analytica case: a conviction would expose him to heavy penalties."

Miguel Trula, Esther. 29 July 2021 "Facebook thinks the networks of the future will look like Ready Player One. And that's why it's creating the Metaverse" https://magnet.xataka.com/en-diez-minutos/facebook-cree-que-redes-futuro-se-pareceran-a-ready-player-one-eso-esta-gestando-metaverso

Mollejo, Veronica. 15 May 2018 "Facebook turns to artificial intelligence to fight hate" https://www.redbull.com/es-es/tecnologia-facebook-inteligencia-artificial-odio

Nielfa, Jaime S. 1January 2022 "7 Artificial Intelligence Tools to Optimise Facebook Ads Campaigns and The Fall of Organic Traffic on Facebook Ads" https://scoreapps.com/blog/es/facebook-ads-herramientas-ia/

Notimundo. November 2021. Facebook was concerned about hate speech violations in Assam ahead of the 2021 elections.

https://noticiasdelmundo.news/facebook-estaba-preocupado-por-las-violaciones-del-discurso-de-odio-en-assam-antes-de-las-elecciones-de-2021/

Olaskoaga, Andrés. 5 January 2021 "Facebook, experiments with artificial intelligence that listens, sees and remembers your personal information" Muy interesante, Ciencia y tecnología. https://www.muyinteresante.com.mx/ciencia-tecnologia/facebook-inteligencia-artificial-recuerda-informacion-personal/amp/

Pastor, Javier 13 April 2018. "Mark Zuckerberg: artificial intelligence will censor, hate messages, before they are published". https://www.xataka.com/robotica-e-ia/zuckerberg-y-el-papel-de-la-inteligencia-artificial-como-moderadora-y-censuradora-de-contenidos

Pastor, Javier. 15 February 2017 "Los algoritmos que control nuestros vidas son tan injustos como nosotros mismos" https://www.xataka.com/privacidad/los-algoritmos-que-controlan-nuestras-vidas-son-tan-injustos-como-nosotros-mismos

Perez, Enrique. 14 October 2021 "The next (and somewhat disturbing) challenge for Facebook's AI is to understand and remember the world as seen through our eyes". https://www.xataka.com/privacidad/siguiente-algo-perturbador-desafio-ia-facebook-entender-recordar-mundo-visto-a-traves-nuestros-ojos

Perez, Enrique. 27 September. 2021 "Facebook halts creation of 'Instagram Kids' for under-13s after controversy with toxicity report." https://www.xataka.com/aplicaciones/facebook-detiene-creacion-instagram-kids-para-menores-13-anos-polemica-informe-toxicidad

Perez, Enrique. 4 October 2021, Updated 8 October 2021 "Facebook is going through one of its most delicate moments after weeks of leaks: everything that has come to light so far". https://www.xataka.com/empresas-y-economia/facebook-atraviesa-uno-sus-momentos-delicados-semanas-filtraciones-todo-que-ha-salido-a-luz-ahora

Pérez de las Heras, Beatriz (2022). Los derechos digitales en la Unión Europea: del liberalismo económico a la protección jurídica. Revista General de Derecho Europeo, ISSN-e 1696-9634, Nº. 57.

Rodriguez, Pablo. 14 December 2021 "Why no scandal will topple Zuckerberg at Meta: more than 200 shareholders have been trying to disempower him since 2018 without success" https://www.xataka.com/empresas-y-economia/que-no-hay-escandalo-que-tumbe-a-zuckerberg-meta-200-accionistas-tratan-restarle-poder-2018-exito#:~:text=No%20hay%20esc%C3%A1ndalo%20de%20Facebook,acumulado%20en%20todos%20estos%20a%20a%C3%B1os.

Rus, Chistian 9 December 2020 "US sues Facebook for anti-competitive practices: asks it to divest itself of Instagram and WhatsApp" https://www.xataka.com/empresas-y-economia/estados-unidos-demanda-a-facebook-practicas-anticompetitivas-pide-que-se-desvincule-instagram-whatsapp

Rus, Chistian. 13 August 2021 "researchers have created a "master face" capable of fooling facial recognition systems"

Rus, Chistian. 14 June 2021 "Facebook says its artificial intelligence can mimic your handwriting just by looking at a word" https://www.xataka.com/robotica-e-ia/facebook-dice-que-su-nueva-inteligencia-artificial-puede-imitar-tu-escritura-solo-ver-palabra

Véliz, Carissa 2021 Privacy is Power: Why and How you should take back control of your data. Ed. Melville House.

Week. Technology. 14 October 2021 "Facebook adjusts its policies to act against mass attacks and sexual harassment" https://www.semana.com/tecnologia/articulo/facebook-ajusta-sus-politicas-para-actuar-contra-los-ataques-masivos-y-el-acoso-sexual/202110/

Well, Georgia; Seetharaman, Deepa and Horwitz, Jeff .14 September 2021 "Revealing documents: Facebook knew Instagram is toxic for teenage girls" The Wall Street Journal. https://www.lanacion.com.ar/el-mundo/documentos-reveladores-facebook-sabia-que-instagram-es-toxico-para-las-adolescentes-nid14092021/

Wells, Georgia; Seetharaman, Deepa and Horwitz, Jeff. 18 October 2021 "Facebook says AI will clean up the platform. Its own engineers have doubts" https://www.latercera.com/que-pasa/noticia/facebook-dice-que-la-ia-limpiara-la-plataforma-sus-propios-ingenieros-tienen-dudas/J3EAKYOGPJDPJNHILN3G2BQVM4/

Part II: Hate Speech and Discrimination

Discrimination on online platforms: legal framework, liability regime and best practices

Laurena Kalaja and Lana Bubalo
GLOBAL DIGITAL HUMAN RIGHTS NETWORK | UNIVERSITY OF STAVANGER

Introduction

The availability of the internet has had huge importance and significant effects on all aspects of people's lives. It has brought many opportunities, connecting the world as never before.[29] At the same time, it has resulted in new forms of human rights infringements, including the right not to be discriminated against.

Online discrimination is understood as denigrating or excluding individuals or groups on the basis of race, gender, sex orientation, age, disability, religion and beliefs through the use of symbols, voice, video, images, text, and graphic representation or the combination thereof, on the internet.

The discrimination on the internet can be direct - such in case of algorithmic bias or discrimination by design. Algorithmic discrimination includes biases incorporated into algorithms and codes that power machine learning and artificial intelligence systems resulting in systematic disadvantage of certain groups or people.[30] This kind of discrimination is becoming increasingly common and requires regulation. Algorithmic bias can have very serious consequences, leading to unfair exclusion of a specific demographic or otherwise target group, for example in employment or housing. For example, Facebook offered "ethnic affinities" as a category which advertisers could use to target their campaigns. After a lot of negative public attention, Facebook discontinued using this designation.[31]

Discrimination by design means that the way the website is made, enables discrimination. For instance, Airbnb's main service requires each guest to create an online profile with certain information, including a genuine name and phone number. It also encourages inclusion of a real photograph. For Airbnb, the authenticity of this profile information is vital to the operation of the service, as it engenders a sense of trust and connection between hosts and guests. Guests' physical characteristics may contain social cues that instill either familiarity and comfort, on the one hand, or suspicion and distrust, on the other. The sense of authentic connection that Airbnb is adamant about cultivating, however, has dangerous consequences in a market long plagued by discrimination against racial and ethnic minorities.[32] After massive criticism, the Airbnb has taken measures to fight against discrimination. As of October 2018, rather than displaying a potential guest's profile photo before the booking is accepted, hosts now receive a guest's photo only after they've accepted the booking request. Additionally, Airbnb hosts explicitly agree to a standard and to

[29] The Select Committee on Communications, Regulating in a digital world, HL Paper 299, 2019. p. 3.

[30] See Lopez, P, Bias does not equal bias: a socio-technical typology of bias in data-based algorithmic systems, Internet policy review, vol. 10, issue 4, 2021.

[31] Facebook Lets Advertisers Exclude Users by Race — ProPublica, last accesses 8.5.2022.

[32] Fair Hous. Council of San Fernando Valley v. Roommates.com, LLC, 521 F.3d 1157 (9th Cir. 2008)

adhere to a nondiscrimination policy that goes beyond what is required by law, in most jurisdictions. Additionally, specially trained teams have been brought in to handle discrimination complaints and enforcement.[33]

Another example of discrimination by design is the site Roommates.com which requires subscribers to express preferences in a dropdown menu that lists gender, sexual orientation, and family status as potential options. A participant had to share such a preference to find a match. In the US, the Fair Housing Act forbids advertising housing with any preference for race, sex or family status. The case was brought to trial by Fair housing groups, who accused Roomates.com of facilitating discrimination. A district court agreed, barring the website from soliciting information on users' sex, sexual orientation or family status. However, the appeal court overturned the decision and found that it would be a serious invasion of privacy, autonomy and security "to prevent people from choosing roommates with compatible lifestyles".[34]

Besides the discrimination by the platforms themselves, discrimination can be indirect- when internet users discriminate other internet users online using intermediaries (such as online platforms[35]). Such discrimination can take many different forms. It can occur in social networking sites, chat rooms, discussion boards, through text messaging, web pages, online videos, music, and online games. This issue opens difficult question on the scope of private authority and public regulation. Should the responsibilities of private tech companies derive from human rights law, terms of service, contracts or something else?[36]

Whilst algorithmic discrimination poses a challenge as it is often "hidden"- meaning that the users are often not aware they are being discriminated against- the discrimination users post on online platforms is more prominent and "visible". As platforms such as Facebook, Instagram and YouTube are not traditional media publishers with editorial control, there is uncertainty about whether they should bear liability for the discriminatory conduct and comments their users post online. It is however uncontested that they do have great power to control the information available to the online users.[37]

Online discrimination may resemble traditional discrimination,[38] but it can have more serious consequences, as the internet plays an essential role in our lives, shaping our conception of the world, our opinions and our values.

Online discrimination, just like any other discrimination is often motivated by hate or prejudice, and it is sometimes not possible to distinguish (online) discrimination and (online) hate speech.[39] Hate

[33] Airbnb Works To Clean Up Its Reputation For Racial Discrimination In New 3-Year Report - Essence last accessed 6.5.2022.

[34] Roommate-matching site does not violate housing laws, court | Reuters last accessed 6.5.2022.

[35] OECD (2019), What is an "online platform?, in An Introduction to Online Platforms and Their Role in the Digital Transformation, OECD Publishing, Paris, 2019, https://doi.org/10.1787/19e6a0f0-en ,last accessed 28.04.2022.

[36] Human Rights Council, Report of the special Rapporteur on the promotion and protection of the freedom of opinion and expression, 2016, p .3.

[37] Levy, K, Barocas, S, Designing Against Discrimination in Online Markets Berkeley Technology Law Journal, vol. 32:1183, 2017.

[38] Gaylord-Harden NK, Cunningham JA, The impact of racial discrimination and coping strategies on internalizing symptoms in African American youth. Journal of Youth and Adolescence. 2009;38(4):532–543. doi: 10.1007/s10964-008-9377-5; Umana-Taylor AJ, Wong JJ, Gonzales NA, Dumka LE. Ethnic identity and gender as moderators of the association between discrimination and academic adjustment among Mexican-origin adolescents. Journal of Adolescence. 2012;35(4): 773–786. doi: 10.1016/j.adolescence.2011.11.003 .

[39] The Equality and Anti-Discrimination Ombud's Report: Hate Speech and Hate Crime, 2015, available at: Hatytringer og hatkriminalitet_Engelsk.indd (ldo.no), last accessed 30.3.2022, p. 6.

speech covers many forms of expressions which advocate, incite, promote, or justify hatred, violence and discrimination against a person or group of persons for a variety of reasons.[40]

Internet has no borders, and discrimination online is a global issue that needs to be addressed at international, regional, and national level. The question is whether the current legislation is sufficient to provide for protection against the discrimination in the online setting. This study therefore aims, *inter alia*, at mapping out the existing legal solutions, accompanying policy measures, community standards and good practice to address and redress discrimination on online platforms.

By analysing existing regulations and community standards, we intend to evaluate which measures are best suited for preventing and redressing online discrimination, and we will eventually provide examples of best practices.

The paper starts with an overview of principle of non-discrimination in international and regional (European) documents. It then looks at the question of liability of social media platforms for the content posted by third parties. Outsourcing the task of defining the infringement and balancing human rights to private companies (privatization of justice) results in several complex legal issues.

The analysis ends with the conclusion on best practices which can contribute to achieve a goal of making the internet a better, more respectful environment where the infringements of human rights are minimized, and vulnerable groups are not exposed to discriminatory practices.

Regulation of online discrimination

There is no doubt the internet is going from the space of freedom to becoming more regulated. The result is that there is more accountability for the users and providers of content, as well as online platforms. The speed of technological change and its transnational character however make the regulation of digital world a challenging task.[41]

Any interference by the state with freedom of expression and information must comply with the rule of law and meet the strict criteria laid down in international human rights law; it must be prescribed by law, pursue a legitimate aim and be proportionate.[42] However, as argued by Jørgensen and Anja Møller Pedersen, the current regulatory schemes are insufficient to provide the standards and compliance mechanisms required to meet these standards.[43] In addition, as it will be shown in the following, most of the sources of law deal with discrimination in general, and not particularly online discrimination, making the legal framework quite fragmented and complex, as the rules on intermediary liability come into play. This eventually leads to legal uncertainty for victims of online discrimination, as well as for the social media platforms which need to comply with the rules. Complex and fragmented laws can increase operational costs, potentially leading them to simplify by being too restrictive.

[40] European Commission against Racism and Intolerance, Hate speech and violence (coe.int), last accessed 30.3.2022.

[41] The Equality and Anti-Discrimination Ombud's Report: Hate Speech and Hate Crime, Hatytringer og hatkriminalitet_Engelsk.indd (ldo.no), last accessed 31.3.2021, p. 7.

[42] Council of Europe, Recommendation CM/Rec(2014)6 of the Committeee of Ministers to member States on a guide to human rights for Internet users - Explanatory Memorandum. Strasbourg: Council of Europe, para. 47.

[43] Jørgensen and Pedersen, p. 2.

International legal framework

The UN Framework recognizes that States have the duty under international human rights law to protect everyone within their territory and/or jurisdiction from human rights abuses. This includes both the positive and negative obligations of the state. Besides not infringing the citizens' rights themselves, States have a duty to have effective laws and regulations in place to prevent and address human rights abuses and ensure access to effective remedy for those whose rights have been abused.[44] Several international documents, such as the UN Charter[45] and the Universal Declaration of Human rights,[46] prohibit discrimination. According to these documents, human rights are universal – to be enjoyed by all people, no matter who they are or where they live.

In addition, two international covenants from 1966 - Covenant on Civil and Political Rights (ICCPR), and the Covenant on Economic, Social, and Cultural Rights (ICESCR) contain general and specific non-discrimination clauses. The principal clause on non-discrimination is found in Article 26 of the ICCPR. It provides an autonomous right of equality and prohibits discrimination in law or in fact in any field regulated and protected by public authorities.[47] This Convention also on art 19(3) allows for restrictions on the freedom of expression, when these are useful, reasonable and desirable.[48]

Other conventions are more specific in terms of discrimination grounds, or they apply only to certain vulnerable groups. The Convention on the Elimination of All Forms of Racial Discrimination (ICERD, 1965) and Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW, 1979), are legally binding universal instruments containing implementation mechanisms. The Committee on the Elimination of Racial Discrimination (CERD) bases its practice on the 'living instrument' doctrine - a key vehicle for evolution and innovation within the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD). This ensures the treaty can respond to contemporary manifestations of racial discrimination while staying within the scope of its provisions.[49] This approach is important for the purpose of extending the scope of these conventions to online discrimination.

We should also mention the UN Guiding Principles on Business and Human Rights (United Nations Human Rights Council, 2011) as a set of guidelines for States and companies to prevent and address human rights abuses committed in business operations. They are a prevailing soft law standard for the human rights responsibility of private actors.[50]The Guiding principles reaffirm that states must ensure that not only State organs, but also businesses under their jurisdiction respect human rights. [51] These Principles

---

[44] The UN Working Group On Business And Human Rights, The UN Guiding Principles On Business And Human Rights An Introduction, Intro_Guiding_PrinciplesBusinessHR.pdf (ohchr.org), last accessed 30.3.2022.

[45] § 26 of the Charter of United Nations, available at: Charter of the United Nations.pdf (undp.org), last accessed 28.3.2022.

[46] Of the thirty articles, some are in one way or another explicitly concerned with equality, and the rest implicitly refer to it by emphasizing the all-inclusive scope of the Universal Declaration of Human Rights.

[47] § 26 of ICCPR reads: "All persons are equal before the law and are entitled without any discrimination to the equal protection of the law. In this respect, the law shall prohibit any discrimination and guarantee to all persons equal and effective protection against discrimination on any ground such as race, color, sex, language, religion, political or other opinion, national or social origin, property, birth or other status."

[48] European Court of Human Rights, App. No. 6538/74, The Sunday Times v The United Kingdom, 26 April 1979, para 59.

[49] Keane, David, Mapping the International Convention on the Elimination of All Forms of Racial Discrimination as a Living Instrument, Human Rights Law Review, Vol. 20, Issue 2, 2020, p. 236.

[50] Jørgensen, Rikke Frank and Møller Pedersen, Anja, Online service providers as human rights arbiters, The Responsibilities of Online Service Providers. Mariarosaria Taddeo; Luciano Floridi (eds.) Springer, Law, Governance and Technology Series, Vol. 31).2017. p. 179-199.

[51] Guiding Principles on Business and Human Rights, Ch. 1 (A) (1).

assert a global responsibility for businesses to avoid causing or contributing to adverse human rights impacts through their own activities, and to address such impacts when they occur and seek to prevent or mitigate adverse human rights impacts that are directly linked to their operations, products of services, even if they have not contributed to those impacts. [52] As a matter of transparency, the Guiding Principles state those businesses should be prepared to communicate how they address their human rights impacts externally, particularly when the concerns are raised by or on behalf of affected stakeholders. [53]

None of these documents mentioned above are specific for online discrimination but are used as general sources of law. There does not exist any international conventions on intermediary liability either.

Regional documents

Council of Europe (CoE)

The most important document on human rights in Europe - European Convention on Human Rights (hereafter Convention or ECHR) from 1950 explicitly forbids discrimination in Art 14.

According to the Court's case-law, the principle of non-discrimination is of a "fundamental" nature and underlies the Convention together with the rule of law, and the values of tolerance and social peace.[54]

The expression "direct discrimination" describes a "difference in treatment of persons in analogous, or relevantly similar situations" and "based on an identifiable characteristic, or 'status'[55] protected by Article 14 of the Convention.[56] Article 14 however is an accessory right, and it applies only in conjunction with other Convention rights.

In addition to article 14, in Article 1 of Protocol 12 to the Convention,[57] it is stated:

1. The enjoyment of any right set forth by law shall be secured without discrimination on any ground such as sex, race, color, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.
2. No one shall be discriminated against by any public authority on any ground such as those mentioned in paragraph 1.

Article 1 of the Protocol 12 hence represents a general prohibition of discrimination[58] and an independent right not to be discriminated against. It confirms that the state has both positive and negative obligations - to secure protection of individuals against discrimination and not to actively discriminate.

---

[52] Ibid. Ch. II (A) 11-13.

[53] Ibid. Ch. II (B) (21).

[54] S.A.S. v. France [GC], Application no. 43835/11 of July 1st 2014, § 149; Străin and Others v. Romania, Application no. 57001/00 of July 25 2005, § 59.

[55] Biao v. Denmark [GC], Application no. 38590/10 of May 24th 2016, § 89; Carson and Others v. the United Kingdom [GC], Application no. 42184/05 of March 16th 2010, § 61; D.H. and Others v. the Czech Republic [GC], Application no. 57325/00 of November 13th 2007, § 175; Burden v. the United Kingdom [GC], 2008, § 60.

[56] Varnas v. Lithuania, Application no. 42615/06 of December 9th 2013, § 106; Hoogendijk v. the Netherlands, Application no. 58641/99 of January 6th 2005.

[57] 20 states have ratification Protocol 12 as of November 2021.

[58] Savez crkava "Riječ života" and Others v. Croatia, Application no. 7798/08, December 9th 2010, § 103; Sejdić and Finci v. Bosnia and Herzegovina [GC], application nos. 27996/06 and 34836/06, December 22nd 2009, § 53.

Article 10 ECHR contains the right to freedom of expression, and in paragraph 2 of Article 10 the Convention acknowledges that the exercise of these freedoms carries with it duties and responsibilities, and 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 (...).

In addition to ECHR, CoE countries have adopted several other conventions with regard to specific protected grounds. The Framework Convention for the Protection of National Minorities of 1994 (FCNM) in its Article 6, encourages Parties to intercultural dialogue and to take appropriate measures to protect persons who are subject to threats or acts of discrimination, hostility or violence due to their ethnic, cultural, linguistic or religious identity. This convention does however not have any specific rules on online discrimination.

Other relevant Council of Europe documents

The Council of Europe Commission against Racism and Intolerance (ECRI) established in 1993 is the CoE's independent human rights monitoring body.[59] Its mandate is combating racism, discrimination, (on grounds of "race", ethnic/national origin, color, citizenship, religion, language, sexual orientation and gender identity), xenophobia, antisemitism and intolerance in Europe.

ECRI issues General Policy Recommendations (GPRs) addressed to the governments of all member states. These recommendations provide guidelines which policymakers are invited to use when drawing up national strategies and policies. One recommendation relevant for this study is ECRI General Policy Recommendation No. 6 (2000) - Combating the dissemination of racist, xenophobic and antisemitic material via the Internet.

Recommendation CM/Rec (2014) 6 of the Committee of Ministers to member states on a Guide to human rights for Internet users states that restrictions to online freedom of expression may apply to expressions which incite discrimination, hatred or violence. These restrictions must be lawful, narrowly tailored and executed with court oversight. So even though the freedom of expression is highly valued right, it is not absolute, and can be restricted in order to protect other interests, such as the right not to be discriminated against.

A distant reference to social media is contined in Recommendation CM/Rec (2011) 7 on a new notion of media. According to this document, the actors operating collective online, shared spaces which are designed to facilitate interactive mass communication should be attentive to the use of, and editorial response to, expressions motivated by racist, xenophobic, anti-Semitic, misogynist, sexist (including as regards LGBT people) or other bias. These actors may be required (by law) to report to the competent authorities' criminal threats of violence based on racial, ethnic, religious, gender or other grounds that come to their attention.[60] The threshold for the platforms to act is high, as it requires the discrimination to reach the level of criminal threats.

Some of the CoE recommendations contain the measures which are meant to help remedy the violations of the human rights online. In appendix to Recommendation Rec (2001) 8 of the Committee of Ministers to

[59] European Commission against Racism and Intolerance (ECRI) - Homepage (coe.int) last accessed 31.3.2022.

[60] § 91.

member states on self-regulation concerning cyber content (self-regulation and user protection against illegal or harmful content on new communications and information services), the member states are encouraged to establish content complaints systems, such as hotlines, which are provided by Internet service providers, content providers, user associations or other institutions. Such content complaints systems should, where necessary for ensuring an adequate response against presumed illegal content, be complemented by hotlines provided by public authorities.[61]

Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries requires the states to guarantee accessible and effective judicial and non-judicial procedures that ensure the impartial review, in compliance with Article 6 of the ECHR, of all claims of violations of Convention rights in the digital environment, including the right not to be discriminated against in the enjoyment of all the rights and freedoms set forth in the ECHR. They should furthermore ensure that intermediaries provide users or affected parties with access to prompt, transparent and effective reviews for their grievances and alleged terms of service violations, and provide for effective remedies, such as the restoration of content, apology, rectification or compensation for damages. Judicial review should remain available, when internal and alternative dispute settlement mechanisms prove insufficient or when the affected parties opt for judicial redress or appeal. [62]

In the most recent Recommendation CM/Rec(2022) 13 of the Committee of Ministers to member States on the impacts of digital technologies on freedom of expression, on 6 April 2022, the issue of algorithmic discrimination is expressly addressed:

"When there are legitimate concerns that their policies may lead to discrimination, internet intermediaries should provide information that allows independent third parties to evaluate whether their policies are implemented in a non-discriminatory way, including by disclosing the datasets upon which automated systems are trained in order to identify and correct sources of algorithmic bias."[63]

Some of the CoE recommendations relate to protection against discrimination of particular groups. For instance, Recommendation CM/Rec (2010) 5 on measures to combat discrimination on grounds of sexual orientation or gender identity includes the obligation to combat inciting hatred or other forms of discrimination against LGBTI+ persons. It covers all forms of discrimination, including online discrimination.

EU law

European union has for decades worked to establish the values of inclusion, non-discrimination, multilingualism and cultural diversity, which are epitomized in EU's motto: "United in diversity", and it is crucial for the EU as a project that these values also are reflected in the online setting. The prevention and response to discrimination are values enshrined in Article 2 of the Treaty of the European Union (TEU).[64] This provision states that:

[61] Chapter IV, 12.

[62] Art. 1.5.1 and 1.5.2.

[63] § 3.6.

[64] Official Journal of the European Union C 326/15 of 26.10.2012.

"The Union is founded on the values of respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities. These values are common to the Member States in a society in which pluralism, non-discrimination, tolerance, justice, solidarity and equality between women and men prevail."

Prohibition of discrimination is contained in EU's Charter of fundamental rights[65] Article 21:

1. "Any discrimination based on any ground such as sex, race, color, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.
2. Within the scope of application of the Treaties and without prejudice to any of their specific provisions, any discrimination on grounds of nationality shall be prohibited."

On the level of secondary legislation, European antidiscrimination legislation is contained in directives,[66] such as Directive 2000/43/EC against discrimination on grounds of race and ethnic origin,[67] Directive 2000/78/EG establishing a general framework for equal treatment in employment and occupation[68], Directive 2006/54/EG, on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast)[69]and the E-Commerce Directive.[70]

The EU regulatory framework on content moderation online is increasingly complex and has been differentiated over the years according to the category of the online platform and the type of content reflecting a risk-based approach. For instance, Audio-Visual Media Services Directive,[71] imposes particular obligations to one category of online platforms, the VideoSharing Platforms. They should take appropriate and proportionate measures, preferably through co-regulation, in order to protect the general public from illegal content. Those measures must be appropriate in the light of the nature of the content, the category of persons to be protected and the rights and legitimate interests at stake and be proportionate taking into account the size of the platforms and the nature of the provided service.

The Counter-Racism Framework Decision (CRFD), which was adopted by the Council alone, seeks to combat particularly serious forms of racism and xenophobia through criminal law but does not define racism and xenophobia nor use the terms racist and xenophobic hate speech.[72]

Instead, the CRFD criminalizes two types of speech - publicly inciting to violence or hatred and publicly condoning, denying or grossly trivializing crimes of genocide, crimes against humanity and war crimes - when they are directed against a group of persons or a member of such a group defined by reference to race,

[65] Charter of Fundamental Rights of the European Union, OJ C 326, 26.10.2012, p. 391–407.

[66] Directives are a method of harmonization of the law, i.e. legal acts that are binding as to the result to be achieved, but that leaves member states discretion as to how to achieve the result.

[67] Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin. OJ L 180, 19.7.2000, p. 22–26.

[68] OJ L 303, 02.12.2000, p. 16-22.

[69] OJ L 204, 26.7.2006, p. 23–36.

[70] OJ L 178, 17.7.2000, p. 1–16.

[71] Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive), OJ L 95, 15.4.2010, p. 1–24.

[72] 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, O.J. [2008] L 328/55.

color, religion, descent or national or ethnic origin. The list of protected grounds is limited to these five characteristics.

While only hate speech which is racial and xenophobic has been made illegal by the CRFD, Member States may go beyond the EU minimum and criminalize other types of hate speech, by referring to a broader list of protected characteristics (a list including e.g., religion, disability, sexual orientation).

In 2017 European commission issued a Communication on tackling illegal content online[73] where it expressed the need for the online platforms to, in light of their central role and capabilities and their associated responsibilities, adopt effective proactive measures to detect and remove illegal content online and not only limit themselves to reacting to notices which they receive.[74] In 2018 a Recommendation on measures to effectively tackle illegal content online[75] was adopted calling the online platforms to act in a diligent and proportionate manner towards the content they host, especially when processing notices and counter-notices and deciding on the possible removal of or disabling of access to content considered to be illegal. 'Illegal content' arguably encompasses a large variety of content categories that are not compliant with EU and national legislation, including discrimination.

According to the E-Commerce Directive**,** which is considered to be foundational legal framework for online services in the EU , the intermediaries benefit from "safe harbors"- which means that they cannot be subject to a general obligation to monitor users' online content, and they are exempt from liability unless they are aware of the illegality and are not acting adequately to stop it. [76] As the social media platforms are considered to be passive and neutral, they are exempted from liability for illicit content posted by their users. However, they cannot rely on the exemptions from liability if they were aware of the facts or circumstances on the basics of which a diligent economic operator should have realized that the publication was unlawful and failed to act expeditiously. [77]

National laws - examples from Australia, Germany and France

On September 8th 2021 the High Court of Australia found that media companies may be liable for the defamatory comments of third parties on social media platforms.[78]

By this, the stronger protection is given to right to reputation vis-a vis right to freedom of expression, as the liability for damages is not limited to the author of the defamatory comments but extends to publishers who allow the defamatory content to be posted on their social media platforms and thereby encourage or facilitate the defamatory publication. Australian law has thereby abandoned the safe harbor principle and opted for a stricter liability of intermediaries with regards to illegal content.

[73] Communication from the Commission to the European parliament, the Council, the European economic and social committee and the Committee of the regions tackling illegal content online -towards an enhanced responsibility of online platforms, COM/2017/0555 final

[74] § 10 of the Communication

[75] Commission Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online. C/2018/1177

[76] See Directive 2000/31 EC (e-commerce directive), art. 12.

[77] C-324/09 of 12 July 2011, § 119.

[78] High Court of Australia in Fairfax Media Publications Pty Ltd v Voller; Nationwide News Pty Ltd v Voller; Australian News Channel Pty Ltd v Voller, S236/2020 S237/2020 S238/2020.

In Europe, Germany's Network Enforcement Act, or NetzDG law[79] represents a key test for combatting hate speech on the internet. Under the law, which came into effect on January 1, 2018, online platforms face fines of up to €50 million for systemic failure to delete illegal content. Supporters see the legislation as a necessary and efficient response to the threat of online hatred and extremism. Critics view it as an attempt to privatize a new 'draconian' censorship regime, forcing social media platforms to respond to this new painful liability with unnecessary takedowns.[80]

In France, recent case law has imposed proactive monitor obligations on intermediaries for copyright infringement even in cases where their liability is not engaged.[81]

This shows that the national legislators are moving towards strengthening the protection of rights online and the position of individuals in relation to big tech companies, and these rules can be applied *mutatis mutandis* to online discrimination as a form of illegal content.

Tech companies' internal regulations and policies

Over the past decade, tech giants such as Google (Alphabet) and Facebook (Meta) have become the biggest companies in the world.[82]

Despite the fact the companies such as Facebook have insisted they are only neutral platforms and have no editorial responsibilities, they have recently been under pressure to filter communication that appears on their platforms, including discriminatory content. As a result, they have introduced more rules to govern speech and participation and instituted several mechanisms for the removal of people and content that transgress these rules.

Terms of service (TOS) which individuals typically must accept as a condition to access the platform, often contain restrictions on content that may be shared. The inconsistent enforcement of terms of service however has also attracted public scrutiny. Some have argued that the world's most popular platforms do not adequately address the needs and interests of vulnerable groups, for example there have been accusations of reluctance to "engage directly with technology related violence against women, until it becomes a public relation issue". [83]

Voluntary codes of conduct for internet service providers have been adopted in several European states (e.g. the Netherlands, UK). In May 2016, European Commission and several platforms[84] agreed on the common "Code of conduct on countering illegal hate speech online"[85] to prevent and counter the spread of

---

[79] » Network Enforcement Act (Netzdurchsetzunggesetz, NetzDG) German Law Archive (iuscomp.org), last accessed 31.3.2021.

[80] The Impact of the German NetzdG law – CEPS

[81] See APC v. Google, Tribunal de Grande Instance [TGI] [ordinary court of original jurisdiction] Paris, Nov. 28, 2013 (Fr.), https://www.legalis.net/jurisprudences/tribunal-de-grande-instance-de-paris-ordonnance-de-refere-28-novembre-2013/ [https://perma.cc/7JA2-37PB]; Nord-Ouest Prod. v. S.A. Daily Motion, Tribunal de grade instance [TGI] [ordinary court of original jurisdiction] Paris, July 13, 2007 (Fr.), https://www.legifrance.gouv.fr/affichJuriJudi.do?oldAction=rechJuriJudi&idTexte= JURITEXT000018861366&fastReqId=728956270&fastPos=2 [https://perma.cc/JX6L-45GZ]

[82] Top 20 Biggest Tech Companies in The World in 2021 - The Teal Mango last accessed 31.3.2021.

[83] See: Human Rights Council, Report of the Special Rapporteur on promotion and Protection to the Right of freedom of opinion and Expression, p. 14.

[84] Code of conduct is joined by Facebook, Microsoft, Twitter and YouTube, Instagram, Snapchat, Dailymotion Jeuxvideo.com. TikTok joined in September 2020. On 25 June 2021, LinkedIn also announced its participation to the Code of Conduct.

[85] The EU Code of conduct on countering illegal hate speech online | European Commission (europa.eu), last accessed 30.4.2022.

illegal hate speech online. The last evaluation shows that on average the companies are now assessing 81% of flagged content within 24 hours and 62.5% of the content deemed illegal hate speech is removed.[86]

The biggest disadvantage of such Codes of conduct are that they are self-regulatory mechanisms and are joined by platforms on voluntary bases. In other words, they are not enforceable.

Facebook has been under fire for giving advertisers the ability to exclude people from their targeting based on race, religion, sexual orientation. Now, it needs advertisers to comply with their updated non-discrimination policy. [87] In addition, Meta Platform Terms and Developer Policies state that one of the prohibited practices is:

"Processing Platform Data to discriminate or encourage discrimination against people based on personal attributes including race, ethnicity, color, national origin, religion, age, sex, sexual orientation, gender identity, family status, disability, medical or genetic condition, or any other categories prohibited by applicable law, regulation, or Meta policy."[88]

Tik Tok's content moderation policy contains unusual measures to protect supposedly vulnerable users. The platform instructed its moderators to mark videos of people with disabilities and limit their reach, as these users are „susceptible to harassment or cyberbullying based on their physical or mental condition".[89]

Online platforms as gatekeepers

Online platforms as private actors are powerful forces in facilitating freedom of expression online.[90] They are also seen as gatekeepers[91] and the first line of defence for protection of users' human rights online. EU regulators for example have outsourced the "first line" protection of human rights to intermediaries who are given responsibilities and tasks to disable or remove alleged illegal content on the internet. The companies have notice and take down procedures and are supposed to react timely and transparently to complaints of discriminatory content on their sites. These procedures should be simple and clear, making it easier for the victims of discrimination to get remedied. Here, the American Digital Millennium Copyright Act (DMCA) can be used as an inspiration. It explicitly regulates who should issue the notification of (copyright) infringement, to whom and what it should contain.

According to the E-Commerce Directive, intermediaries are exempted from liability for third party content (so called Safe harbors).[92] This has been confirmed in the ECtHR practice in the case of *MTE and Index v. Hungary*[93] where the ECtHR had to decide whether a non-profit self-regulatory body of Internet content providers (MTE) and an Internet news portal (Index) were liable for offensive comments posted

[86] The EU Code of conduct on countering illegal hate speech online | European Commission (europa.eu), last accessed 30.04.2022.

[87] Review compliance for Facebook's Non-discrimination Policy | Facebook Business Help Centre

[88] 3 a (i), available at: Platform Terms - Facebook for Developers

[89] Discrimination: TikTok curbed reach for people with disabilities (netzpolitik.org), last accessed 31.3.2021.

[90] Laidlaw, Emily, Regulating Speech in Cyberspace: Gatekeepers, Human Rights, and Corporate Responsibility, Cambridge University Press, 2015, p. i.

[91] Ibid.

[92] 12-14 of Directive 2000/31/Ec Of The European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (Directive on electronic commerce)

[93] Magyar Tartalomszolgáltatók Egyesülete and Index.hu zrt v. Hungary, Application no. 22947/13, February 2nd 2016.

on their websites. The Court found that the intermediaries had no liability for the infringement of rights, and that "the notice-and-take-down-system could function in many cases as an appropriate tool for balancing the rights and interests of all those involved".[94]

However, it was pointed out in literature that this system allows platforms to tolerate discrimination, and this can perpetuate inequalities in our society.[95] Given the current state of technology, it should not be impossible for the tech companies to police user content take as a part of their due diligence.

It has been suggested to introduce the "systemic duty of care"[96] model- a legal standard for assessing a platforms overall system for handling online content. The idea is that platforms should improve their systems for reducing online harms, including detecting and removing illegal content. The systemic duty of care could be based on one of two models: A "prescriptive" model which defines precisely the measures a platform must take, and "flexible model in which the measures are left undefined.[97]

It is believed that the legal basis for this could be the E-Commerce Directive, which references to potential duties to "detect and prevent" illegal activities.[98]

Conclusion - suggestions and best practices

The online discrimination is often only a manifestation of the existing disparity and discrimination that exists in the "real world". In other words, digital rights violations often affect those who are already marginalized.

Online discrimination can be seen as a more serious form of discrimination, as the internet can make discrimination more visible and hurtful. As the issue of digital discrimination is likely to persist in the future, the platforms have the responsibility to prevent and minimize the effects of this form of human rights infringement, under the control of the state. Even though the automated systems for content detection or notice and take down systems enable prevention and quick reaction to infringements, these measures suffer from the lack of procedural safeguards and judicial review.

Online discrimination requires effective responses on international, regional, national levels, setting standards for the tech companies who are the gatekeepers and first lines of defense against this practice. The current situation in Europe regarding the liability of platforms is very diffuse and unclear.

Outsourcing the difficult task to protect against discrimination and balance different interests online is problematic as the procedural safeguards cannot be ensured. The platforms are not judicial organs and cannot be said to have a legal expertise to evaluate requests against general legal criteria. The Interamerican Commission on Human Rights has observed that private actors "lack the ability to weigh rights and interpret the law in accordance with freedom of speech and other human rights standards."[99] In addition,

---

[94] Magyar Tartalomszolgáltatók Egyesülete and Index.hu zrt v. Hungary, Application no. 22947/13, February 2nd § 9.

[95] Sylvain, O., Discriminatory Designs on User Data, Emerging threats, 2018, available at: Discriminatory Designs on User Data | Knight First Amendment Institute (knightcolumbia.org) last accessed 01.05.2022.

[96] Keller, D, Systemic duties of care and intermediary liability, Center for internet and society, 2020, available at: Systemic Duties of Care and Intermediary Liability – Daphne Keller – Inforrm's Blog, last accessed 8.5.2022.

[97] Ibid.

[98] Recital 48.

[99] Inter-American Commission in Human Rights, Freedom of Expression and the Internet, pp. 47-48.

there is problem of lack of transparency and to a large extent different attitudes and approaches by the various providers.[100] In addition, intermediaries that operate in diverse range of markets inevitably face "complex value judgements", issues with cultural sensitivity and diversity and "difficult decisions about the conflict of laws".[101]

In other words, even though the alleged infringement of right not to be discriminated against handled by the platform is a quicker and more efficient way to justice, it lowers legal certainty. Therefore, controlling the access to content and services has to be subject to judicial review.

The online platforms should, as part of their social responsibly due diligence, contribute to prevention and protection of the right not to be discriminated against in order to secure the equality of all citizens. They should be incentivized to uncover the illegal content, rather then be punished for not preventing it. This way, the potential "over-removing" in order to avoid liability would be avoided. The applications should be designed to elicit and ensure that certain kinds of content do not even occur in the first place.

Even though the safe harbor principle still applies in Europe and the USA, platforms can hardly be said to be neutral and passive in the modern world. Intermediaries are not mere conduits that purport to provide free and uninhibited forum for social interaction. They are implicated in every user utterance or act, even if they do not moderate posts.[102] They have the ability to control over the information published on the platform, and they have financial profits form the information they convey.[103] In addition, intermediaries design their platforms in ways that shape the form and substance of their users' content.[104] Their role is no longer to transmit or store material on behalf of the users-rather it fulfils an active role in the organization and functioning of the websites.[105] It is in control of what users see, much like a newspaper editor - and like editors should have some form of liability.[106]

Given all said, what could be effective practices and policies that address, mitigate and/or prevent online discrimination?

- Online discrimination can have grave consequences for public safety and social inclusion and should be expressly addressed in international, legal and national regulations, and these sources of law should be harmonized
- States, tech companies and NGOs should work together on raising awareness of the problem of discrimination online, so people can recognize discriminatory practices and know their rights
- More research about online discrimination is needed so this practice can be recognized and better addressed
- Tech companies ought to share best practices in detecting and avoiding discriminatory practices

[100] Van der Sloot, p. 224.

[101] Taylor, E, The Privatization of Human Rights: Illusions of Consent, Automatisation and Neutrality, GCIG Paper No. 24 (2016).

[102] Sylvain, O, Intermediary Design Duties, Connecticut Law Review, vol 50, nr. 1 2018, p. 226.

[103] Sylvain, O., Discriminatory Designs on User Data, Emerging threats, 2018, available at: Discriminatory Designs on User Data | Knight First Amendment Institute (knightcolumbia.org) last accessed 01.05.2022.

[105] Van der Sloot, B, Welcome to the Jungle:the Liability of Internet Intermediaries for Privacy Violations in Europe, JIPITEC, 6(2015), p. 212.

[106] Keller, D, op.cit.

- Tech companies ought to cooperate on developing the automated systems of content control instead of developing parallel systems, which would be more cost efficient and result in more harmonized systems
- Filtering algorithms would require human review to prevent human rights violations and discrimination.
- The existing mechanisms for reputational and copyright protection such as notice and take down procedures and the right to be forgotten can analogously be applied in case of online discrimination
- Online platforms should have independent bodies consisting of law experts evaluating the reported cases of discrimination in order to achieve better balancing of rights
- There is a need for additional transparency measures for online platforms, including on the algorithms used. Platforms that feature user-generated content should offer users a clear explanation of their approach to evaluating and resolving reports of hateful and discriminatory content, highlighting their relevant terms of service
- Greater ease for reporting cases of online discrimination (user-friendly mechanisms and procedures)
- Platforms should enforce sanctions of their terms of service in a consistent, timely and fair manner
- Platforms should abide by duty of care, going beyond notice-and-takedown based legal models
- Internet intermediaries can no longer be considered passive and neutral transmitters of information, and there should not be exempt from liability for online discrimination
- Legislative framework for handling of requests to take down discriminatory content should be put in place
- Procedural protections should be built into platforms notice-and-takedown systems
- Rules should incentivize intermediaries and users to detect illegality, while minimizing the risks and the costs of errors and safeguarding a balance between the different human rights at stake
- Tech companies need to ensure algorithm transparency and neutrality
- A balance between citizens and tech companies must be struck in designing the liability rules
- Setting up a detailed and harmonized European notice and take down procedure would provide more legal certainty

Online Hate Speech - User Perception and Experience Between Law and Ethics

Gregor Fischer-Lessiak, Susanne Sackl-Sharif and Clara Millner
UNIVERSITY OF GRAZ | UNIVERSITY OF MUSIC AND PERFORMING ARTS | ANTIDISCRIMINATION OFFICE STYRIA

Introduction

Online hate speech (OHS) is a virulent social problem that has been challenging democratic discourses in the past years. The storming of the US Capitol in January 2021 is the most prominent high-level case of online disinformation and OHS leading to real-life consequences. As well, by the end of 2021, anti-vaccination propaganda and hate speech against medical staff have gained momentum (Gleicher et al, 2021).

Due to the great relevance of the topic, it is unsurprising that OHS research has been intensifying in recent years. Many studies deal with the question of how OHS can be evaluated from a communication studies or media studies perspective and/or discuss technical conditions of social media and societal changes related to OHS (e.g. Pörksen, 2018; Sponholz, 2018; Zannettou et al, 2020). Other studies address the content of OHS, its impact on those affected as well as coping strategies (e.g. Anderson et al, 2014; Brodnig, 2016; Lumsden and Morgan, 2017). When it comes to the regulation of OHS, which is the focus of this paper, a lot has been written lately about platforms' and states' duties to regulate and counter OHS effectively (e.g. Brown, 2020; Davidson et al, 2017; Waseem and Hovy, 2016). But a central aspect, namely users' opinions on OHS and its regulation, remains under-researched. Especially in the German-speaking area, there are only a few studies on this topic, e.g. Geschke et al (2019) that explore how users perceive state- and platform-made norms on OHS and their implementation.

To contribute to this debate using empirical data, we conducted an online survey (November/December 2020, 157 respondents, Austria) to gather the opinions and experiences of users with OHS based on five OHS model postings. In this paper, we interpret and contextualize their responses based on legal (criminal and human rights law) and sociological approaches. First, we provide the theoretical backdrop. Second, we present our empirical research design. Third, we analyse how respondents assessed our model postings legally and ethically. This includes respondents' willingness to take action against these postings. Fourth, we present cross-case analyses and conclusions.

Considerations on 'hate', the law, and the user

Human rights and online hate

In human rights law, the discussion on the management/moderation of OHS is generally framed around freedom of expression and its boundaries. While there is no authoritative definition of 'hate speech', the term is widely used as an umbrella term. Lately, the United Nations Strategy and Plan of Action on Hate Speech (United Nations, 2020: 8) offers the following definition of 'hate speech': "Any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion,

ethnicity, nationality, race, colour, descent, gender or other identity factor. This is often rooted in, and generates, intolerance and hatred, and in certain contexts can be demeaning and divisive."

This Action Plan builds on a plethora of international and regional human rights documents that stipulate state duties to take action against hate speech. For example, Art. 4 (a) ICERD explicitly imposes state duty to prohibit expression that promotes racial hatred, hereby setting boundaries to the freedom of expression to protect the rights of targets/target groups of hate speech. These state obligations apply substantively online and offline. The fulfilment thereof, however, can be procedurally challenging in internationalized online environments (Brown, 2020). Given the lack of immediate state governance, users are confronted mainly with media providers' business ethics, their terms of service and the latter's implementation. To illustrate this fact, we provide the example of Austrian hate speech and platform governance.

Austria: criminal law fulfils positive obligations

In human rights law, states enjoy a certain leeway, a so-called margin of appreciation, in deciding how to combat hate speech, allowing them to take into account their historical backgrounds as well as legal traditions. For example, denial of the Shoah is outlawed in many EU member states by a variety of norms. In spite of the EU's harmonization efforts (Council Framework Decision 2008/913/JHA, 2008), these norms still differ from state to state – and beyond EU borders, there is even less uniformity. Austria has decided within its margin of appreciation that there are public interests – especially the prevention of a reinvigorated National Socialist movement – justifying the exclusion of these expressions from legal protection, and their prosecution. Postings and comments on social media that fulfil the criteria of these provisions hence lead to law enforcement action just like expressions in the 'analogue' world.

While some argue that social media platforms could be treated as accessories to hate speech crimes in cases of non-deletion of OHS (e.g. Austrian Ministry of Justice, 2016), social media platforms have not (yet) been held criminally liable this way. Contrarily, platforms have profited from OHS, legal and illegal forms alike, as it produces high interaction numbers raising platforms' ad revenue. Social media companies' algorithms have even learned to accelerate the spread of OHS (Zannettou et al, 2020). At the same time, ethical considerations built into terms of service have been meandering between recognition for the problematics of OHS and overstressing US-American freedom of speech doctrine (Kang and Isaac, 2019).

Human rights considerations are more and more included in community standards (e.g. Facebook, 2021a). A strong, strictly legal duty of platforms to moderate according to international human rights law, however, has not emerged yet. While community standards ban hate speech, they lack efficacy until now.

Laws, platform rules, ethics and the user

Given these shortcomings, states have tried to enhance platform accountability by other means. Since 2021, for example, the Austrian Communication Platforms Act (CPA, BGBl. I Nr. 151/2020), links the moderation practices of large private social media networks to national hate speech provisions. If platforms fail to provide effective reporting tools and transparency reports on their moderation, they might face harsh financial penalties. Austrian national law may hence be able to exert immediate effects on the practices of social media platforms (e.g. Facebook, 2021b). As well, the law may provide for new avenues of cooperation between law enforcement and social media. In the past, investigations against original posters of OHS have been hampered by their non-cooperation and hesitance to provide national authorities with user data needed to trace OHS suspects under national criminal law (Haider and Millner, 2021).

Only lately, platforms have moved towards more comprehensive implementation of due process in their OHS rules by the establishment of court-like entities, e.g., Facebook's/Instagram's/Meta's Oversight Board (2021). The corporation herein sought to supplement business ethics with legal considerations based on global freedom of expression standards. This work is ongoing. At the same time, and seemingly contrarily, Facebook is challenging the Austrian Communication Platforms Act. The company's motion is currently pending review by the Austrian Administrative Court of Appeal. In this lawsuit, the law's compliance with the EU E-Commerce-Directive (2000/31/EC, 2000) will have to be assessed. Whatever the outcome of the case may be, platforms are not going to be able to hold off extended moderation duties forever. US lawmakers are considering reforms of platform responsibility (Cole et al, 2021), and the EU is preparing new legislation momentarily that is expected to harmonize platform duties in Europe (Digital Services Act, 2020). Amending policies and rethinking business ethics now would help platforms anticipate these new legal developments. More immediate reasons to do so exist: Users affected by OHS report, inter alia, disengagement from societal debates, psychological trauma and even physical effects (e.g. GREVIO, 2021: §12).

In the following, we present our empirical research of user experiences with OHS on social media platforms. We include their views on the applicable laws, their ethical considerations on freedom of expression vis-a-vis OHS and their opinions on reporting mechanisms offered by platforms.

Research design

In November and December 2020, we conducted an online survey to gather the public's perceptions of and experiences with OHS. The survey focused on five main topics: definition of OHS, platforms and contents, affected persons and coping strategies, OHS perpetrators, and counter-speech strategies. In this paper, we focus on the evaluation of five model cases that resemble real OHS postings sourced from the data of an Austrian OHS reporting app offered by the Antidiscrimination Office Styria. In the selection of cases, we tried to integrate different OHS target groups, contents, and degrees of intensity. In this context, we were guided by previous empirical results on OHS (e.g. Geschke et al, 2019) and by legal considerations. Respondents were asked to assess the postings with regards to their (perceived) illegality and their (ethical) worthiness of remaining online. We provided 11 items on a 5-point rating scale as well as the opportunity to enter further information in an open response field. The original questionnaire was published in German, the official language in our survey area. All questionnaire responses described below were translated by the authors.

Our purposive sampling strategy focused on two groups of people: 1) young people, as they are particularly frequent users of social media and more likely to be exposed to OHS (Gadringer et al, 2021) and 2) members/employees of organisations dealing with OHS in their work practice. The sample hence mainly includes persons frequently confronted with social media and/or OHS in possession of expert knowledge. Our sample comprises 157 persons. About half of them are between 20 and 29 years old (46%), 5% are younger than 20, 37% are between 30 and 49, and 12% older than 50. Almost 30% are students, almost 20% work in the educational and social sector, 13% as (office) employees and 6% in STEM professions outside the education sector. More than half of the respondents identified as female (59%), 37% as male, and 4% as other genders.

As the complexity of OHS requires an open-explorative and interdisciplinary research approach, the analysis of the five model cases includes social science and legal evaluations. Our social sciences analysis

included usual statistical descriptive rations such as arithmetic mean or standard deviation, but also the clustering of open-ended questions according to frequency. Our legal analysis is based largely on Austrian criminal law doctrine, accompanied by human rights law considerations. Each case is analysed individually before essential similarities and differences are worked out in a cross-case analysis.

Case studies

In this chapter, we describe and discuss the main characteristics of our five model cases. We assess these cases based on Austrian criminal law and compare this analysis with the responses to our questionnaire. Table 1 gives an overview of the ratings of the 11 items per case (5=complete agreement; 1=complete disagreement). The items can be clustered into three content groups: a) online hate speech vs. expression of opinion (items 1-2), b) evaluation of posting according to punishability (items 3-4), c) responses/strategies (items 5-11).

Table 1: Overview of case studies evaluations, complete model postings below

When I see this posting publicly on social media...	case 1 virgins	case 2 scum	case 3 cunt	case 4 refugees	case5 siblings
... I perceive it as online hate speech.	M=4.30 SD=0.35	M=4.76 SD=0.41	M=4.82 SD=0.41	M=3.55 SD=0.27	M=4.32 SD=0.35
... it's a normal expression of opinion for me.	M=1.31 SD=0.40	M=1.17 SD=0.42	M=1.12 SD=0.41	M=2.24 SD=0.29	M=1.42 SD=0.38
... I think it's punishable.	M=2.69 SD=0.27	M=4.08 SD=0.33	M=3.69 SD=0.28	M=2.01 SD=0.31	M=2.41 SD=0.28
... I think it should be punishable.	M=3.29 SD=0.26	M=4.38 SD=0.36	M=4.24 SD=0.34	M=2.44 SD=0.27	M=2.97 SD=0.26
... I think it should be deleted.	M=4.26 SD=0.34	M=4.82 SD=0.39	M=4.83 SD=0.41	M=3.42 SD=0.26	M=4.17 SD=0.33
... I do nothing.	M=2.99 SD=0.26	M=2-21 SD=0.30	M=2.24 SD=0.29	M=2.87 SD=0.26	M=2.77 SD=0.31
... I report it to the social platform.	M=3.29 SD=0.26	M=3.92 SD=0.31	M=3.88 SD=0.30	M=2.37 SD=0.28	M=2.92 SD=0.26
... I report it via NGO app.	M=2.17 SD=0.34	M=2.57 SD=0.31	M=2.43 SD=0.31	M=1.76 SD=0.38	M=1.99 SD=0.35
... I report it to the police.	M=1.37 SD=0.40	M=2.01 SD=0.32	M=1.66 SD=0.36	M=1.32 SD=0.41	M=1.36 SD=0.16
... I answer it publicly.	M=2.03 SD=0.31	M=2.42 SD=0.28	M=2.49 SD=0.28	M=2.40 SD=0.28	M=2.09 SD=0.31
... I answer in a private message.	M=1.53 SD=0.37	M=1.59 SD=0.37	M=1.62 SD=0.36	M=1.63 SD=0.36	M=1.40 SD=0.39

Case 1: '72 virgins'

You see a picture of a (Muslim) man and a flock of sheep with the caption: "72 virgins just for you; ElitePartner Academics & Singles with standards; Syria edition."

The law

This publication fulfils the requirements for punishability as incitement to hatred (§ 283 (1) no. 1, second case Austrian Criminal Code – ACC). The posting at hand must be assessed in its entirety: Content-wise,

the published picture must be treated as an integral part of the publication. The picture and its caption are directed against men of Muslim faith, a protected group (criterion: religion) under § 283 ACC (Plöchl 2020: § 283 no. 8). The posting appeals to the recipients' feelings, instigates hatred against persons of Muslim faith and can evoke a strong feeling of antipathy by attributing a tendency towards sodomy to Muslim men. The pseudo-humoristic context of the posting expresses additional contempt against the target group of the posting (Plöchl 2020: § 283 no. 19).

The data

This model case is clearly perceived as hate speech: 83% of the respondents completely or somewhat agree that this posting constitutes OHS and only 3% completely or somewhat agree that this is a normal expression of opinion. Regarding punishability, this model case has an intermediate position in our case sample: 35% of the respondents (rather) agree that this posting is punishable, 55% (rather) agree that this posting should be punishable. This case prompted high ratings related to rather passive responses/strategies. 79% of the respondents (rather) agree that this posting should be deleted, 46% (rather) think that they would do nothing if they saw it online. If respondents were to take action themselves, they would most likely report the posting to the platform.

Case 2: 'Parasite scum'

Please rate the following posting. Comment on an article about religious holidays/festivities: "This parasite scum vanished 6-million-fold through chimneys without a trace."

The law

This posting does not explicitly name a group attribute ("This parasite scum […]"). It is still deducible from the context – religious festivities in the original posting – that it is directed against persons based on their religion. Without this context, a target group could not be determined. The terminology used within the posting against an identifiable, protected group of the population, is per se sufficient to fulfil the requirements of § 283 (1) no. 2 ACC. The statement constitutes verbal abuse infringing upon human dignity. This interpretation can be based on a general linguistic understanding. The term 'scum' is an intentional, disparaging designation of a part of an entirety (here: society as a whole) which is considered as inferior. The term 'parasite' signifies a lifeform which lives at the expense of another and was already used in National-Socialist propaganda in a stigmatising fashion against Jews* and other minorities (Musolff, 2011). By considering the posting as a whole, the semantic content of the incriminated expression becomes specifically obvious as a ridiculing trivialisation of the murder of Jews*. This constitutes not only a pejorative degradation, but also a statement that is able to induce feelings of hatred against persons based on their religion within the scope of § 283 (1) no. 1 second case ACC (see OGH 23.5.2018, 15 Os 33/18v = EvBl 2018/143).

The data

This case has a very high OHS rating: 95% of the respondents (rather) agree that this posting is OHS and only 3% (rather) agree that this is a normal expression of opinion. This case also has the highest punishment ratings: 78% of the respondents (rather) think this posting is already punishable, 85% think the posting should (rather) be punishable. This case has the highest ratings related to active responses/strategies. Many

respondents would (rather) report the posting to the platform (73%), to the reporting app (34%) and to the police (18%). Furthermore, 33% of the respondents would (rather) answer to this posting publicly, 94% think it should be deleted. Only 23% of the respondents would (rather) do nothing.

Case 3: 'Dirty cunt'

Comment to a female user: "Dirty cunt, you sleazy wench. Yikes, you hate-consumed cunt."

The law

This posting fulfils the requirements of criminal defamation and insult under §§ 111, 115 ACC. As required by § 111 ACC, a user is accused of possessing a despicable characteristic/disposition as being "hateful" in a manner perceivable for third parties. This accusation is able to reduce the target's reputation and esteem she enjoys among her fellow human beings (Tipold 2016: § 111 no. 4). The term "sleazy wench" may not be punishable as an accusation of dishonourable conduct or of conduct against common decency, however, it signifies contempt against the targeted person. In the context of another abusive word used in the posting, namely "dirty cunt", "sleazy wench" further serves to humiliate a female person (Rami 2021: § 115 no. 8). The posting must be assessed in its entirety. By the cited misogynistic expressions, a female user is assailed based on her gender. The underlying goal of the posting is to articulate disdain against this user as required by § 115 ACC. The gender-based insults used in the posting are hence punishable under §§ 111, 115 ACC, however, they do not reach the threshold of incitement to hatred (against women) under § 283 ACC (Tipold 2016: § 115 no. 3).

Lately, international organisations, such as the Council of Europe and its independent expert body GREVIO, have been paying closer attention to gender-based violence and OHS. In its first general recommendation on the Istanbul Convention, GREVIO noted (2021: 19): "Sexist behaviour such as sexist hate speech, which often constitutes a first step in the process towards physical violence, may also escalate to or incite overtly offensive and threatening acts, including sexual abuse or violence or rape, thus falling within the remit of Article 40 of the Istanbul Convention." Article 40 stipulates state duties to combat sexual harassment – accordingly, further regulatory action in this area may be warranted, especially given the rising numbers of women* being targeted by OHS. Without such reform, it would be desirable for courts' interpretations of § 283 in cases of hate speech against women* to take this fact into account.

The data

This case has a very high OHS rating: 96% of the respondents (rather) agree that this posting is OHS and only 4% of the respondents (rather) agree that this is a normal expression of opinion. This case also has the second highest punishability ratings: 67% of the respondents (rather) think this posting is already punishable, 82% of the respondents stated the posting should (rather) be punishable. This case has high ratings related to active responses/strategies. Many respondents would (rather) report the posting to the platform (72%) or to the reporting app (31%), 35% of the respondents would (rather) answer to this posting publicly. Besides, 97% of the respondents think the posting should be deleted and only 25% would (rather) do nothing.

Case 4: 'Male refugees'

Comment on an article dealing with questions about refugee movements: "But one may still say that the many young men are more violent than families!?"

The law

The cited posting addresses the societal group of refugees and other displaced persons and reproduces the prejudice that the flight of predominantly male persons had led to rising numbers in "foreign crime". The posting aims to instil fear and to reinforce antipathy against refugees. By claiming to ask a problematic question, the poster tries to trivialize the content and insinuates that this statement is at the margins of freedom of expression. If the contextualisation of criminality and origin/gender is intended to instil hostile sentiments against protected groups, § 283 (1) no. 1 second case ACC could be fulfilled. In the case at hand, however, the posting is not formulated sufficiently to deduce a tendentious incitement to hatred and contempt. Aversion, rejection, or contempt are not uttered to the extent legally required by § 283 ACC (Plöchl 2020: § 283 no. 19).

The data

This case has by far the lowest OHS ratings: 63 % of the respondents (rather) perceived this posting as OHS, 23% as a normal expression of opinion. This case also has the lowest punishment ratings: 13% of the respondents (rather) agree that this posting is punishable, 25% of the respondents (rather) agree that this posting should be punishable. This case has rather high ratings related to passive responses/strategies. 38% of the respondents would (rather) do nothing if they see the posting online, only 27% would report it to the platform. Interestingly, however, this posting has the highest value in terms of the respondents' response behaviour: 36% of the respondents would (rather) answer publicly to that posting.

Case 5: 'Siblings'

Comment on people with a migration history: "With most people from your area, the parents are also siblings!"

The law

The poster obviously intends to incentivise negative attitudes towards persons based on their origins and probably also their religion. Recipients could deduce from this statement that incest is common among persons with migration histories. A context between origin, and possibly also religion, and consanguinity is subtly alleged. The narrative that marriage between relatives is common in families with migration histories is commonly used by right-wing groups (Hödl, 2010; Deutscher Bundestag, 2018). These groups do not contribute to discussions on the historical roots of marriage within families or criticise such practices, but exclusively use this narrative for purposes of propaganda (Müller, 2012). This interpretation may show the underlying motivation of the poster, however, the statement per se is not sufficiently formulated to warrant for punishability under § 283 ACC (incitement to hatred) or any other provision of Austrian criminal law.

The data

This model case is clearly perceived as OHS: 85% of the respondents stated that they completely or somewhat agreed that this is OHS and 7% of the respondents completely or somewhat agreed that this is a normal expression of opinion. This case has the second lowest punishment ratings: 25% of the respondents (rather) agree that this posting is punishable, 41% of the respondents (rather) agree that this posting should be punishable. This case has rather high ratings related to passive responses/strategies. 36% of the respondents would (rather) do nothing if they see the posting online, 46% would report it to the platform and 20% would report it to the reporting app.

Cross-case analysis

Our cross-case comparison shows that, except for case 4 'male refugees', all cases were predominantly perceived as OHS and not as 'normal' expression. It is noticeable that OHS against women* or anti-Semitic content were more likely to be perceived as OHS than attacks against people with a refugee, Muslim or Syrian background. This also affects the issues of punishability and deletion of postings: The perceived necessity of OHS regulation and removal is higher in case 2 'parasite scum' and case 3 'dirty cunt' than in all other cases of the sample. Case 4 'male refugees' and case 5 'siblings' are, in contrast, not punishable under Austrian criminal law but respondents expressed a strong (ethical) desire for these content pieces to be removed. Interestingly, the punishable case 1 'virgins' only displays marginally higher ratings in perceived and preferred punishability than the (unpunishable) cases 4 and 5. From these considerations, it seems possible to derive initial indications that topics that are closer to oneself are more likely to be perceived as OHS than topics that are more distant from one's own reality of life. Most probably due to Austrian history, there is more awareness of and sensibility towards anti-Semitism than towards discrimination against Muslims to give just one example. Discrimination based on ethnicity or religion, in contrast to gender and political opinion, was hardly ever mentioned as a basis of respondents' own experiences with online discrimination. This hints towards target-group dependence of OHS perception among users. It would be beneficial to conduct further research into the dependency of receptions of target groups in societies at large vis-à-vis users' readiness to label content as illegal hate speech.

Furthermore, it is interesting to see who should regulate OHS from the user's point of view. Respondents seem to perceive that Austrian criminal law lacks rigour, as punishability ratings under the – perceived – current criminal law regime are overall lower than those of *preferred* punishability. This can be interpreted as a desire for (more) state intervention in the regulation of OHS. Geschke et al (2019) have shown a similar desire for state intervention in their representative study on OHS in Germany: 75% of respondents (N=7,349) agreed with the demand that the state should consistently enforce existing laws against insults, hate speech and defamation on the internet. Community-driven responses to OHS, i.e., counter-speech and reporting, remain at low levels in all cases. In the few instances in which a posting is reported, it is more likely to be reported to the platform itself and not to the local hate speech reporting app offered by the Antidiscrimination Office Styria or the police. Thus, in addition to the state, platforms themselves are perceived as responsible for the regulation of OHS or helpful in the fight against OHS. In general, it can also be stated that more active strategies are applied when postings are strongly perceived as OHS, as shown by the analysis of case 2 'parasite scum' and case 3 'dirty cunt'. We can hence conclude for our sample that counter-speech happens – if at all – rather after, instead of before the escalation of OHS into illegality. This is also shown by the recommendations for deleting postings: Except for case 4 'male refugees', respondents show an overall tendency towards endorsing deletion of postings regardless of how they are

perceived under criminal law. Thus, it is considered better to make content invisible than to become active and influence the discourse itself. This shows the idea of counter-speech as a mechanism against OHS lacks support among our respondents. The self-regulation of OHS among users, as in the *marketplace of ideas*, is not likely under such circumstances as also Knauder and Romanin (2021) indicated.

Conclusions, recommendations and outlook

Legally and ethically, we conclude that OHS will not be held at bay by states, platforms, or users alone, but by an interplay of these actors. Their different scopes of action can mutually reinforce each other. States can refine national norms and their application, platforms can reform their standards and reinforce moderation practices, and users can contribute with counter-speech, moderation, and, where applicable, by reporting to platforms, NGOs, and the police. We have shown that some of these responses are considered to be more effective than others by users.

Given the complexity of human expression – e.g., a multitude of languages, humour, sarcasm, and irony – especially the detection of emerging and grey area forms of OHS will need flexible, participatory models of governance. Societal climates can support or hinder the detection of OHS and illegality even among expert respondent groups, as we have shown above. Large-scale, integrated, and multi-/interdisciplinary studies on the interrelation of societal discourse and surges in OHS, as well as on the legal and ethical views of internet users, are hence needed. Conducting such studies will require transparency of platforms regarding their moderation practices. At the same time, by engaging professionals in the field as well as the public at large, media competence could be raised – and the perceived need for stricter hate speech laws, as well as expenditure in law enforcement lowered in the future. Human-rights-based, high-quality and human moderation by platforms will be key therein, as well as the participation of targets of OHS/advocacy groups to detect emerging surges in OHS. Ethics, human rights law, expert knowledge from NGOs as well as state officials and, finally, user participation by means of low-threshold tools could help, in their interplay, to realise less hate-driven social media for all – an endeavour that has become even more important during COVID-19.

References

Literature

Anderson, A. et al (2014) "The 'Nasty Effect': Online Incivility and Risk Perceptions of Emerging Technologies", *Journal of Computer-Mediated Communication,* 19:3, pp. 373–387.

Brodnig, I. (2016) *Hass im Netz. Was wir gegen Hetze, Mobbing und Lügen tun können*, Brandstätter, Wien.

Brown, A. (2020) "Models of Governance of Online Hate Speech", [online], Council of Europe, https://rm.coe.int/models-of-governance-of-online-hate-speech/16809e671d (Retrieved: November 26, 2021).

Cole, C.J. et al (2021) "INSIGHT: Social Media Reform. Can the U.S. Learn From France?", [online], Bloomberglaw, https://news.bloomberglaw.com/us-law-week/insight-social-media-reform-can-the-u-s-learn-from-france (Retrieved: November 26, 2021).

Davidson, T. et al (2017) "Automated Hate Speech Detection and the Problem of Offensive Language", *Proceedings of the International AAAI Conference on Web and Social Media 11,* pp. 512–515.

Facebook (2021a) Community Standards, [online], Facebook, https://transparency.fb.com/policies/community-standards/hate-speech (Retrieved: November 26, 2021).

Facebook (2021b) KoPl-G Transparenzbericht, [online], Facebook, https://about.fb.com/de/wp-content/uploads/sites/10/2021/10/Facebook-KoPl-G-Transparenzbericht-Oktober-2021.pdf (Retrieved: November 26, 2021).

Gadringer, S. et al (2021) *Digital News Report 2021. Detailergebnisse für Österreich*, University of Salzburg, Salzburg.

Geschke et al (2019) *#HASS IM NETZ. Der schleichende Angriff auf unsere Demokratie*, IDZ, Jena.

Gleicher et al (2021) "Adverserial Threat Report", [online], Meta, https://about.fb.com/wp-content/uploads/2021/12/Metas-Adversarial-Threat-Report.pdf (Retrieved: December 2, 2021).

Haider, I. and Millner, C. (2021) "Hasspostings im Strafverfolgungssystem", *Online Hate Speech. Perspektiven aus Praxis, Rechts- und Medienwissenschaften* (ed. G. Fischer / C. Millner / S. Radkohl), NWV, Vienna, pp. 91–139.

Hödl, K. (2010) "Sarrazin und der Zeitgeist: Sarrazin argumentiert zweifellos rassistisch", [online], DerStandard, https://www.derstandard.at/story/1282978632580/sarrazin-und-der-zeitgeist (Retrieved: December 2, 2021).

Kang, C. and Isaac, M. (2019), "Defiant Zuckerberg Says Facebook Won't Police Political Speech", [online], New York Times, https://www.nytimes.com/2019/10/17/business/zuckerberg-facebook-free-speech.html (Retrieved: November 26, 2021).

Knauder, B. and Romanin, A. (2021) "Politikwissenschaftliche Perspektive aud die Regulierung von Online Hate Speech und deren Einflüsse auf die demokratische Gesellschaft", *Online Hate Speech. Perspektiven aus Praxis, Rechts- und Medienwissenschaften* (ed. G. Fischer / C. Millner / S. Radkohl), NWV, Vienna, pp. 57–66.

Lumsden, K. and Morgan, H. (2017) "Media framing of trolling and online abuse: Silencing strategies, symbolic violence, and victim blaming", *Feminist Media Studies*, 17:6, pp. 926–940.

Müller, T. (2012) "Damit alles in der Familie bleibt", [online], Wiener Zeitung, https://www.wienerzeitung.at/nachrichten/politik/oesterreich/430859-Damit-alles-in-der-Familie-bleibt.html?em_cnt_page=1 (Retrieved: December 2, 2021).

Musolff, A. (2011) "Metaphorische *Parasiten* und 'parasitäre' Metaphern: Semantische Wechselwirkungen zwischen politischem und naturwissenschaftlichem Vokabular", *Metaphern und Gesellschaft: Die Bedeutung der Orientierung durch Metaphern* (ed. M. Junge), Springer, Wiesbaden, pp. 105–119.

Oversight Board (2021) Governance, [online], Oversight Board, https://oversightboard.com/governance/ (Retrieved: November 26, 2021).

Plöchl, F. (2020) "§ 283 StGB", *Wiener Kommentar zum Strafgesetzbuch* (2nd ed., F. Höpfel / E. Ratz), Manz, Vienna, https://rdb.manz.at (Retrieved: December 12, 2021).

Pörksen, B. (2018) *Die große Gereiztheit. Wege aus der kollektiven Erregung,* Hanser, Munich.

Rami, M. (2021) "§§ 111, 115 StGB", *Wiener Kommentar zum Strafgesetzbuch* (2nd ed., F. Höpfel / E. Ratz), Manz, Vienna, https://rdb.manz.at (Retrieved: December 12, 2021).

Sponholz, L. (2020) *Hate Speech in den Massenmedien. Theoretische Grundlagen und empirische Umsetzung*, Springer VS, Berlin.

Tipold, A. (2016) "§§ 111, 115 StGB", *Leukauf/Steininger StGB* (4th ed., C. Aichinger et al), Linde, Vienna, https://rdb.manz.at, (Retrieved: December 12, 2021).

Waseem Z. and Hovy, D. (2016) "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter", *Proceedings of the NAACL Student Research Workshop*, pp. 88–93.

Zannettou, S. et al (2020) "Measuring and Characterizing Hate Speech on News Websites", *Proceedings of the 12th ACM Conference on Web Science,* pp. 125–134.

Legal Sources and Recommendations

Austrian Ministry of Justice, Decree on the Agreement with Facebook on Deleting Hate Speech, German: *Erlass vom über die Vereinbarung mit Facebook zur Löschung von Hasspostings und Informationserteilung,* 20 July 2016, BMJ-S884.024/0014-IV/2016.

Austrian National Council, Communication Platforms Act (CPA), German*: Bundesgesetz über Maßnahmen zum Schutz der Nutzer auf Kommunikationsplattformen, Kommunikationsplattformen-Gesetz, KoPl-G,* BGBl. I Nr. 151/2020.

Austrian Supreme Court of Justice (Oberster Gerichtshof/OGH), 23.5.2018, Case no 15 Os 33/18v = EvBl 2018/143.

Council Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law, 2008/913/JHA, 28 November 2008, OJ L 328.

Deutscher Bundestag, Drucksache 19/1444 – „Schwerbehinderte in Deutschland" – Kleine Anfrage der Abgeordneten Höchst/Gminder/Pohl/Hartmann und der Fraktion der AfD, Berlin (2018).

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

GREVIO (2021), *General Recommendation No. 1 on the digital dimension of violence against women*, Council of Europe, Strasbourg, https://rm.coe.int/grevio-rec-no-on-digital-violence-against-women/1680a49147 (Retrieved: November 26, 2021).

International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) 1965, (resolution 2106 (XX)), opened for signature 21 December 1965, entered into force 4 January 1969.

Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, 2020, COM/2020/825 final.

United Nations (2020) "United Nations Strategy and Plan of Action on Hate Speech", [online], United Nations, https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf (Retrieved: November 26, 2021).

The Impact of Online Hate Speech on Muslim Women: some evidence from the UK Context

Kyriaki Topidi
EUROPEAN CENTRE FOR MINORITY ISSUES

"Technology is neither good or bad; nor is it neutral."[107]

Introduction

The recent ECRI General Policy Recommendation No.5 on preventing and combating anti-Muslim racism and discrimination (ECRI, 2022) in its preamble draws a troublesome picture on the features of anti-Muslim discrimination patterns, including in their digital dimensions. It notes 'the prevalence of false accusations affecting Muslim communities as a whole without distinguishing between Muslims and religiously disguised extremists' (ECRI, 2022, 7). It also points out the intersectional dimension of anti-Muslim hatred and the 'growing gendered abuse against Muslim women, particularly those that publicly manifest their faith' (Ibid, 7).

Starting, therefore, from continuing attempts in Europe to draw an 'essentialized' picture of Muslims as a group characterized by foreignness, backwardness, threat and cultural incompatibility with core European values (Ibid, 10), this contribution focuses on the implications of a 'racialized' representation of Muslims online, placing emphasis on women. An intersectional analysis of the impact of online hate speech is justified by the fact that this category of persons experience harm as a consequence of the interaction of their gender, religious affiliation and ethnic origin as identity markers that single them out as 'targets' for online hate speech.

Against the backdrop of the exponential growth of online hate speech targeting Muslims,[108] the aim of the analysis is to illustrate, using evidence from the UK context where data is to some extent available, the dimensions of a gendered reading of anti-Muslim hatred online, especially in connection to the harm that such speech creates. By securitizing Muslim communities, online hate speech sets that basis for serious limitations of the freedom of expression, freedom of association and political participation of a vulnerable segment of contemporary European societies (ECRI, 2022, 21).

Digital Hate and its Impact: Intersectionality as a Methodological Lens

The concept of *intersectionality* is linked to the assumption that an individual has multiple identities. As a conceptual frame, it suggests that by adopting an intersectional approach, research abandons single-axis perspectives (e.g. gender, ethnicity, religion, social class, etc.) in order to explain more comprehensively an

[107] Melvin Kranzberg, "Technology and History: 'Kranzberg's Laws", Technology and Culture, 27, No.3 (1986): 545.

[108] See indicatively the results of the 2021 Council of Europe survey of the Secretary General's Special Representative on Antisemitic, anti-Muslim and other forms of religious intolerance and hate crimes.

individual's experiences and socio-legal positionality (Crenshaw, 1991). The consideration of multiple demographic categories has been initially devised to describe ethnic minority women and their employment prospects through the notion of disadvantage and oppression. More than that, these categories have served as evidence of intersecting systems of power (in the form of racism and sexism) that create social inequalities (Collins, 2015). Previous research has addressed intersectionality as a paradigm (Bilge, 2010; Hancock, 2007), as a theory of generalized identity (Nash, 2008) as well as a methodological approach (McCall, 2005; Yuval-Davis, 2006).

Intersectionality, however, as proposed by Crenshaw, has the potential to provide a frame for two additional sets of inquiries: first, how the individual effects of discrimination on the basis of single characteristics separately (e.g. gender and ethnic origin) have cumulative effects through their intersection on right-holders; and second, the implications of applying intersectionality against the risk of over-generalisation (Hancock, 2007). More recently, on the basis of Holvino's (2010) framework of *simultaneity*, intersectional studies have proposed to take account of the multiple levels of norms, processes and structures that reproduce and maintain inequality, often in historically informed patterns of systematic disadvantage. In this light, the intersectionality of gender, ethnicity and socio-economic position are shown to create variations among women, even when they belong to the same ethnic group (Acker, 2996; Holvino, 2010). When analyzing the impact of online hate speech on Muslim women in the UK, the intersectional analytical trajectory matters because it highlights two neglected aspects: first, the intersectional process in itself (i.e. the interplay of several multiplicative factors of vulnerability such as gender, religion or race) and second, the extent to which there is conflation of intragroup differences.[109]

In general terms, the variations in which discrimination is manifested depend on the combination of various identity markers and are neither easily captured by statistical data nor are they addressed within anti-discrimination legislation. In stricter legal terms, no explicit legal references to intersectional discrimination exist within the European Convention on Human Rights, with a limited exception under Article 14 ECHR which has the potential to address intersectional discrimination.[110]

At the individual level, an intersectional lens acknowledges the violation of one's individual right to equal treatment but at the structural level, if ignored, it reinforces discrimination within and between legally protected categories. This implies that structural forms of discrimination operate on the basis of deep-seated social hierarchies and call for more contextual analysis. So far, the individual dimension of discrimination has prevailed in the European context but is inadequate to address the socio-legal impact on online hate speech on Muslim women (CIJ/ENAR, 2019, 13). This is because such an approach makes abstraction of the broader context of discrimination (i.e. the societal system that is favourable to hate speech online), as well as of the role played by other relevant actors/institutions in mapping the effects of laws and policies on the question.

The aim of the study is therefore to stress the need to highlight intragroup differences among Muslim women online against over-generalisations. Reversing or mitigating the harm of online hateful expression caused to these women, as an exercise in inclusivity, is nevertheless complex. McCall (2005, 1782) has argued

[109] Both dimensions are directly sourced from K. Crenshaw's work (1989, 1993).

[110] See also the B.S. v Spain case finding a violation of Article 3 and 14 ECHR for an isolated application of intersectionality. EU law and the CJEU are also silent on the issue. See a contrario the UN CEDAW General Recommendation No. 28, para.18 that stipulates: "The discrimination of women based on sex and gender is inextricably linked with other factors that affect women, such as race, ethnicity, religion or belief, health, status, age, class, caste and sexual orientation and gender identity. Discrimination on the basis of sex or gender may affect women belonging to such groups to a different degree or in different ways to men."

in favour of an intra-categorical approach when applying an intersectional lens because it allows to 'uncover the differences and complexities of experience' at the intersection of multiple categories. At the same time, the study will also extend the intersectional analysis beyond the 'target' subjects (i.e. Muslim women) to address the role of other powerful actors with the ability to influence the degree/type of harm caused online (McBride et al., 2015, 13). Ultimately, the justification of an intersectional approach responds to the need to engage with the 'missing' voices, moving beyond the acknowledgement of their absence in the debates. More than that, the goal is to question the relationship between race, gender and religion in the online space in order to address the processes of marginalization and discrimination at play.

In terms of identity practice, intersectionality scholarship has already established in the UK context, that regardless of their individual characteristics, visible gender and religious/ethnic markers activate stereotypes against immigrant women, in particular Muslims, emphasizing their perceived submissiveness and weakness (Hwag & Beauregard (2021), 8; Ali et al. (2017); Kamenou (2008); Wyatt and Silvester (2015)), as well as their invisibility.

The internet and female Muslim identity in the UK: A brief background note

The increase of online Islamophobic incidents and hate speech in Britain is related to the rise of far-Right discourse, as well as to Brexit and lastly to the role of media news and social media in normalizing hateful acts (Hopkins et al, 2020, 6).[111] As such, a clearer link is gradually being established in this respect between the general economic, social and political climate and the online ecosystem. The majority of female Muslims in recent surveys (80 per cent) live in wide-ranging fear of experiencing Islamophobia (physically or verbally) (Hopkins et al., 2020, 12-13), often with repercussions on their family members (e.g. the victims' children at school).

Such fear based on gendered Islamophobia is well documented in the UK context: for example, Arab or South Asian origin increases the perception of the victims as 'risks' in responses in NE England (35.3 and 25.8 per cent respectively), while wearing a headscarf or having brown skin also attracts fear among respondents (94.6 and 80.2 per cent respectively) (Hopkins et al., 2020, 14). Experiences of fear, exclusion and immobility are therefore far from uncommon among Muslim women. In that sense, online hate speech targeting Muslim women operates against an offline environment which is already characterized by tense community relations, rooted within more entrenched forms of racism and discrimination in society. The consolidation of social media in the hands of a limited number of actors has further affected the ways in which information and expressions are shared and received by their users, with implications for this particular category of users (Aguilera-Carnero and Azeez, 2016, 22).

The UK is one of the few European countries that collect data about ethnic minorities. Protection against racism in the UK extends to the online domain and therefore incitement to racial and religious hatred or violence is also illegal online (Feldman & Littler, 2014). Despite some prosecutions under the Public Order Act (1986), targeted discrimination and prejudice online are often related to 'trigger events' around which online hate speech incidents increase drastically (PRISM, UK Report, 9, stating the example of the assassination of British soldier Lee Rigby in 2013 by two Muslim men.) A second important feature relates to the active involvement of far-right parties in spreading hostile messages online, with data suggesting an

[111] The Report indicates that a majority of the Muslim respondents surveyed in North-East England (84.3 per cent) had experienced directly Islamophobia, with 15.5 per cent having experienced online abuse (Ibid, at 7).

important majority of hate speech online incidents linked to the British far right (Feldman et al., 2013: 21). The use of social media by far-right parties is essentially designed to facilitate the flow of their opinions and arguments, also through mainstream media and reach their supporters (PRISM UK Report, 11).

Those with 'visible' Muslim identity online are targeted, often repeatedly, and by the effect of such repetitive targeting, the line between the offline and online impact of hate speech becomes blurred, especially in connection to intimidation and abuse (Awan & Zempi, 2015,6). Hostile comments, racist posts, memes and images, often circulated through fake ID profiles, constitute the basic framework for hateful online expressions, particularly around the 'trigger events'.[112] The use of pictures magnifies the range of abuse and suffering of victims online (Awan & Zempi, 2015, 20). Within this context, it is worth highlighting how race and religion become interlinked in attacking Muslim women online in the UK (Awan & Zempi, 2015, 13) and end up reinforcing each other. Stigmatization, 'othering' and stereotypization happens through abusive and provocative language often including offline threats against the victims and their families. However, moving beyond the victim/perpetrator frame in the case of online hate speech, may have long term benefits in combatting online hate speech.

In the UK, religious hate crime figures declined by 5 per cent in 2020, and still, Islamophobic hate constituted 50 per cent of total incidents (Home Office, 2020). At the same time, it is becoming clearer that Islamophobic online discourses impact hate crimes, within and beyond the UK context (Levin, 2016; Mueller and Schwarz, 2020). In a sense, it can be argued that both online and offline hate speech represent the evidence of the internalization of certain cultural discourses propagated by some politicians too (Burnett, 2017). Far right and populist groups' online discourses function as enabling factors towards the creation of digital ecosystems reinforcing exclusion (Perry and Scrivens, 2018). They do so by promoting narratives of danger and untrustworthiness (Home Office, 2020).

In 2016, the Home Secretary established the National Online Hate Crime Hub and in 2017 a Home Affairs Select Committee ran an inquiry on hate crime leading to a Law Commission review seeking to address the inadequacy in legislation on online hate speech. More recently, OFCOM (2018) found in the UK an expansive development of hate speech online, with almost half of UK internet users seeing hateful content in the past year, although around 50 per cent of anti-Muslim hate was only produced by 6 per cent of users, often classified as politically anti-Islam (Demos, 2017).[113] The question on the correlation between online hate speech on race/religion with police recorded religiously aggravated crimes remains positive (Williams et al., 2020) in Britain, although it is unclear whether online hate speech precedes or follows offline hate crime.[114]

The online (technological) context also appears to perpetuate 'climates of unsafety' in the vicious cycle of victimization (Williams et al., 2020, 113). In this sense, algorithmic tools towards content moderation and classification of hate speech can only be finite in their contribution towards the reduction of hate speech, especially as cultural and linguistic context of online discourse shifts.

[112] See for example the #KillAllMuslims twitter hashtag after the Paris shooting in January 2015.

[113] According to Hope Not Hate (2019) figures, 5 out of 10 far rights social media activists with the biggest online reach in the world were British.

[114] Other factors such as demographic factors, unemployment rates, etc matter as well (Williams et al., 2020, 112).

The English Legal Framework on Online Hate Speech

Hate speech, for the purposes of the present study, is understood as "covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin." (Council of Europe, 1997).

In the digital space, however, the concept of 'cyber hate' or its co-terminous 'hate speech online' are of particular focus in the present context and refer to "any use of electronic communications technology to spread anti-Semitic, racist, bigoted, extremist or terrorist messages or information. These electronic communication technologies include the internet (i.e. websites, social networking sites, Web 2.0 user-generated content, dating sites, blogs, online games, instant messages and e-mail) as well as other computer – and cell phone-based information technologies (such as text messages and mobile phones)."[115]

It is an offence, in England and Wales, to incite hatred through hate content based on the grounds of race, religion and sexual orientation. The dissemination of the hateful material can happen through words, pictures, video music and can take a variety of forms such as messages calling for racial/religious violence, glorification of violence on the basis of race/religion or within chat forums where people are invited to commit hate crimes.

While there is no specific legal definition available on online hate speech in the British context, the Crown Prosecution Service has provided some guidance on the qualification which will determine prosecution: speech which is motivated by any form of discrimination/hostility against the victim's ethnic or national origin, gender, disability, age, religion or belief, sexual orientation or gender identity can form the object of criminal prosecution.[116] In fact, according to Home Office Data, between 2017 and 2018, there was a 40 per cent increase in online offences flagged as hate crimes. [117]

While no single piece of hate crime legislation exists, a number of offences, based on hostility aimed at one of the 'protected characteristics' (e.g. race, religion, sexual orientation)[118] provide for harsher sentences. Legislative references in this respect include the Crime and Disorder Act 1998 and the Criminal Justice Act 2003 which consider hostility as an aggravating factor of an offence.[119] A distinction can be made between hatred and hostility with the former being more serious (Williams, 2019, 35) and carrying a public order dimension through attacks targeting an entire group. No legal definition of hostility is available. Other related offences to hate speech include communications' offences and harassment (e.g. messages that are offensive/indecent, menacing or false).

---

[115] Anti-Defamation League, Responding to Cyberhate – Toolkit for Action, http://www.adl.org/internet/Bindel_final.pdf

[116] On a European level, the approach of the CPS echoes to some degree the general definition of hate speech within the Council of Europe's 1997 Recommendation (97) 20 which defines hate speech : "(...) as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin."

[117] The number of offences rose to 1605 in 2018, although the data available were treated with caution [Cf. Home Office, Hate Crime England and Wales 2017/18 Report].

[118] According to CPS, protected characteristics are identified within "[a]ny criminal offence which is perceived by the victim or any other person, to be motivated by hostility or prejudice, based on a person's disability or perceived disability; race or perceived race; or religion or perceived religion; or sexual orientation or perceived sexual orientation or a person who is transgender or perceived to be transgender" (Williams, 52).

[119] Hostility in both Acts is framed as follows: at the time of committing the offence or immediately before or after doing so, the offender demonstrated towards the victim hostility based on protected characteristics (see e.g. section 145 and 146 of the CJA 2003).

Additionally, within the Public Order Act 1986 are punished offences that stir up racial and religious hatred. Stirring religious hatred occurs when someone says something, including online, which is threatening, with the intention to stir up religious hatred. Stirring hatred, in this specific legislative context, should target an entire group, and not simply one person, and pose a threat to public order. Put simply, this offence entails more than voicing an opinion or causing offence.[120] Finally, the Communications Act 2003 and the Malicious Communications Act 1988 cover the sending of offensive, menacing, indecent or obscene messages and treat them as offences with no requirement for the communication to be received so that sending it suffices. Section 1 of the MCA 1988 criminalizes the sending of communications to another person with the purpose of causing 'distress and anxiety'. The message must also be indecent, grossly offensive, a threat or false information. Section 127 of the CA 2003 similarly carries the requirement of the message to cause 'inconvenience or needless anxiety to another' in order to qualify under the offence. The CA 2003 excludes communication sent over a 'private' network. For both Acts, the terms of the offences are generally ambiguous (e.g. the notion of 'gross offensiveness').[121] Overall, in practice, the majority of online hate speech expressions are pursued under one of the communications offences.

In criminal law terms, it is the targeted nature of (online) speech that stigmatizes its victims, along with an eventual incitement of others to hate in a threatening/abusive way, that distinguishes it from offensive or controversial speech, otherwise protected under Article 10 ECHR (Williams, 2019, 14-15). It is worth comparing this approach with the definitions of online hate speech adopted by social media platforms. For example, the 2019 Twitter Rules and Policies approach hate speech as online speech that "(…) promote(s) violence against or directly attack[s] or threaten[s] other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability or serious disease". According to the same Twitter rules "accounts whose primary purpose is inciting harm towards others on the basis of these categories'" are also not allowed. Twitter's Policy also forbids hateful imagery (e.g. symbols associated with hate groups or images depicting others as less than human or non-human). Similarly, the 2019 Facebook Community Standards approach online hate speech on the basis of three tiers of severity, according to their degree of offensive character and impact of the speech in question.[122]

In terms of conceptualization, the negative impact of online hate speech on its victims is identified in both the English criminal law context as well as within the tech companies' terms and policies of use. The effectiveness in identifying and removing as well as prosecuting such expressions, because of their harmful impact is less established. Tech companies are still failing to remove large quantities of such speech[123], while the criminal law system's response is limited as the reporting of online hate offences is low.[124] The emerging positive correlation between hate speech with religiously aggravated offences on the streets confirms the salience of the issue (Muller and Schwarz, 2017, 2018; Williams et al., 2019).

---

[120] Section 29J POA 1986 stipulates that nothing in the Act " (…) prohibits or restricts discussion, criticism or expressions of antipathy, dislike, ridicule, insult or abuse of particular religions, or the beliefs or practices shall not be taken of itself to be threatening."

[121] The Law Commission, Abusive and Offensive Online Communications: A Scoping Report, London: Crown Copyright, 2018 at 6.

[122] Tier 1 characterizes violent, dehumanizing speech; Tier 2 implies inferiority, contempt/disgust or uses slurs; Tier 3 speech excludes or negatively targets people (reference?).

[123] See the UK Digital, Culture, Media and Sport Committee's Report on 'Disinformation and "fake news"' (2019) on the limited results in preventing such damaging expressions.

[124] See the 2018 Inspectorate of Constabulary and Fire and Rescue Service Report 'Understanding the Difference: The Initial Police Response to Hate Crime' that stressed the inadequacy on online offence responses.

The Role of Platforms in spreading hate against Muslim Women

Law is not, however, an exclusive or even determinant tool for regulating online hate speech: other important parameters of regulation are market considerations, social norms and the architecture of the internet per se. The actors with the power to control what is possible and available online indirectly contribute in shaping how users behave in a more visible way. Social media -and media in general – have difficulty to depict Muslim women as neither oppressed nor dangerous (ENAR, 2016,3). Presented as not having agency, such women appear online most often as victims, usually involved in incidents or legal procedures (ENAR, 2016, 13 citing evidence from Italy, France, Denmark).

More specifically, the cycle of hatred is perpetuated by the role of social media within which are also reflected the negative representations of Muslims on broadcast television. Social media have also been blamed for allowing the normalization of xenophobia and propaganda disinformation about minority groups more generally (Evolvi, 2018; Farkas et al. 2017). The use of Twitter in this context as a platform in relation to anti-racist politics is indicative[125]: more recently, popular social media platforms have allowed the circulation of gifs and memes perpetuating stereotypes (Sharma, 2013), racism and misogyny (Ringrose,2018) likely to escape moderation and ultimately portraying minority groups as shared enemies (Wahl- Jorgensen, 2019, 110). In a sense, such use of the platforms suggests that these latter are likely to enhance the agency of groups of haters that are well-organized and with stronger ties among them (Poole et al., 2021, 1438). In addition, two dominant streams fuel this cycle further: the online discourse of a number of politicians and the political system and the contribution of far-right political views in normalizing such discourse.

When it comes to content moderation, remaining neutral is not an option for platforms: attempts to do so have led to increasing amounts of disinformation, threatening democracies and inter-group relations (Suzor, 2019, 22). The initial refusal of technology providers to deal with hate speech in its socio-legal dimension belongs, therefore, to a by-gone era. Abuse and harm contained in content hosted by the platforms in question are tightly linked to the architecture of commercial internet and cannot be exclusively placed in the remit of state criminal law systems. This happens also because, often, online abuse may not raise to the standard required to qualify as a criminal law offence and yet produces harm that leads to the exclusion of entire groups of people from opportunities offered by the internet (Suzor, 2019, 31).

Online abuse is conducive to the silencing of minority voices. While major social media networks prohibit online hate speech, as mentioned, the effectiveness of such internal rules can be questioned. From the platforms' perspective, internal rules against hate speech are either not enforced or are applied in discriminatory ways, depending on the identity/position of the speaker. Even when platforms enforce such rules, people spreading hate against minorities can choose to act in groups, instrumentalizing platforms' reporting functions to silence further minority groups (Suzor, 2019, 36-37), with women being a widespread target (Ibid, 146). As fundamentally, online hate speech in the form of name-calling, threats or bullying constitute both a cause and effect of a broader web of inequality (Suzor, 2019, 32). This is because such speech contributes and/or perpetuates the disempowerment and lack of autonomy of those that it targets.

Well-known techniques for screening content online include preemptive filtering, post-publication review and automated detection. All three of them carry the implications of either filtering insufficiently or

[125] Twitter's use is also linked to the decline of investigative journalism within mainstream media and its emergence as a source of news in itself (Poole et al., 2020, at 1417).

exceedingly online materials that impact the capacity of speech of minority groups. Without a full understanding of how platforms are moderating content, it becomes challenging to identify and reverse patterns of exclusion. The use of flagging moderation systems combined with blocking and/or filtering tools are therefore not sufficient to reverse the worrying trend. Relying on users to flag breaches of internal rules on hate speech is also not enough as in a good proportion of cases, the content flagged mirrors dominant biases in society (Suzor, 2019, 143). Ultimately, online abuse and hate speech become much more than an issue of content of classification: they are essentially the reflection of a culture of systemic discrimination that platforms should not encourage.

In connection to the last point, in the UK context, a YouGov poll found that 74 per cent of the British public knew little or nothing about Islam, with 64 per cent getting their information from the media (ENAR, 2016, Forgotten Women- UK National Report). Similarly, UK Gallup found that 30 per cent of the British public perceives the hijab as a threat and 16 per cent would not want a Muslim as a neighbour (ENAR, 2016, UK National Report). The visibility of an individual's Muslim identity, both offline and online, is key to anti-Muslim hate crime and yet the victims of online anti-Muslim hate are less inclined to report incidents to the criminal justice system (Awan & Zempi, 2020, 3).

At the same time, Facebook, Twitter and Instagram include considerable amounts of anti-Muslim hate speech against women (Awan, 2014). In light of the main features of anti-Muslim hate speech online targeting women, ICTs can no longer be treated in isolation, from both a normative and a regulatory perspective, but rather as embedded in a society (Schroder & Ling, 2013, 790). After all, the market and public pressure influence the policies of platforms (Suzor, 2019, 108), beyond strict legal considerations.

The consideration of the broader context leading to the production of online hateful speech may provide an alternative starting point to understand and then contain hate speech online. Instead of tackling individual expressions online, it could be useful to focus also on the context of online speech production, with emphasis on what an individual user is producing, sharing or engaging with (Vidgen et al., 2021, 16). A behavioural approach to online hate can be relevant when designing the responses to hate speech (e.g. quarantines, warning, counter speech, etc), especially given the limited number of 'online haters' producing the bulk of anti-Muslim hate speech. A more articulate understanding of the impact of such speech on its victims can also guide further in making decisions on resource allocation, especially in connection with victim support (Vidgen et al., 2021, 16). It should be noted nevertheless that patterns of online Islamophobic expressions vary across time, in quantity and quality, although it remains relatively consistent that a limited number of users is responsible for most hateful content online (Vidgen et al., 16).

Typology and impact of anti-Muslim hate speech: Gendered Experiences

Hate speech in general terms undermines the pursuit of social cohesion and inclusion within multicultural societies. Using vilification messages against group identities, it transforms attacks against ideas to attacks against persons. Common themes within Islamophobic content encountered online include attacks on moderate Islam on the basis of distrust towards Muslim communities in general; arguments that Muslims are not a race, but a religion and therefore vilification against them cannot be racist; or calls for the wiping out of Muslim culture (Oboler, 2013, 18-20). A basic typology of anti-Muslim hate speech, according to Oboler's categorization (2013, 14 et seq.), treats Muslims as security threats (e.g. Muslims as terrorists), as cultural threats (e.g. availability of halal food), as economic threats (e.g. Muslims as 'drains' to welfare

systems), as violent threats as well as objects of dehumanization (e.g. Muslims compared to animals) and demonization (e.g. collective allegations of criminality).

Muslim women's experiences online are equally highly contextual and varied although in direct link and continuity with the offline world. Consequently, a conceptualization of online communications that begins and ends online would be largely outmoded (Wellman & Hampton, 1999). More than that, according to Varisco (2010), "there is no pure Islamic presence, separated from other relevant forms of identity, in cyberspace any more than there is (…) in the real world (Ibid, 176).

Still, within social media, conceived as networks of interaction, anti-Muslim online hate speech acquires differentiated characteristics when compared with offline forms: anonymity, immediacy, impact and reach amplify the impact of hateful expressions (Brown, 2018). The following features, according to UNESCO (2015), differentiate online harmful expressions from offline ones: their permanence, as speech can move across platforms and be repeatedly linked; their itinerancy, across various online spaces; the possibility to be anonymous when expressing hate and finally the transnational reach of online hate speech. This differentiated impact applies especially to 'casual' forms of Islamophobia that make space for more divisive and extreme forms and expression of hate (Vidgen et al., 2021, 3). Far-right actors are particularly involved in creating 'walls of hate' (Awan, 2016),[126] although available research on online hate speech suggests that it varies across time, users, context and geography (Ganesh, 2018). This may be due to online spaces being decentralized, complex and disrupted (Margetts, John, Hale, Yasseri, 2015).

The visibility of the veil online amplifies its symbolism offline as a reminder of women's 'powerlessness, vulnerability and oppression' (Chakraborti and Zempi, 2012, 280). At the same time, the veil operates as an identification tool towards Islamophobic victimization. The issue of Muslim female 'visibility' is in itself complex: on one hand perpetrators rely on the symbolic visibility of such women to identify them as targets (e.g. visual signs such as the veil) but at the same time, online representation of the same group is victimized in largely invisible ways being represented as without agency, escaping regulation and public condemnation (Allen, 2015, 299). In sum, Muslim women's online visibility as victims of hate speech only serves the goal of 'erasure' in social value and relevance (Sayyid, 2003).

From notions of safety, physical and emotional security as well as self-worth, the impact of collective victimization has been 'escaping' the radars of legal and policy responses (Bowling, 1999; Chakraborti, 2010). In fact, Victim Support Data (2013) demonstrated that the degree of impact of crime on its victim is not necessarily commensurate to the gravity of the crime in criminal law terms. Ultimately, hate crimes, including online hate speech, are not directed towards an individual victim only but also against the community to which they are presumed to belong (Chana, 2020, 73). Due to this dual dimension, online hate speech can be framed as being beyond a reaction to 'trigger' events. It can also be considered as forming part of a wider cycle of systemic victimization, stigmatization and intimidation (Bowling, 1999, 18).

The 1997 Runnymede Trust Report on *Islamophobia: A Challenge for Us All* had identified in an earlier context the impact of Islamophobic incidents as detrimental on personal and community identities, including on one's individual and community worth (Allen, 2015, 289). Later in 2005, the Open Society Institute's Report *Muslims in the UK: Policies for Engaged Citizens* (Choudhary, 2005) highlighted the

[126] For example, among Islamophobic tweets sent between 2016/17, 15 per cent were sent by 1 per cent of the most active users and 50 per cent by the 6 per cent of most active users (Demos, 2017). See also Awan and Zempi (2017) showing how far right actors exploit virtual environment and world events to incite hatred.

emergence of stereotypes about Muslims and Islam, attributed in the public sphere to those most visible. Building on the observation of how dominant representations of Muslim women as oppressed/subjugated are, the Report showed the parallel development of a sense of justification developed among perpetrators attacking veiled Muslim women as representative of all Muslims, ignoring their individual dimensions and position as humans and subjects of legal rights/citizens.

'Islamophobia', a neologism encountered in 1996 within the Runnymede Trust established Commission on British Muslims, is premised on a monolithic view of Islam based on an essentialized interpretation of the communities as inferior and alien to Western values (Aguilera-Carnero et al., 2016, 24). It has been defined as 'indiscriminate negative attitudes or emotions directed at Islam or Muslims' (Bleich 2011). Cyber-Islamophobia, of particular interest here, is present through both blogs and social media but equally within online traditional media outlets. Cyber-hate, as such, covers abusive online material, which can and does lead in some cases to offline violence, cyber violence, cyber stalking as well as online harassment, through the use of images, videos and text. Anti-Muslim hate is premised on hostility based on a person's real or perceived Muslim religion (Awan, 2014) and is presented in the English case, as English patriotic speech which in practice dehumanizes Muslims (Awan & Zempi, 2015).

Normatively, online hate speech embodies the digital clash between freedom of expression, autonomy, human dignity and ultimately equality (Gargliardone, 2014). In other forms, it can also be ambiguously presented, without hateful language but disguising hate in its tone and/or context. The emotional and behavioural impact of online hate speech can be both short term (e.g. isolation, resentment, shame, anger) and/or long term (e.g. low self-esteem, development of prejudice against hate speaker's group, concealment of identity) (Williams, 17).

Against women in particular, the internet is creating opportunities for new ways, tools and means to perpetrate crimes (Banks, 2001, 163). The most common forms of online abuse against women include their depiction as unintelligent, hysteric or ugly (Jane, 2014), which are increasingly normalized due to lack of removal, prosecution[127] and reporting. Earlier empirical data from Twitter, for example, indicate the presence of stereotyping against Muslims (Aguilera-Carnero et al., 2016) as violently opposed to other religions as well as secularism. The use of linguistically generic references to demonize the entire community is also very present, creating new patterns of racism online (Van Dijk, 2000).

The harm that derives from online hate speech is well captured by Waldron's definition of the term as 'profound disrespect, hatred, and vilification of the members of minority groups' (Waldron, 2012). In fact, the normative contradiction between hate speech and free speech can be filtered through the focus on the harm created by certain categories of speech: Benesch in her work has introduced the concept of 'dangerous speech' precisely to express the amplifying violence by one group against another (Benesch, 2013,1). Labelled as a 'gendered crisis', affecting disproportionately women (Elmir, 2016), it is beyond the focus of this analysis, however, to debate the advantages/disadvantages of the criminalization of hate speech in the British legal order. Precisely because of its impact and given the multi-stakeholder constellation relevant for online content moderation, the online space operates as an unequal environment of suppression, with visible patterns of hierarchy of human worth. It operates against the background of 'online misogyny', a term employed to signify a form of gendered types of 'trolling' (also called e-bile) that is facilitated and

127 Online abuse against women tends to be tackled only when escalating to death threats or when abuse continues 'offline; (Jane, 2014).

normalized through digital platforms. (Ringrose, 2018).[128] According to the Law Commission Scoring Report on Abusive and Offensive Online Communication, the qualitative effect of hate speech has a wide range and includes psychological effects, emotional harms and physiological harms, exclusion from the online space, economic harms and wider societal harms affecting groups and social inclusion more broadly.[129]

Vulnerability is thus an integral part of Muslim women's lives due to the growing normalization effects of hateful expressions (and crimes) online. This leads them to operate under the weight of continuous risk assessments both online and offline in order to minimize, and if possible, avoid further harassment (ENAR, 2016, 26). Online abuse also results in attempts by victims to become less 'visible' by removing their veils (Zempi & Chakraborti, 2014; Perry & Alvi, 2012) so as to decrease their vulnerability. Perceived as attacks on their identity and with implications for their self-esteem, belonging and safety, the impact of anti-Muslim hate crime may exceed that of other types of crime (Awan & Zempi, 2015, 25). Verbal abuse and hate speech remain the most common incidents quantitatively.[130] Hate images and posts are, within this context, characterized by loaded generalisations (ENAR Report, 2016, 27). The effects of these gendered offences extend to self-censor online (with implications for free speech exercise); vulnerability; anxiety or terror (Lewis- Hastelen, 2011). This type of hate speech therefore suggests the connection between such vitriolic and harmful expression with the silencing and exclusionary patterns operating against female online (and offline) agency, self-control and ultimately empowerment (Barlow & Awan, 2016, 7). Put simply, "perpetrators (usually men) attack Muslim women not only because they belong to an ethnic, racial or religious minority, or because they are women, but because of their combined intersectional identities" (ENAR, 2016, 31).

The primary implications of virality of such content lead to self-censorship as mentioned,[131] and as crucially to the creation and perpetuation of digital infrastructures for the promotion of hate (Malik, 2017). These structures rely on both online and offline expression avoidance strategies by victims, although ironically, anti-Muslim expressions online are framed often as freedom of expression instances (Ekman, 2015). Visual representations of 'Muslimness' through clothing serve to signal potential targets for anti-Muslim hatred, within a flattening process of racialization by proxy of religious affiliation (Ganesh, 2016, 32).

Such harm caused by hate speech, according to Waldron (2012), is often left out of balancing exercises in regulatory and policy assessments of responses to such expressions. And yet, "members of the vulnerable groups targeted are expected to live their lives, conduct their business, raise their children (…) in a social atmosphere poised by their sort of speech (…)."

Moving Forward: The Law's Response and examples of Best Practices

Internet governance proposals place a disproportional focus on the legal aspects connected to states' contributions in legal rules on content moderation (Suzor, 2019, 113). Criminal law systems, constituting an integral part of the regulation of online hate speech, are less adequately designed, however, in the UK

[128] The example of Twitter is illustrative as a structure which allows public users to address each other but also do keyword and hashtag searches.

[129] The Law Commission, Abusive and Offensive Online Communications: A Scoping Report, London: Crown Copyright, 2018, at 4.

[130] Tell MAMA 2013-14 data found that 82 per cent of online incidents were about verbal abuse/hate speech.

[131] Hampton et al (2014) have found that self-censorship affects by priority those holding minority opinions.

context (and elsewhere) to address online abuse and even less the disproportionate targeting of specific groups such as women and minority religious groups.[132] The UN Special Rapporteur on violence against women noted in 2018 the "inadequate and substandard responses from intermediaries concerning online gender-based violence."[133] Moving forward, adopting human rights standards for platforms, as corporate entities, is a way to reduce risks (and negative press) while improving current exclusionary practices. In simpler terms, for platforms, taking on human rights responsibilities signifies taking responsibility for the consequences of their choices on users (Suzor, 2019, 131). This implies the introduction of monitoring systems assessing the effects of their policies and practices on the human rights of their users. Beyond privacy concerns and free speech, the impact of hate speech is fundamentally a choice on the type of equality these companies endorse. Disregarding social context and impact of hate speech shows preference for a formal type of equality, over a substantive one.

In the UK context, state supported attempts to address anti-Muslim online hate speech can be conceived on two levels: as institutional and as normative.

Institutional Responses

According to the UK Law Commission's initial 2018 report related to Offensive Online Communications,[134] online hate crime has an acknowledged strong misogynistic dimension with damaging effects which the current criminal law content does not capture (Williams, 2019, 44). The following institutional initiatives are some of the measures introduced recently to mitigate the impact of online hate speech, including on women, with varying degrees of state involvement:

- The National Online Hate Crime Hub, established in 2017, functions institutionally as a central point for the collection of all reports of online hate crime. It is designed to improve the victim experience and increase rates of successful prosecution (Williams, 2019, 43), especially given the prevailing reluctance of victims of online hate to report incidents. Alongside the Online Hate Crime Hub, the UK Council for Internet Safety (UKCIS) caters inter alia for women and girls victims of online harms and hate speech.[135]
- Third-party reporting centres have been put forward as an instrument to overcome mistrust towards the policy felt by members of religious and racial minorities and a barrier to reporting further hate crime incidents (Macpherson, 1999). Such centres were introduced to improve accessibility of the criminal justice system for all (Home Office, 2016). In practice, their establishment, granted only to existing organizations, created an asymmetric web of reporting venues with mitigated impact (Chana, 2020, 75-76).
- 'Tell MAMA' (Measuring Anti-Muslim Attacks) is a service launched in February 2012. It has been facilitated by the civil society organization Faith Matters. Tell MAMA records instances of Islamophobic hate crimes and incidents, including those happening online. It has been the first entity of its kind in Europe and is supported by the government in Britain. Victims can report

[132] The Law Commission, Abusive and Offensive Online Communications: A Scoping Report, London: Crown Copyright, 2018, at 8-9.

[133] Dubravska Simonovic, 'Report of the Special Rapporteur on Violence against Women, its Causes and Consequences on Online Violence Against Women and Girls from a Human Rights Perspective, UN General Assembly, June 14, 2018, at para. 73.

[134] The Law Commission, Abusive and Offensive Online Communications: A Scoping Report, London: Crown Copyright, 2018.

[135] https://www.gov.uk/government/organisations/uk-council-for-interntet-safety .

incidents through freephone numbers, by email, SMS, Twitter, Facebook and receive support by trained staff. Tell MAMA passes information to the police using an online reporting system.[136]

– In one of its earlier reports, covering the period between April 2012-April 2013, TellMAMA noted that 74 per cent of Islamophobic incidents registered (out of a total of 584) happened online, targeting Muslim women in considerable proportion (58 per cent of the total number of which 80 per cent concerned veiled/visibly recognizable Muslim women) (Copsey et al, 2013). The report also noted low levels of reporting of incidents to the police (with 63 per cent of all incidents going unreported). In a more recent 2017 Report, TellMAMA found the continuation of the increase on online Anti-Muslim incidents.

Normative Responses

The 2019 White Paper on Online Harm, published by the Department for Digital, Culture, Media and Sports and the Home Office proposed a new statutory duty of care for tech companies. The proposed legislation (Online Harm Bill) is intended to 'bring much needed clarity to the regulatory landscape (…)'[137] through proportionate, risk-based and tightly defined means of implementation, moving beyond liability models. This is necessary given the great diversity of online services and harms it proposes to cover. [138] To meet the 'duty of care' the companies concerned[139] will need to understand the risk of harm to individuals on their services and put in place systems to improve their safety.[140] It also requires companies to set codes of practice, establish transparency as well as effective /accessible complaint mechanisms for users and the obligation to collaborate with law enforcement in cases of illegal online harms. The current Bill also categorizes content as illegal, harmful and legal but harmful. The notions of harm and harmful content are not, however, clearly defined, although references to harm provoked through existing criminal offences (e.g. hate crime) are covered. This is likely to be insufficient as 'harmful' content is highly context dependent. The fulfilment of the tech companies' duty is to be overseen by an independent regulator, already designated as being Ofcom. The regulator in question would dispose of a range of powers against companies breaching the duty of care that include the issuing of fines,[141] the disruption of their activities and the liability of individual members of the senior management of such companies.

The draft legislation is partially responsive to online abuse and intimidation suffered by women and those from minority backgrounds,[142] as it embraces equality concerns, in particular misogyny and persecution of minorities. It has been criticized nevertheless for its implications on freedom of speech as well as for

[136] https://tellmamauk.org .

[137] UK Government, Online Harms White Paper: Full government response to the consultation, December 2020, Command Paper Number 354, Crown Copyright, at para. 17.

[138] Range of services covered include social media services, consumer cloud storage sites, video sharing platforms, online forums, dating services, online instant messaging services, peer-to-peer services, videogames which enable interaction with other users and online marketplaces.

[139] Only companies with direct control over the content and activity on a service will owe the 'duty of care'.

[140] Means of redress offered to companies are foreseen to be content removal, sanctions against offending users, reversal of wrongful content removal or sanction or changes to company processes and policies.

[141] Fines of up to 18 million pounds or 10 per cent of a company's annual turnover (whichever is higher) can be imposed.

[142] UK Government, Online Harms White Paper: Full government response to the consultation, December 2020, Command Paper Number 354, Crown Copyright, see Box 9 -Anonymous Abuse.

uncertainty on the treatment of content that is not illegal but still harmful. Evidence-based considerations of harm caused by digital content also remain underdeveloped.

References

Aguilera- Carnero, C. and Azeez, A.-H. (2016), 'Islamonausea, not Islamophobia': The many faces of cyber hate speech, *Journal of Arab and Muslim Media Research, 9*(1), 21-40.

Ali, F.; Malik, A.; Pereira, V. et al. (2017) A relational understanding of work-life balance of Muslim migrant women in the West: Future research agenda, *International Journal of Human Resource Management, 28*(8): 1163-1181.

Allen, C. (2015) 'People hate you because of the way you dress': Understanding the invisible experiences of veiled British Muslim Women victims of Islamophobia, *Int'l Review of Victimology, Vol.21* (3), 287-301.

Allen, C.; Isakjee, A.; Young, O. (2013) "Maybe we are hated": The experience and impact of anti-Muslim hate on British Muslim women, Institute of Applied Social Studies.

Awan, I. (2014) Islamophobia on Twitter: A typology of online hate against Muslims on social media, *Policy and Internet, 6*, 133-150.

Awan, I. and Zempi, I. (2015), Virtual and physical world anti-Muslim hate crime, *The British Journal of Criminology*.

Banks, K. (2001), Leave the internet alone, gender equality and ICT, *APWIN Journal*, 3, 147-173.

Barlow, C. and Awan, I. (2016) "You need to Be Sorted Out with a Knife": The Attempted Online Silencing of Women and People of Muslim Faith within Academia, *Social Media and Society,* October-December 2016: 1-11.

Benesch, S. (2013) Dangerous Speech: A proposal to prevent group violence: The dangerous speech project, http://www.worldpolicy.org/content/dangerous-speech-along-the-path-to-mass-violence

Bowling, B. (1999), Violent Racism: Victimization, Policing and Social Context, Oxford: Oxford University Press.

Burnett, J. (2017) Racial violence and the Brexit state, *Race and Class, 58*: 85-97.

Center for Intersectional Justice /ENAR, (2019) Intersectional discrimination in Europe: Relevance, Challenges and Ways Forward, https://www.intersectionaljustice.org/img/intersectionality-report-FINAL_yizq4j.pdf

Central for Intersectional Justice, Intersectionality at a Glance in Europe – Fact Sheet, April 2020, available at https://www.intersectionaljustice.org/img/2020.4.14_cij-factsheet-intersectionality-at-a-glance-in-europe_du2r4w.pdf

Chakraborti, N. and Zempi, I. (2012) The Veil under attack: Gendered dimensions of Islamophobic Victimization, *Int'l Review of Victimology, 18*(3): 269-284.

Chana, S. (2020) 'Working Towards a Better Understanding of Islamophobia', British Journal of Community Justice, Vol. 16(2), 72-91.

Chetty, N.; Altathur, S. (2018) Hate Speech review in the context of online social networks, *Aggression and Violent Behavior, 40*: 108-118.

Choudhury, T. (2005) Muslims in the UK: Policies for Engaged Citizens, London: Open Society Institute.

Copsey, N.; Dack, J.; Littler, M. et al. (2013), Anti-Muslim Hate Crime and the Far Right, Middlesbrough: Centre for Fascist, Anti-Fascist and Post-Fascist Studies.

Crenshaw, K. (1989) Demarginalizing the Intersection of Race and Sex, *University of Chicago Legal Forum,* 139-167.

Crenshaw, K. (1993) Mapping the margins: Intersectionality, identity politics and violence against women of color, *Stanford Law Review, 43*: 1241-1299.

DEMOS (2017), Anti-Islamic Content on Twitter

Ekman, M. (2015) Online Islamophobia and the politics of fear: Manufacturing the green scare, *Ethic and Racial Studies, 38*(11): 1986-2002.

Elmir, R. (2016) Muslim women bear the burden of Islamophobia, *Gulf News: The Views,* 18 September.

European Network Against Racism (ENAR) (2016), Forgotten Women: The impact of Islamophobia on Muslim women, https://www.enar-eu.org/Forgotten-Women-the-impact-of-Islamophobia-on-Muslim-women/

Evolvi, G. (2018) #Islamexit: Inter-group antagonism on Twitter, *Information, Communication and Society, 22*(3): 386-401.

Farkas, J.; Schou, J. and Neumayer, C. (2017) Cloaked Facebook pages: explaining fake Islamist propaganda in social media, *New Media and Society, 20*(5):1850-1867.

Feldman, M. and Littler, M. (2014) Tell MAMA Reporting 2013/14 – Anti-Muslim Overview, Analysis and 'Cumulative Extremism'

Gagliardone, I; Gal, D.; Aleves, T.; Martinez, G. (2015) Countering Online Hate Speech

Ganesh, B. (2016) The Geography of Anti-Muslim Hatred, Annual Report 2015, London: Tell MAMA.

Hampton, K.; Shin, I. and Lu, W. (2016) Social media and political discussion: when online presence silences offline conversation, information, communication and society, 1468

Home Office (2016), Action Against Hate: The UK Government's Plan for Tackling Hate Crime, London: Home Office.

Home Office (2020) Hate Crime, England and Wales: 2019 to 2020, https://www.gov.uk/government/statistics/hate-crime-england-and-wales-2019-to-2020

Hope Not Hate (2019), State of Hate 2019, Hope Not Hate

Hopkins, P; Clayton, T. and Tell MAMA (2020), Islamophobia and Anti-Muslim Hatred in North- East England, https://www.tellmamauk.org/wp-content/uploads/2020/06/ISLAMOPHOBIA-AND-ANTI-MUSLIM-HATRED-IN-NORTH-EAST-ENGLAND-090620.pdf

Hwang, S. and Beauregard, A.T. (2021), Contextualising intersectionality: A qualitative study of East Asian female migrant workers in the UK, *Human Relations,* 1-26.

Jane, E. (2014) "You 're an ugly, whorish, slut": Understanding e-bile, *Feminist Media Studies, 14,* 531-546.

Jubany, O. and Roiha, M., (2016) Backgrounds, Experiences and Responses to Online Hate Speech: A Comparative Cross-Country Analysis, PRISM https://sosracismo.eu/wp-content/uploads/2016/07/Backgrounds-Experiences-and-Responses-to-Online-Hate-Speech.pdf

Kamenou, N.; Netto, G. and Fearfull, A. (2013) Ethnic minority women in the Scottish labour market: Employers' perceptions, *British Journal of Management 24*(3): 398-413.

Kim, J. (2019) Ethnic capital, migration and citizenship: A Bourdieusian perspective, *Ethnic and Racial Studies, 42*(3): 357-385.

Law Commission, Abusive and Offensive Online Communications: A Scoping Report, London: Crown Copyright, 2018.

Levin, B. (2016) Special Status Report: Hate Crime in the United States, San Bernardino: California State University.

Lewis- Hasteley, H. (2011) "You should have your tongue ripped out": The reality of sexist abuse online, *New Statesman,* 3 November 2011.

Macpherson, W. (1999) The Stephen Lawrence Inquiry: Report of an Inquiry by Sir William Macpherson of Cluny, HMSO, London.

Malik, N. (2017) London won't let hate prevail, *Gulf News: The Views,* 25 March

Mc Bride, A; Hebson, G.; Holgate, V.J. (2015) Intersectionality: Are we taking enough notice in the field of work and employment relations?, *Work, Employment and Society, 29*(2), 331-341.

McCall, L. (2005) The complexity of intersectionality, *Signs – A Journal of Women in Culture and Society, 30*(3): 1771-1800.

Muller, K. and Schwarz, C. (2017), 'Fanning the Flames of Hate: Social Media and Hate Crime', Working Paper, University of Warwick.

Muller, K. and Schwarz, C. (2018) ,Making America Hate Again? Twitter and Hate Crime Under Trump', Working Paper, University of Warwick.

Muller, K. and Schwarz,C. (2020), From Hashtag to Hate Crime: Twitter and Anti-Minority Sentiment, SSRN, 1-47.

Murphy, A. (2021) Political Rhetoric and Hate Speech in the case of Shamima Begum, *Religions, 12*: 834

Oboler, A. (2013) Islamophobia on the Internet: The growth of online hate targeting Muslims, Online Hate Prevention Institute Report, IR 13-7,

https://www.researchgate.net/publication/271706783_Islamophobia_on_the_Internet_The_growth_of_online_hate_targeting_Muslims

OFCOM (2018), News Consumption in the UK: 2018, OFCOM.

Pennington, R. (2018) Making Space in Social Media: #MuslimWomensDay in Twitter, *Journal of Communication Inquiry, Vol.42*(3), 199-217.

Perry, B. and Alvi, S. (2012) "We are all Vulnerable": The Terrorism Effects of Hate Crime, *Int'l Review of Victimology, 18*(1): 57-71.

Perry, B. and Scrivens, R. (2018) A Climate for Hate? An Exploration of the Right-Wing Extremist Landscape in Canada, *Critical Criminology, 26*: 169-187.

Poole, E.; Giraud, E.H.; de Quincey, E., (2021) Tactical interventions in online hate speech: The case of #stopIslam, *New Media and Society, Vol.23*(6), 1415-1442.

Ringrose, J. (2018) Digital feminist pedagogy and post-truth misogyny, *Teaching in Higher Education 23*(5): 647-656.

Runnymede Trust (1997) Islamophobia: A Challenge for Us All, London: Runnymede Trust

Sayyid, B.S. (2003) A Fundamental Fear: Eurocentrism and the Emergence of Islamism, London: Zed Books.

Schroder, R. and Ling, R. (2013) Durkheim and Weber on the Social Implications of New Information and Communication Technologies, *New Media and Society, 16*(5), 789-805

Seglow, J. (2016) Hate speech, dignity and self-respect, *Ethical Theory and Moral Practice, 19*(5): 1103-16.

Sharma, S. (2013) Black Twitter? Racial hashtags, networks and contagion, *New Formations, 78*:46-64.

Suzor, N. (2019) Lawless, Cambridge University Press, 2019

UK Government (2020) Online Harms White Paper: Full government response to the consultation, December 2020, Command Paper Number 354, Crown Copyright

UNESCO (2015) Countering online hate speech

Van Dijk, T. (2000), New(s) Racism: A Discourse Analytical Approach, in Cottle, S. (ed.) Ethnic Minorities and the Media, Buckingham: Oxford University Press, 33-49.

Varisco, D.M. (2010) Muslims and the Media in the blogosphere, *Contemporary Islam, 4*, 157-177.

Victim Support (2013), How Crime can affect you, http://www.victimsupport.org.uk/help-for-victims/how-crime-can-affect-you

Vidgen, B.; Yasseri, T.; Margetts, H. (2021), Islamophobes are not all the same! Study of far rights actors across Twitter, *Journal of Policing, Intelligence and Counter Terrorism*

Wahl-Jorgensen, K. (2019) *Emotions, Media and Politics*, London: Polity Press.

Waldron, J. (2012) The Harm in Hate Speech, London: Harvard University Press.

Williams, M. (2019) Hatred Behind the Screens: A Report on the Rise of Online Hate Speech, Mischcon de Reya, https://hatelab.net/wp-content/uploads/2019/11/Hatred-Behind-the-Screens.pdf

Williams, M.L.; Burnap, P; Javed, A; Liu, H.; Ozalp, S. (2019), Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime, *British J, Criminol. , 60*, 93-117.

Wyatt, M. and Silvester, J. (2015) Reflections on the labyrinth: Investigating black and minority ethnic leaders' career experiences, *Human Relations 68*(8): 1243-1269.

Zempi, I. and Chakraborti, N. (2014) Islamophobia, Victimisation and the Veil, Basingstoke: Palgrave Macmillan

Part III: Protecting Rights on Platforms

Pandemics and Platforms: Private Governance of (Dis)Information in Crisis Situations

Matthias C. Kettemann and Marie-Therese Sekwenz[143]
UNIVERSITY OF INNSBRUCK, LEIBNIZ INSTITUTE FOR MEDIA RESEARCH | HANS-BREDOW-INSTITUT

Introduction

What role do online platforms play in managing and governing information during the pandemic? Chinese platforms cooperated substantially with the governments' message (and message control) on COVID-19, but also US-based platforms like Twitter and Facebook that had employed a hands-off approach to certain types of disinformation in the past invested considerably in the tools necessary to govern online disinformation more actively. Facebook, for instance, deleted Facebook events for anti-lockdown demonstrations while Twitter had to rely heavily on automated filtering (with human content governance employees back at home). This contribution will assess these practices, their impact and permanence in light of the author's research on the important role of intermediaries as normative actors, including their establishment, through terms of service and content governance practices, of a private order of public communication.

State responsibilities and private duties regarding online communication

Online just as offline, states have an obligation to respect, protect and ensure human rights for everyone on their territory or under their control.[144] This extends the duties states have from the analog world into the digital one, especially as being 'online' is now the new normal and the internet of platforms and contents is enriched by an internet of things (like smart cars) and an internet of bodies (like intelligent wearables). Even as new approaches to norm entrepreneurship online emerge,[145] rights that people have offline are still their rights in online environments.

Online just as offline, states have a primary responsibility and ultimate obligation to protect human rights and fundamental freedoms.[146] But what are these requirements international law imposes on states to ensure rights online? A key international legal basis for freedom of expression is Article 19 of the Universal Declaration of Human Rights, which is largely considered to reflect customary law. In addition, in 1976 the International Covenant on Civil and Political Rights (ICCPR) was adopted, which in its Article 19 reiterates the text of the Universal Declaration and then clarifies (in para. 2) that everyone "shall have the

[143] This contribution was first published in Matthias C. Kettemann and Konrad Lachmayer (eds.), Pandemocracy in Europe. Power, Parliaments and People in Times of Covid-19 (London: Hart, 2021).

[144] This section draws on Kettemann/Benedek, Freedom of expression online, in Mart Susi (Hrsg.), Human Rights, Digital Society and the Law. A Research Companion (London: Routledge, 2019), 58-74 and Benedek/Kettemann, Freedom of Expression on the Internet (Strasbourg: Council of Europe, 2014, 2nd ed. 2020).

[145] Radu/Kettemann/Meyer/Shahin, 'Normfare: Norm entrepreneurship in internet governance', Telecommunications Policy, Volume 45, Issue 6, 2021, https://doi.org/10.1016/j.telpol.2021.102148.

[146] Just see European Court of Human Rights, Beizaras and Levickas v. Lithuania, (Application no. 41288/15), 15 January 2020.

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."

Accordingly, the right goes beyond the freedom of the press and the freedom of the media to include individual expression in the widest sense. However, the right, with the exemption of the freedom of opinion, is not absolute or without limits. Under certain clearly defined conditions it can be restricted. In its biannual resolution on human rights on the internet in 2012, 2014 and 2016, the Human Rights Council affirmed, with references to Articles 19 of the UDHR and the ICCPR, the special role of freedom of expression online: "the same rights that people have offline must also be protected online, *in particular freedom of expression*, which is applicable regardless of frontiers and through any media of one's choice [...]."[147]

An evaluation of freedom of expression standards in international law from a European perspective (must) also consider similar regional standards such as the protections of Article 10 (1) of the European Convention on Human Rights (ECHR), enshrining "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." Note the reference to the non-interference "by public authority": States are obliged to protect freedom of expression both as a free-standing right and as an essential "enabler" of other rights through the internet. As former UN Special Rapporteur for Freedom of Expression, Frank La Rue, wrote, "by acting as a catalyst for individuals to exercise their right to freedom of opinion and expression, the internet also facilitates the realisation of a range of other human rights".[148]

The ECtHR case of K.U. v. Finland[149] confirms that states have an obligation, under the European Convention of Human Rights, to ensure that the human rights of persons under their jurisdiction are protected – offline just as online. If social network service providers fail to introduce safeguards (in the case of K.U. v. Finland, to protect the privacy rights of a child), states need to enforce a legal protection framework.[150] Just as real as the primary responsibility of states, however, is the observation that a lot of the discourse relevant for the constant opinion-forming work of democratic modernity takes place in private spaces.

The key questions regarding how to enable, moderate and regulate speech today therefore have to be asked and answered with a view to digital and private spaces.

The vast majority of communicative spaces on the internet are privately held and owned.[151] This is due to the powerful role of intermediaries, companies that enable our online activity.[152] States are therefore not the only actors in ensuring human rights online. As the 2018 Recommendation of the Council of Europe

[147] Human Rights Council Resolution 32/13, The promotion, protection and enjoyment of human rights on the Internet, UN Doc. A/HRC/RES/32/13 of 18 July 2016, para. 1 (emphasis added).

[148] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, UN Doc. A/HRC/17/27 of 16 May 2011, paras. 22 and 23. But the internet also brings about new challenges to these same human rights.

[149] ECtHR, K.U. v. Finland (2 December 2008), Application No. 2872/02.

[150] See Benedek/Kettemann, Freedom of Expression on the Internet (Strasbourg: Council of Europe, 2014, 2nd ed. 2020), pp. 92, 110.

[151] On why would need public social media, too, see Lukas B. Wieser Social Media im demokratischen Verfassungsstaat – Warum wir öffentlich-rechtliche soziale Medien brauchen. In M. Becker, M. Hofer, E. Paar, & C. Romirer (Hrsg.), Gesellschaftliche Herausforderungen – Öffentlich-rechtliche Möglichkeiten (S. 239-288). Wien: Verlag Jan Sramek.

[152] Cf. Kettemann/Schulz, Setting Rules for 2.7 Billion. A (First) Look into Facebook's Norm-Making System: Results of a Pilot Study (Hamburg: Working Papers of the Hans-Bredow-Institut, Works in Progress # 1, January 2020), https://leibniz-hbi.de/uploads/media/Publikationen/cms/media/5pz9hwo_AP_WiP001InsideFacebook.pdf.

on internet intermediaries notes, a "wide, diverse and rapidly evolving range of players, commonly referred to as "internet intermediaries", facilitate interactions on the internet between natural and legal persons by offering and performing a variety of functions and services. Some connect users to the internet, enable the processing of information and data, or host web-based services, including for user-generated content. Others aggregate information and enable searches; they give access to, host and index content and services designed and/or operated by third parties."[153]

Network effects and mergers have led to the domination of the market by a relatively small number of key intermediaries. As the 2018 Recommendation warned, these few companies have growing power: "[the] power of such intermediaries as protagonists of online expression makes it imperative to clarify their role and impact on human rights as well as their corresponding duties and responsibilities, including as regards the risk of misuse by criminals of the intermediary's services and infrastructure."[154]

Internet intermediaries have duties under international and national law. In line with the UN Guiding Principles on Business and Human Rights and the "Protect, Respect and Remedy" Framework, intermediaries should respect the human rights of their users and affected parties in all their actions. This includes the responsibility to act in compliance with applicable laws and regulatory frameworks. Internet intermediaries also develop their own rules, usually in form of terms of service or community standards that often contain content-restriction policies. This responsibility to respect within their activities all internationally recognized human rights, in line with the United Nations Guiding Principles on Business and Human Rights, exists independently of the states' ability or willingness to fulfil their own human rights obligations.[155]

States have also misused intermediaries in the past to introduce filters and enforce laws that violate international human rights commitments. Therefore, as the Recommendation notes, any norms applicable to internet intermediaries, regardless of their objective or scope of application, "should effectively safeguard human rights and fundamental freedoms, as enshrined in the European Convention on Human Rights, and should maintain adequate guarantees against arbitrary application in practice."[156]

Due to the multi-layered nature of the regulatory framework governing services provided by or through intermediaries, their regulation is challenging. As they operate in many countries and data streams, especially for cloud-based services, and often cross many countries and jurisdictions, different and conflicting laws may apply.[157] This is exacerbated by, as the 2018 Council of Europe recommendation identified, "the global nature of the internet networks and services, by the diversity of intermediaries, by the volume of internet communication, and by the speed at which it is produced and processed."[158]

In line with the UN Guiding Principles on Business and Human Rights and the 'Protect, Respect and Remedy' Framework ('Ruggie Principles'), a convincing approach posits that intermediaries need to behave in a certain way to keep their 'social licence' to operate the quasi-public sphere. Such a 'licence' necessitates

[153] Council of Europe, Recommendation CM/Rec (2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries, preambular para. 4.

[154] Ibid., preambular para. 7.

[155] Ibid., para. 2.1.1.

[156] Ibid., para. 2.1.2.

[157] Ibid., preambular para. 6.

[158] Ibid., preambular para. 9.

commitments to human rights of their users and affected parties in all their actions (including the formulation and application of terms of service) in order to address and remedy negative human rights impacts directly. For example, in order to identify and prevent adverse human rights impacts, business enterprises need to carry out human rights-due diligence. This should involve meaningful consultation with potentially affected groups and other relevant stakeholders, taking appropriate action, monitoring the effectiveness of the response and communicating their action as part of their accountability obligations.[159]

There is substantial literature on the duties of private entities in international law, especially with regard to the duties of transnational corporations[160] and private military contractors.[161] Much of it is applicable to internet standard-setters, but also to internet content companies, such as search engine providers and social networking services.[162]

Platforms in Pandemic Times

In a study[163] and subsequent analysis[164] of platform behaviour during the year of the rising Covid-19 pandemic 2020, we have identified a number of key shared commonalities among more than 40 states. Dominant platforms have been able to defend, or even solidify, their position, but communicative practices on those platforms are changing. State authorities increasingly use platforms to communicate and inform, and platforms support these approaches willingly. In the following, we look specifically at selected platforms and study their reaction to (dis)information related to Corona to assess whether we can see an emergence of a cross-platform commitment to counter Corona-related disinformation.

Facebook

During the pandemic Facebook continued to remain one of the leading platforms. With its two point seven billion daily users on its main platform alone.[165] With data traffic for messaging services, video and voice calls throughout the time of the pandemic was an important space for online speech during the pandemic.[166]

---

[159] See Ruggie J. (7 April 2008), Human Rights Council, Report of the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises, Protect, respect and remedy: a framework for business and human rights, UN Doc. A/HRC/8/5 and Guiding principles on business and human rights, implementing the United Nations "Protect, respect and remedy" framework, Annex to the Final Report of the Special Representative to the Human Rights Council, UN Doc. A/HRC/17/31 and adopted by the Human Rights Council (16 June 2011) by Resolution 17/4, Guidelines 17-21. See Benedek/Kettemann (2020), 85f.

[160] Especially after the adoption of the UN Guiding Principles on Business and Human Rights. See Radu Mares (ed.), The UN Guiding Principles on Business and Human Rights. Foundations and Implementation (Leiden: Nijhoff, 2011); and, for a comprehensive analysis, Wesley Cragg (ed.), Business and Human Rights (Cheltenham: Edward Elgar, 2012). For the international trade dimension relevant for aspects of ICTs, see Alistair M. Macleod, Human rights and international trade: normative underpinnings, in ibid., 179-196.

[161] Cf. Lindsey Cameron, Vincent Chetail, Privatizing War. Private Military and Security Companies under Public International Law (Cambridge: CUP, 2013), 288-382 (arguing that PMSCs can be bound both as companies and as the sum of their individual employees.). See also the body of scholarship cited in ibid., 269, note 22.

[162] Council of Europe, Committee of Ministers (4 April 2012), Recommendation CM/Rec(2012)3 on the protection of human rights with regard to search engines and Recommendation CM/Rec(2012)4 on the protection of human rights with regard to social networking services.

[163] Kettemann/Fertmann, 'Viral Information: How States and Platforms Deal with Covid-19-related Disinformation: an Exploratory Study of 18 Countries' (Hamburg: Verlag Hans-Bredow-Institut, 2021), GDHRNet Working Paper #1, 126.

[164] Kettemann et al., Healthy Conversations? Selected Trends in Covid-19-Related (Dis)Information Governance on Platforms, in: Kettemann/Fertmann (eds.), Viral Information: How States and Platforms Deal with Covid-19-related Disinformation: an Exploratory Study of 18 Countries (Hamburg: Verlag Hans-Bredow-Institut, 2021), GDHRNet Working Paper #1.

[165] John Clement, 'Facebook MAU Worldwide 2020' (Statista, 2020) <https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/> accessed 3 December 2020.

[166] Kiran Khan and others, 'The COVID-19 Infodemic: A Quantitative Analysis Through Facebook' (2020) 12 11.

Before the pandemic, Facebook claimed not wanting to be an "arbiter of truth".[167] While this was never accurate, and Facebook has always influenced how online communication takes place on this platform, the reaction to COVID-19 was much stronger than any other single issue addressed by automated and human content moderation.

According to the report by 'Avaaz' Facebook projected three point eight billion pieces of content that were classified as misleading health content to its users[168]. While the amount of content on the platform has increased, its content moderation was more difficult during the pandemic.[169] Because of global lockdown constraints, Facebook had to rely even more on automated content moderation[170]. Facebook also changed the community standards and defined content related to anti-vaccine statements[171], or advertising claims for medical face masks, hand sanitizer, disinfectant wipes and COVID-19- test kits, as forbidden by its terms of service which also can be seen as a shift in the company's approach.[172]

In March 2020, Facebook introduced an 'Information Hub'[173] for most users to provide health information by trusted authorities like the 'Center for Disease Control and Prevention' or the 'World Health Organization' matched with content from hand-picked journalists, politicians or other selected content about the pandemic. Facebook makes also use of pop-ups as a user-interface-design decision to additionally remind users to wear facemasks or to provide further information about the pandemic. Another information-related action was the investment of 100 million dollars to support fact-checking and journalism on the Corona crisis.[174] The financial support by Facebook also included donations for relief efforts[175], healthcare workers[176], small businesses[177] or supporting health crisis helplines.[178]

[167] Tom McCarthy, 'Zuckerberg Says Facebook Won't Be "arbiters of Truth" after Trump Threat' The Guardian (28 May 2020) <https://www.theguardian.com/technology/2020/may/28/zuckerberg-facebook-police-online-speech-trump> accessed 3 December 2020.

[168] AVAAZ, 'Facebook's Algorithm: A Major Threat to Public Health' <https://secure.avaaz.org/campaign/en/facebook_threat_health/> accessed 3 December 2020.

[169] Facebook, 'Community Standards Enforcement Report, November 2020' (About Facebook, 19 November 2020) <https://about.fb.com/news/2020/11/community-standards-enforcement-report-nov-2020/> accessed 11 December 2020.

[170] 'Keeping People Safe and Informed About the Coronavirus - About Facebook' <https://about.fb.com/news/2020/10/coronavirus/> accessed 24 November 2020.

[171] Jin Kang-Xing, 'Supporting Public Health Experts' Vaccine Efforts' (About Facebook, 19 October 2020) <https://about.fb.com/news/2020/10/supporting-public-health-experts-vaccine-efforts/> accessed 3 December 2020.

[172] Facebook, 'Information about Ads about Social Issues, Elections or Politics and COVID-19' (Facebook Business Help Center, 2020) <https://www.facebook.com/business/help/213593616543953> accessed 15 January 2021 and Facebook, 'Banning Ads and Commerce Listings for Medical Face Masks' (6 March 2020) <https://about.fb.com/news/2020/12/coronavirus/>.

[173] Salvador Rodriguez, 'Facebook Is Encouraging Everybody to Take Social Distancing Seriously' CNBC (18 March 2020) <https://www.cnbc.com/2020/03/18/coronavirus-facebook-launches-information-center-at-top-of-news-feed.html> accessed 3 December 2020.

[174] Facebook, 'Investing $100 Million in the News Industry' (30 March 2020) <https://about.fb.com/news/2020/12/coronavirus/>.

[175] Facebook, 'Matching $20 Million in Donations to Support COVID-19 Relief Efforts' (13 March 2020) <https://about.fb.com/news/2020/12/coronavirus/>.

[176] Facebook, 'Donating $25 Million to Support Healthcare Workers' (30 March 2020) <https://about.fb.com/news/2020/12/coronavirus/>.

[177] Facebook, 'Investing $100 Million in Small Businesses' (17 March 2020) <https://about.fb.com/news/2020/12/coronavirus/>.

[178] Facebook, 'Connecting People to Well-Being Tips and Resources' (9 April 2020) <https://about.fb.com/news/2020/12/coronavirus/>.

According to Kahn et al. 22,3 per cent of their investigated Facebook posts contained misinformation about COVID-19.[179] Facebook furthermore opened some data silos to the public and researchers[180] as part of the 'Data for Good' program.[181] To increase the use of this data Facebook had to further adapt its terms of service to the situation.[182] This data-support includes a COVID-19 map and dashboard with data about global symptom surveys, as well as information about datasets that mirror the movement range, or other mobility-related information of Facebooks users. This data can be used for research that e.g., takes a close look at the friendship-boundaries of Facebook users in two countries to predict the likelihood of the creation of coronavirus hotspots.[183]

Facebook had to send home content moderators on the 16th of March 2020.[184] This situation caused by the lockdown led to a high increase in artificial intelligence supported content moderation.[185] While the old moderation system was going through the amount of content chronologically, the use of a variety of algorithms (this includes machine learning approaches, filtering, ranking and sorting) now uses criteria[186] to sort through the content and prioritize it.[187] This change within the moderation system should help remove harmful content quicker than the chronological system did.

Nevertheless, Facebook remained a key platform for the spread of misinformation.[188] This claim is based on the high number of interactions related to the content in question compared to other platforms. A study also highlighted the connection between YouTube and Facebook, which are more strongly correlated through content shares than other platforms. The authors therefore come to the conclusion that misinformation is more likely to become viral if it is shared through Facebook.

Twitter

The company reports a total reach of its monetizable daily active users (mDAU) of 164 million in the first quarter of 2020, which is a growth of 23 per cent in comparison to the corresponding values in 2019.[189]

---

[179] Khan and others (n 2).

[180] Facebook, 'Data for Good: New Tools to Help Health Researchers Track and Combat COVID-19' (About Facebook, 6 April 2020) <https://about.fb.com/news/2020/04/data-for-good/> accessed 3 December 2020.

[181] Facebook, 'Our Work on COVID-19' (Facebook Data for Good) <https://dataforgood.fb.com/docs/covid19/> accessed 1 December 2020.

[182] Facebook, 'Protecting Privacy in Facebook Mobility Data during the COVID-19 Response' (Facebook Research, 3 June 2020) <https://research.fb.com/blog/2020/06/protecting-privacy-in-facebook-mobility-data-during-the-covid-19-response/> accessed 3 December 2020.

[183] Theresa Kuchler, Dominic Russel and Johannes Stroebel, 'The Geographic Spread of COVID-19 Correlates with the Structure of Social Networks as Measured by Facebook' [2020] arXiv:2004.03055 [physics, q-bio] 1.

[184] 'Keeping People Safe and Informed About the Coronavirus - About Facebook' (n 5).

[185] James Vincent, 'Facebook Is Now Using AI to Sort Content for Quicker Moderation' The Verge (13 November 2020) <https://www.theverge.com/2020/11/13/21562596/facebook-ai-moderation> accessed 15 January 2021.

[186] The criteria used are: virality, severity and how likely it is for the content to violate the Facebook Community Standards.

[187] Sílvia Majó-Vázquez and others, 'Volume and Patterns of Toxicity in Social Media Conversations during the Covid-19 Pandemic' 12.

[188] Aleksi Knuutila and others, 'Covid-Related Misinformation on Youtube' 7.

[189] Statista, 'Twitter Global MDAU 2020' (Statista) <https://www.statista.com/statistics/970920/monetizable-daily-active-twitter-users-worldwide/> accessed 17 January 2021 and Hans Rosenberg, Shahbaz Syed and Salim Rezaie, 'The Twitter Pandemic: The Critical Role of Twitter in the Dissemination of Medical Information and Misinformation during the COVID-19 Pandemic' (2020) 22 Canadian Journal of Emergency Medicine 418.

While the traffic on the platform has risen in numbers the problems via moderation, misinformation and fake news became even more problematic for COVID-19 related content.[190] Twitter took several measures to overcome the challenges of the pandemic. It supported verified information sources and tried to make them easy to access[191] in order to protect the debate on its platform.[192] Twitter strengthened its organization-relationships and fostered public engagement on its platform.[193] Twitter also focussed on the research aspects as a fourth pillar of handling the pandemic.[194] Furthermore, Twitter decided to focus on the safety of partners and employees.[195] In order to provide valuable information to its users Twitter developed a COVID-19 tab in its 'Explore'[196] function. Here users have easy access to reliable sources and hand-picked page highlights from public health experts. Through the use of verified accounts misleading speech or misinformation should be tackled on the microblogging platform.[197]

Pulido et al. found out that during the pandemic misinformation increased in presence while it is retweeted less likely, compared to scientific content or evidence, which create more engagement within the online environment. [198] The COVID-19 search prompt is another design decision Twitter took in order to curb the spread of misinformation.[199] This search prompt should also correct misspellings within the search function and promote search results from credited sources like the 'World Health Organization' in relation to COVID-19.[200] The second cluster of actions against the pandemic amplified the need of clarifying statements about misleading information and how the company deals with it.[201]

Twitter published its three key questions which are taken into consideration for COVID-19 content removal decisions, an important element of justification governance. First, 'Is the content advancing a claim

---

[190] Anatoliy Gruzd and Philip Mai, 'Going Viral: How a Single Tweet Spawned a COVID-19 Conspiracy Theory on Twitter' (2020) <https://journals.sagepub.com/doi/full/10.1177/2053951720938405> accessed 24 November 2020 and Rosenberg, Syed and Rezaie (n 40).

[191] Twitter, 'Helping People Find Reliable Information: Staying Safe and Informed on Twitter' (18 May 2020) <https://blog.twitter.com/en_us/topics/company/2020/covid-19.html> accessed 17 January 2021.

[192] Twitter, 'Protecting the Public Conversation' (14 July 2020) <https://blog.twitter.com/en_us/topics/company/2020/covid-19.html> accessed 17 January 2021.

[193] Twitter, 'Partnering with Organizations and Public Engagement' (10 April 2020) <https://blog.twitter.com/en_us/topics/company/2020/covid-19.html> accessed 17 January 2021.

[194] Twitter, 'Empowering Research of COVID-19 on Twitter' (29 April 2020) and Twitter, 'Twitter Developer Labs' (2020) <https://developer.twitter.com/en/products/labs> accessed 17 January 2021<https://blog.twitter.com/en_us/topics/company/2020/covid-19.html> accessed 17 January 2021.

[195] Jennifer Christie, 'Keeping Our Employees and Partners Safe during #coronavirus' (12 May 2020) <https://blog.twitter.com/en_us/topics/company/2020/keeping-our-employees-and-partners-safe-during-coronavirus.html> accessed 17 January 2021.

[196] Twitter, 'Coronavirus: Staying Safe and Informed on COVID-19 Tab in Explore' (18 May 2020) <https://blog.twitter.com/en_us/topics/company/2020/covid-19.html> accessed 17 January 2021.

[197] Twitter, 'COVID-19 Account Verification' (20 March 2020) <https://blog.twitter.com/en_us/topics/company/2020/covid-19.html> accessed 17 January 2021.

[198] Cristina M Pulido and others, 'COVID-19 Infodemic: More Retweets for Science-Based Information on Coronavirus than for False Information' (2020) 35 International Sociology 377.

[199] Twitter, 'Global Expansion of the COVID-19 Search Prompt' (4 March 2020) <https://blog.twitter.com/en_us/topics/company/2020/covid-19.html> accessed 17 January 2021.

[200] World Health Organization, 'World Health Organization (WHO) (@WHO) / Twitter page' (Twitter, 2020) <https://twitter.com/WHO> accessed 17 January 2021.

[201] Twitter, 'Broadening Our Guidance on Unverified Claims' (22 April 2020) <https://blog.twitter.com/en_us/topics/company/2020/covid-19.html> accessed 17 January 2021 and Twitter, 'Clarifying How We Assess Misleading Information' (14 July 2020) <https://blog.twitter.com/en_us/topics/company/2020/covid-19.html> accessed 17 January 2021.

of fact regarding COVID-19?' Second, 'Is the claim demonstrably false or misleading?' The third question risen by Twitter is: 'Would belief in this information, as presented, lead to harm?'

The first question demands the existence of more than an opinion and rather seeks for content that covers some degree of factual truth. The expression has to have the power to influence the behaviour of other users on the platform in order to fulfil the criteria Twitter has set. The second question analyses the degree of truth of the statement or otherwise it will classify the Tweet as misleading.[202] The Tweet either contains already falsified information[203] or the claim could confuse users through the process of visibility and sharing pattern.[204] The third question tries to minimize the harm that misinformation could cause through its platform. Twitter explicitly names content that could increase the likelihood of exposure to the virus or information that could lead to capacity bottlenecks within the public health system. When a Tweet meets all three of the forementioned questions and criteria Twitter grants itself the right to block or remove the content in question.

On 11 May 2020 Twitter updated its 'Terms of Service' for the placement of warning labels on Tweets that come with a reduced visibility for others.[205] Twitter's ads policy had to be renewed in order to meet the COVID-19 needs on the platform. The update restricted content that could cause panic, and content that could influence prices or the advertising of products that might be short in stock like face masks or hand sanitizers. Twitter also widened its understanding of harm on its platform.[206] Now the term also addresses speech that directly challenges the guidance from authoritative sources that contain public health information.

The first layer of the moderation process of Twitter is automated and Twitter's systems questioned one and a half million accounts that were under suspicion of amplifying COVID-19 discussion through spamming or other manipulative behaviours. Tasks related to judgement of the content itself had to be changed due to the pandemic. Twitter clarified its use of automated systems on the 16th of March 2020.[207] Twitter reported the automated surfacing of the uploaded content on its platform through the help of data trained on previous moderation decisions taken by its human moderation team. While misleading or false claims around COVID-19 often demand for additional context, the human moderation team of Twitter will take review decisions 'by hand'.[208] Twitter also informs its users of longer waiting periods for content

[202] An example given by Twitter includes statements like: „The National Guard just announced that no more shipments of food will be arriving for two months — run to the grocery store ASAP and buy everything" or "5G causes coronavirus — go destroy the cell towers in your neighbourhood!".

[203] This process of falsification is supported by subject-matter experts.

[204] Twitter gives the following examples: „Whether the content of the Tweet, including media, has been significantly altered, manipulated, doctored, or fabricated; Whether claims are presented improperly or out of context; Whether claims shared in a Tweet are widely accepted by experts to be inaccurate or false."

[205] Yoel Roth and Nick Pickles, 'Updating Our Approach to Misleading Information' (11 May 2020) <https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html> accessed 17 January 2021 and "Tweets that are labelled under this expanded guidance will have reduced visibility across the service. Reducing the visibility of Tweets means that we will not amplify the Tweets on a number of surfaces across Twitter. However, anyone following the account will still be able to see the Tweet and Retweet„.

[206] Twitter, 'Broadening Our Definition of "Harm"' (1 April 2020) <https://blog.twitter.com/en_us/topics/company/2020/covid-19.html> accessed 17 January 2021.

[207] Twitter, 'An Update on Our Content Moderation Work' (23 March 2020) <https://blog.twitter.com/en_us/topics/company/2020/covid-19.html> accessed 17 January 2021.

[208] Twitter, 'Coronavirus' (n 54).

moderation, while also giving the user a right to appeal.[209] Furthermore, Twitter announced to change its hierarchy of the global 'content severity triage system'. It now prioritizes content that might be classified as a rule violation, because this contravention is attributed as the highest risk by the platform to cause harm to its users.[210] The company also reported to have implemented a daily assurance check of its moderation system.[211] On 3 March 2020, Twitter also reminded its users of the 'zero-tolerance approach' the platform has towards manipulation.[212]

The third category of measures include the Twitter questions and answers that supported public engagement and promoted actions like 'Clapping for our healthcare heroes'[213] or '#AsktheGov'[214] where elected leaders were able to answer questions of Twitter users. Twitter announced a global software solution hackathon to fight the pandemic.[215]The company also donated one million dollars to the 'Committee to Protect Journalists' and the 'International Women's Media Foundation'.

As a further response to the crisis, Twitter tried to keep the public conversation alive while also using valuable information about the pandemic through the user data. In order to do that, Twitter created 'Twitter Developer Labs'[216] to grant access of real-time data to developers and researchers. Open research data is used for projects that take a closer look at trends and COVID-19 related discriminatory conversation.[217] There are other examples for valuable insight through Twitter's data to determine the amount or magnitude of misinformation.[218]

---

[209] Twitter, 'Appeal an Account Suspension or Locked Account' (Help Center, 2020) <https://help.twitter.com/forms/general> accessed 17 January 2021.

[210] @Vijaya and Matt Derella, 'An Update on Our Continuity Strategy during COVID-19' (16 March 2020) <https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html> accessed 17 January 2021.

[211] ibid.

[212] Twitter, 'Our Zero-Tolerance Approach to Platform Manipulation' (4 March 2020) <https://blog.twitter.com/en_us/topics/company/2020/covid-19.html> accessed 17 January 2021.

[213] Twitter, 'Clapping for Our Healthcare Heroes' (7 April 2020) <https://blog.twitter.com/en_us/topics/company/2020/covid-19.html> accessed 17 January 2021.

[214] Twitter, '#AsktheGov & #AsktheMayor Twitter Q&As' (2 April 2020) <https://blog.twitter.com/en_us/topics/company/2020/covid-19.html> accessed 17 January 2021.

[215] DEVPOST, 'COVID-19 Global Hackathon 1.0' (COVID-19 Global Hackathon 1.0, 2020) <https://covid-global-hackathon.devpost.com/> accessed 17 January 2021.

[216] Twitter, 'Twitter Developer Labs' (n 47).

[217] Maria Renee Jimenez-Sotomayor, Carolina Gomez-Moreno and Enrique Soto-Perez-de-Celis, 'Coronavirus, Ageism, and Twitter: An Evaluation of Tweets about Older Adults and COVID-19' (2020) 68 Journal of the American Geriatrics Society 1661.

[218] Matthew D Kearney, Shawn C Chiang and Philip M Massey, 'The Twitter Origins and Evolution of the COVID-19 "Plandemic" Conspiracy Theory' (2020) 1 Harvard Kennedy School Misinformation Review; Ramez Kouzy and others, 'Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter' (2020) 12 Cureus; Anna Kruspe and others, 'Cross-Language Sentiment Analysis of European Twitter Messages during the COVID-19 Pandemic', Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020 (Association for Computational Linguistics 2020) <https://www.aclweb.org/anthology/2020.nlpcovid19-acl.14> accessed 24 November 2020; Richard J Medford and others, 'An "Infodemic": Leveraging High-Volume Twitter Data to Understand Early Public Sentiment for the Coronavirus Disease 2019 Outbreak' (2020) 7 Open Forum Infectious Diseases <https://academic.oup.com/ofid/article/7/7/ofaa258/5865318> accessed 24 November 2020; Akif Mustafa, Subham Mohanta and Shalem Balla, 'Public Reaction to COVID-19 on Twitter: A Thematic Analysis' [2020] EPRA International Journal of Multidisciplinary Research (IJMR) 2455.

[218] Gautam Kishore Shahi, Anne Dirkson and TA Majchrzak, 'An Exploratory Study of COVID-19 Misinformation on Twitter' [2020] ArXiv; Gautam Kishore Shahi, Anne Dirkson and TA Majchrzak, 'An Exploratory Study of COVID-19 Misinformation on Twitter' [2020] ArXiv; Gautam Kishore Shahi, Anne Dirkson and TA Majchrzak, 'An Exploratory Study of COVID-19 Misinformation on Twitter' [2020] ArXiv and Karishma Sharma and others, 'COVID-19 on Social Media: Analyzing Misinformation in Twitter Conversations' [2020] arXiv:2003.12309 [cs].

Kouzy found that 24.8 per cent of tweets contained misinformation, while not only the tweet is of interest but also its author. Kouzy found that the rate of misinformation increased to 33.8 per cent when the author was an informal individual or posted within a group account setting. This finding is also mirrored within the usage of verified accounts. Where 31.0 per cent of the unverified accounts were classified as misinformation, while only 12.6 per cent of verified accounts contained misinformation. The company also focussed on parameters like site reliability in the pandemic due to an increase in service demand.[219] Metrics can provide a valuable insight into numbers and statistics or in this case of sentiment analysis. According to Ordun et al.[220] the information related to COVID-19 was about 50 Minutes faster retweeted compared to other Chinese networks.

Kruspe,[221] Mustafa et al.[222] and Proharel[223] used Twitter data to employee a sentiment analysis of the tweets to find out more about people's moods. But not only ordinary Twitter users are under investigation – Rufai and Bunce analysed tweets from leaders of G7 countries where the majority of tweets were classified as 'informative' content (82.8 per cent) by the researchers while the G7 leaders also used their twitter accounts to boost the moral of their citizens (nine point four per cent) of Tweets. [224]

Twitter reported to have taken into account several measures to support its employee's safety through mandatory[225] work from home whenever possible, while also assuring contractual fulfilment in cases where home office solutions are not possible.[226] In order to smoothen the change in working conditions the company also provided reimbursement toward home office related costs and additional resources for parents in the form of financial help for COVID-19 related additional day-care expenses.

YouTube

YouTube has a current user base of two billion that consumes one billion hours of content daily.[227] YouTube had some prior knowledge and experience for how to deal with pandemics.[228] Misleading information amounts to a fourth of classified COVID-19 related misleading content, which reached up to 62 million users around the globe.[229]

---

[219] @Vijaya and Derella (n 66).

[220] Catherine Ordun, S Purushotham and Edward Raff, 'Exploratory Analysis of Covid-19 Tweets Using Topic Modeling, UMAP, and DiGraphs' [2020] ArXiv 1.

[221] Kruspe and others (n 76).

[222] Mustafa, Mohanta and Balla (n 78).

[223] Bishwo Prakash Pokharel, 'Twitter Sentiment Analysis During Covid-19 Outbreak in Nepal' (Social Science Research Network 2020) SSRN Scholarly Paper ID 3624719 <https://papers.ssrn.com/abstract=3624719> accessed 24 November 2020.

[224] Sohaib Rufai and Catey Bunce, 'World Leaders' Usage of Twitter in Response to the COVID-19 Pandemic: A Content Analysis' (2020) 42 Journal of public health (Oxford, England) 1.

[225] Twitter reported on the Updated April 1, 2020 to Send home content moderators

[226] For contractors and hourly workers who are not able to perform their responsibilities from home, Twitter will continue to pay their labor costs to cover standard working hours while Twitter's work-from-home guidance and/or travel restrictions related to their assigned office are in effect. March 11, 2020

[227] YouTube, 'YouTube in Numbers' (2020) <https://www.youtube.com/intl/en-GB/about/press/> accessed 14 December 2020.

[228] Kaustubh Bora and others, 'Are Internet Videos Useful Sources of Information during Global Public Health Emergencies? A Case Study of YouTube Videos during the 2015–16 Zika Virus Pandemic' (2018) 112 Pathogens and Global Health 320.

[229] Heidi Li and others, 'YouTube as a Source of Information on COVID-19: A Pandemic of Misinformation?' (2020) 5 BMJ Global Health 1.

YouTube uses a search algorithm coupled with a recommendation system that makes use of 'collaborative filtering' in order to individually sort content according to user preferences.[230] Research in user behaviour sheds light on the importance of the ranking order of YouTube's search results. Gudivada et al. found out, that users usually only consider the top 20 search results for consumption, therefore the algorithmic recommendation of YouTube is responsible for approximately 70 per cent of content consumed by users on their platform.[231] Furthermore, Li et al. claim that during the c-Corona-crisis the content of credited sources on the platform are under-represented compared to other content creators.[232]

YouTube used several measures to curb the spread of Corona-related disinformation its platform. YouTube implemented the following key strategies: authoritative voices, providing helpful information, boosting remote learning, removing misinformation, reducing the spread of borderline content through the creation of a COVID-19 'Medical Misinformation Policy', while also providing infrastructure to its users to stay connected.[233]

With YouTube's efforts for making authoritative voices more visual, the company displayed information panels of health organisations connected to search results related to COVID-19 queries. According to YouTube, this promoted content had around 100 billion views.[234] COVID-19 related content also has high engagement, while content that also is politicized raises on average around 9000 comments for a video and factual content gained 3000 comments on average.[235] Furthermore, the consumption of news (compared to the numbers of the previous year) on the platform soared up to 75 per cent.[236] Marchal et al. found out that four-fifths of channels on YouTube sharing information are professional news agencies.[237] Nevertheless, content containing misinformation reached high volumes of shares on social media platforms and add up to the sum of shares of the five biggest English media and news sites.[238]

The company also increased the visibility of non-profit organisation and governments through free ad inventory. Another change in the user interface is the news shelf for COVID-19 related information to highlight news from authoritative sources and health agencies[239] while also building a fact-checker network that can place warning labels on content that also reduces the visibility of the video.[240]

---

[230] James Davidson and others, 'The YouTube Video Recommendation System' (2010).

[231] VN Gudivada, D Rao and J Paris, 'Understanding Search-Engine Optimization' (2015) 48 Computer 43.

[232] Li and others (n 91); Nahema Marchal, Hubert Au and Philip N Howard, 'Coronavirus News and Information on YouTube': 5 and Nahema Marchal and Hubert Au, '"Coronavirus EXPLAINED": YouTube, COVID-19, and the Socio-Technical Mediation of Expertise' (2020) 6 Social Media + Society 2056305120948158, 19.

[233] YouTube, 'Youtube Response During Coronavirus - How YouTube Works' (Youtube Response During Coronavirus - How YouTube Works, 2020) <https://www.youtube.com/howyoutubeworks/our-commitments/covid-response/> accessed 12 December 2020.

[234] ibid.

[235] Marchal, Au and Howard (n 95).

[236] Casey Newton, 'How YouTube's Moderators Are Keeping up with Changing Guidance around COVID-19' The Verge (29 April 2020) 19 <https://www.theverge.com/interface/2020/4/29/21239928/youtube-fact-check-neal-mohan-interview-misinformation-covid-19> accessed 12 December 2020.

[237] Marchal, Au and Howard (n 95).

[238] Knuutila and others (n 35).

[239] YouTube, 'Youtube Response During Coronavirus - How YouTube Works' (n 97) and Newton (n 100) 19.

[240] Ibid and Knuutila and others (n 35).

On 13 July 2020, YouTube first launched a feature called 'Depression and Anxiety Information Panels'[241] that uses information and guidelines provided by the 'Centre for Disease Control' (CDC).[242] One of the latest changes to YouTube's information channels on the 17 November now also corners content about misinformation on vaccines for COVID-19.[243] The platform started in 2019 to limit its recommendation for borderline content.[244] Borderline content makes up for around one per cent of the content on YouTube and describes cases that almost meet the criteria of deletion according to the 'Community Guidelines'.[245] Furthermore, YouTube promotes content for fundraising through a specific tag and a donation button.[246]

According to YouTube over almost eight million videos were removed by the platform between July and September this year.[247] The platform now exercises more intensive oversight over, and strives to limit the reach of, content that contains medical misinformation or discredits authoritative health authority's guidance in one of the following categories: treatment,[248] prevention,[249] Corona diagnostics[250] and/or transmission.[251]

YouTube, in contrast to Facebook, monetises COVID-19 related content.[252] This is a change in the platform's monetarisation approach that prohibited the utilisation of sensitive events it followed only month before.[253] On 16 March 2020, the company announced that it will use more automated content

---

[241] YouTube, 'Health Information Panels' (2020) <https://support.google.com/youtube/answer/9795167> accessed 14 December 2020.

[242] YouTube, 'Update to COVID-19 Information Panels' (11 June 2020) <https://support.google.com/youtube/answer/9777243?hl=en-GB> accessed 17 January 2021.

[243] YouTube, 'Update to COVID-19 Information Panels' (17 November 2020) <https://support.google.com/youtube/answer/9777243?hl=en-GB>.

[244] YouTube, 'Continuing Our Work to Improve Recommendations on YouTube' (blog.youtube, 2019) <https://blog.youtube/news-and-events/continuing-our-work-to-improve/> accessed 14 December 2020.

[245] YouTube, 'YouTube Community Guidelines & Policies - How YouTube Works' (YouTube Community Guidelines & Policies - How YouTube Works, 2020) <https://www.youtube.com/howyoutubeworks/policies/community-guidelines/> accessed 14 December 2020.

[246] Sarah Perez, 'YouTube Launches a Suite of Fundraising Tools' (TechCrunch, 2018) <https://social.techcrunch.com/2018/08/30/youtube-launches-a-suite-of-fundraising-tools/> accessed 14 December 2020 and YouTube, 'Youtube Response During Coronavirus - How YouTube Works' (n 97).

[247] 'YouTube Community Guidelines Enforcement – Google Transparency Report' <https://transparencyreport.google.com/youtube-policy/removals?hl=en> accessed 24 November 2020.

[248] YouTube gives the following examples: „Content that encourages the use of home remedies in place of medical treatment such as consulting a doctor or going to the hospital, Content that encourages the use of prayer or rituals in place of medical treatment, Content that claims that there's a guaranteed cure for COVID-19, Claims about COVID-19 vaccinations that contradict expert consensus from local health authorities or WHO, Content that claims that any currently-available medicine prevents you from getting the coronavirus" or "Other content that discourages people from consulting a medical professional or seeking medical advice".

[249] YouTube gives the following examples: „Claims that there is a guaranteed prevention method for COVID-19, Claims that an approved COVID-19 vaccine will cause death, infertility, or contraction of other infectious diseases, Claims that an approved COVID-19 vaccine will contain substances that are not on the vaccine ingredient list, such as fetal tissue, Claims that an approved COVID-19 vaccine will contain substances or devices meant to track or identify those who've received it, Claims that an approved COVID-19 vaccine will alter a person's genetic makeup, Claims that any vaccine causes contraction of COVID-19, Claims that a specific population will be required (by any entity except for a government) to take part in vaccine trials or receive the vaccine first".

[250] YouTube gives the following example: "Content that promotes diagnostic methods that contradict local health authorities or WHO".

[251] YouTube gives the following examples: „Content that claims that COVID-19 is not caused by a viral infectionContent that claims COVID-19 is not contagious, Content that claims that COVID-19 cannot spread in certain climates or geographies, Content that claims that any group or individual has immunity to the virus or cannot transmit the virus, Content that disputes the efficacy of local health authorities' or WHO's guidance on physical distancing or self-isolation measures to reduce transmission of COVID-19"; see also YouTube, 'COVID-19 Medical Misinformation Policy - YouTube Help' (n 114).

[252] YouTube, 'Monetising COVID-19-Related Content' (2 April 2020) <https://support.google.com/youtube/answer/9777243?hl=en-GB> see also Marchal, Au and Howard (n 95).

[253] Sarah Perez, 'YouTube Warns of Increased Video Removals during COVID-19 Crisis' <https://techcrunch.com/2020/03/16/youtube-warns-of-increased-video-removals-during-covid-19-crisis/> accessed 12 December 2020.

moderation and informed its platform users about the fact that more false positives and false negatives will be visible.[254] According to their enforcement report, YouTube removed 99 per cent of comments[255] through automated filtering.[256] Furthermore, YouTube defines exceptions for removal in cases of educational, documentary, scientific or artistic settings. The platform grants itself the power to remove content that violates a provision of its 'Community Guidelines', where YouTube also informs the uploading-user of the content removal per mail. Users, that violate the company's rules for the first time will only be warned, while YouTube will strike against the user's channel for further violations. When a user has reached three strikes YouTube will delete the channel.[257]

According to Priyanka et al. users are a central player in the creation or sustainment of misinformation. The authors argue that independent user content, which accounts for 11 per cent of total video content, is seven times less likely useful information about COVID-19 compared to academic institution content.[258]

The platform is a popular host for remote learning. YouTube launched 'Learn@Home', an extension to its 'Learning Hub' and is supported by several educational content creators and services like e.g., 'Khan Academy'.[259]

The removal of content on the platform is one way to target misinformation, but here the technological eco-system is more entwined than expected. In the deletion process of a video, YouTube had a longer removing time of several hours that was also viral on Facebook and Twitter.[260] According to Knuutila et al. YouTube needed 41 days to remove misleading videos that gained 149,825 views on average according to their sample.[261] As mentioned in section a.), the authors describe that the audience for misleading content of COVID-19 on YouTube is closely correlated[262] to (and on a large scale caused by) Facebook shares.

This entwined ecosystem was studied by Cinelli et al. for several platforms including YouTube.[263] The authors discovered that users have a specific timing pattern for content consumption. Furthermore, 'mainstream social media' only grants a small fraction of interaction to questionable content.[264] The questionable content on the platform can reach different degrees of visibility. In order to compare the platform's approach, the authors used the coefficient of relative amplification.[265] According to their

[254] Ibid; see also YouTube, 'Actions to Reduce the Need for People to Come into Our Offices' (Google, 16 March 2020) <https://blog.google/inside-google/company-announcements/update-extended-workforce-covid-19/> accessed 10 December 2020.

[255] The total amount of comments removed between July and September this year add up to 1,140,278,887 comments on the platform.

[256] 'YouTube Community Guidelines Enforcement – Google Transparency Report' (n 115).

[257] YouTube, 'Community Guidelines Strike Basics' (2020) <https://support.google.com/youtube/answer/2802032> accessed 14 December 2020.

[258] Priyanka Khatri and others, 'YouTube as Source of Information on 2019 Novel Coronavirus Outbreak: A Cross Sectional Study of English and Mandarin Content' (2020) 35 Travel Medicine and Infectious Disease 1.

[259] Khan Academy, 'Khan Academy' (2020) <https://www.youtube.com/user/khanacademy> accessed 14 December 2020.

[260] Statt (n 53).

[261] Knuutila and others (n 35).

[262] With a positive correlation of 0,7 for the variables „views on YouTube" and „Shares on Facebook" .

[263] The authors investigated: Twitter, YouTube, Gab, Reddit.

[264] M Cinelli and others, 'The COVID-19 Social Media Infodemic' [2020] Scientific reports 10.

[265] The coefficient of amplification is a metric to capture the amplification on a platform for the fraction of average engagement for unreliable posts to reliable posts.

findings, YouTube amplifies unreliable content less compared to reliable content with a ratio of four out of ten.

Telegram

Telegram is a Russian instant messaging service and was founded in 2013 by Pavel Durov.[266] Pavel Durov also founded the Russian social network 'VKontakte' which can be seen as a *pendant* to Facebook.[267] The service has more than 200 million[268] active users. Germany, Austria and Switzerland together account for eight million users on a daily basis.[269] The service's popularity can be explained through the one-to-many messaging option which also provides for the creation of groups reaching up to 200.000 members. Messages send within those groups can only be seen if searched for or appear within the group for every user.[270] A user can stay anonymous while posting to other users. Telegram therefore can create a wide reach for an individual user, while the user's personality can be hidden. Furthermore, the platform, in contrast to Facebook or Twitter, does not use a recommendation system nor an algorithmic timeline.[271]

The service is available within the EU or the United Kingdom for users that are 16 according to the company's terms of service.[272] Telegrams terms of service are very brief. A user has to avoid practices that: 'Use our service to send spam or scam users, promote violence on publicly viewable Telegram channels, bots, etc. or post illegal pornographic content on publicly viewable Telegram channels, bots, etc.'.[273]

Through this open formulation of the online behaviour of users, Telegram grants its online population an ample understanding of free speech. Telegram therefore is an *El Dorado* for extremist groups like the Islamic state[274] or the far right.[275] Nevertheless, Telegram announced cooperation with the EUROPOL to counter terrorist propaganda online.[276] Because of the *laissez-faire* approach the company has towards content moderation and fake news, it poses a serious threat for COVID-19 misinformation.[277]

---

[266] Anna Baydakova, 'Telegram CEO Donates 10 BTC to Pandemic Relief Effort' (CoinDesk, 28 May 2020) <https://www.coindesk.com/telegram-ceo-donates-10-btc-to-pandemic-relief-effort> accessed 12 December 2020.

[267] Katsiaryna Baran and Wolfgang Stock, 'Facebook Has Been Smacked Down. The Russian Special Way of SNSs: Vkontakte as a Case Study' (2015).

[268] Manish Singh, 'Telegram, Nearing 500 Million Users, to Begin Monetizing the App' (TechCrunch, 23 December 2020) <https://social.techcrunch.com/2020/12/23/telegram-to-launch-an-ad-platform-as-it-approaches-500-million-users/> accessed 8 January 2021.

[269] BR, 'Hildmann, Naidoo & Co.: Warum Verschwörungsfans Telegram nutzen' (BR24, 8 May 2020) <https://www.br.de/nachrichten/netzwelt/hildmann-naidoo-and-co-warum-verschwoerungsfans-telegram-nutzen,RyOCmN4> accessed 12 December 2020.

[270] Aleksi Knuutila and others, 'Junk News Distribution on Telegram: The Visibility of English-Language News Sources on Public Telegram Channels' 1.

[271] ibid.

[272] Telegram, 'Terms of Service' (Telegram) <https://telegram.org/tos> accessed 12 December 2020.

[273] ibid.

[274] Ahmad Shehabat, Teodor Mitew and Yehia Alzoubi, 'Encrypted Jihad: Investigating the Role of Telegram App in Lone Wolf Attacks in the West' (2017) 10 Journal of Strategic Security 1; Ahmet Yayla and Anne Speckhard, 'Telegram: The Mighty Application That ISIS Loves' [2017] International Center for the Study of Violent Extremism (ICSVE) 10.

[275] Alexandre Bovet and Peter Grindrod, 'The Activity of the Far Right on Telegram v2.11' (2020) 11, researchgate.net.

[276] EUROPOL, 'Europol and Telegram Take on Terrorist Propaganda Online' (Europol, 2019) <https://www.europol.europa.eu/newsroom/news/europol-and-telegram-take-terrorist-propaganda-online> accessed 8 January 2021.

[277] Knuutila and others (n 140).

Telegram has a much less strict approach to governing COVID-19 information than other major platforms. Yet, Pavel Durov started to promote verified channels[278] on his platform.[279] Those channels can be verified if an active official channel, bot or a public group is concerned and another platform (Twitter, Facebook, Instagram or YouTube) already has verified a similar account.[280] If the user has no verified account on any of those platforms, an undisputed page on Wikipedia that is in accordance with its 'Notability Guidelines'[281] also is accepted by Telegram. Ordinary user accounts cannot be verified. These are reserved for 'big and active official channels and bots'.[282] Therefore, Telegram expands their cooperation with worldwide[283] health ministries[284]. Telegram also allowed for notification of users by verified channels do address COVID-19[285].

Hui Xian Ng and Loke Jia were researching group behaviour and misinformation on Telegram in relation to the COVID-19.[286] Most activity could be measured at midday or between eight to ten pm. According to them zero point zero five per cent of overall content could be classified as misinformation. The corresponding answers to misinformation on the platform express scepticism to overall zero point four per cent. The authors found that activity within the group increased, when governments announces were made. Whereas the soar in confirmed COVID-19 cases did not influence the activity level upon the platform as much. Hui Xian Ng and Loke Jia also found, that the sentiment of the user's content could be labelled within a rather negative spectrum which correlates to governmental communication.

Conclusion and Outlook

Private Ordering of COVID-19-related Content

During the pandemic all of the platforms mentioned above took some measures related to COVID-19 while the amount of action differs. Telegram is based on a very broad understanding of free speech. Its one-to-one and one-to-few communication channels are rightly protected by law, but the groups and other one-

[278] Telegram, 'Telegram Channels' (Telegram, 29 January 2018) <https://telegram.org/tour/channels> accessed 8 January 2021.

[279] E Hacking News, 'Pavel Durov: The World Will Not Be the Same after the COVID-19 Pandemic' (E Hacking News - Latest Hacker News and IT Security News) <https://www.ehackingnews.com/2020/04/pavel-durov-world-will-not-be-same.html> accessed 12 December 2020.

[280] Telegram, 'Page Verification Guidelines' (Telegram) <https://telegram.org/verify?setln=en> accessed 8 January 2021.

[281] Wikipedia, 'Notability', Wikipedia (2020) <https://en.wikipedia.org/w/index.php?title=Wikipedia:Notability&oldid=995288718> accessed 8 January 2021.

[282] Telegram, 'Page Verification Guidelines' (n 152).

[283] Ministerio de Salud Pública de Cuba, 'Canal Oficial Del Ministerio de Salud Pública de La República de Cuba Para Ofrecer Información Sobre La #COVID19.' (Telegram) <https://t.me/MINSAPCuba> accessed 8 January 2021; Ministry of Georgia, 'StopCoV.Ge ᲒᲔ' (Telegram) <https://t.me/StopCoVge> accessed 8 January 2021; German Federal Ministry of Health, 'Corona-Infokanal Des Bundesministeriums Für Gesundheit' (Telegram) <https://t.me/Corona_Infokanal_BMG> accessed 8 January 2021; Government of India, 'MyGov Corona Newsdesk' (Telegram) <https://t.me/MyGovCoronaNewsdesk> accessed 8 January 2021; Italy Ministry of Health, 'Ministero Della Salute' (Telegram) <https://t.me/MinisteroSalute> accessed 8 January 2021 Italy Ministry of Health, 'Ministero Della Salute' (Telegram) <https://t.me/MinisteroSalute> accessed 8 January 2021; Russian Ministry of Health, 'СТОПКОРОНАВИРУС.РФ' (Telegram) <https://t.me/stopcoronavirusrussia> accessed 8 January 2021.

[284] Telegram, 'Coronavirus Info Telegram' (Telegram, 26 March 2020) </s/corona?before=23> accessed 4 January 2021 and Leong Dymples, 'Responding to COVID-19 with Telegram' (East Asia Forum, 1 May 2020) <https://www.eastasiaforum.org/2020/05/01/responding-to-covid-19-with-telegram/> accessed 12 December 2020.

[285] Telegram, 'Coronavirus News and Verified Channels' (Telegram, 2020) <https://telegram.org/blog/coronavirus> accessed 4 January 2021.

[286] Lynnette Hui Xian Ng and Yuan Loke Jia, 'Is This Pofma? Analysing Public Opinion and Misinformation in a COVID-19 Telegram Group Chat' (2020).

to-many communication facilities leave room for largely unregulated online speech which can turn problematic.[287] This gap between Telegram and the other platforms grew when measures and moderation on other social networks or messaging services became stricter. Facebook, Twitter and YouTube all have taken a selection of different means to tackle COVID-19.

The 'Organisation for Economic Co-operation and Development' (OECD) provides four recommendations to handle the pandemic: first ‚supporting a multiplicity of independent fact-checking organisations'; second, 'ensuring human moderators are in place to complement technological solutions'; third, 'voluntarily issuing transparency reports about COVID-19 disinformation'; fourth, 'improving users' media, digital and health literacy skills'.[288]

The first recommendation was *in nuce* supported by Facebook, Twitter and YouTube. The second recommendation was only partly deployed through the platforms and was not implemented when lockdowns were in place. The third recommendation was of special importance, because only with increased transparency the phenomenon of misinformation can be studied properly and tackled across platforms. The fourth recommendation is also partly employed by Facebook, Twitter and YouTube.

The European Commission also provided recommendations to digital companies.[289] It stressed the visibility of trusted content by authoritative sources, the awareness of users for content that is displayed to them, the detection of harmful content and the reduced advertising for disinformation.[290] Platforms largely incorporated the recommendations.

Misinformation can only be tackled effectively if measures are taken coherently upon platforms. With a general increase in users and views this year the platforms have a severe duty to prevent users from harm through their offered services. This increase in numbers also will lead to a gain in profit for most of the platforms. Content moderation is at the core of company's service and has changed for Facebook, Twitter and YouTube. The working conditions for moderators at Facebook are problematic especially during the pandemic. Most had to work from home or were unable to work. That is why the usage of automated systems for content moderation soared for Facebook, Twitter and YouTube. Automated systems have drawbacks compared to human content moderation and could foster the spread of misinformation online. On average 25 per cent of content relating to COVID-19 could be classified as misleading on all platforms.[291] This amount further increased up to 31 per cent when the users stayed anonymous.[292]

The recommendation algorithms employed by the platform act as 'digital curators' on platforms and are responsible for most of the content consumed by users.[293] Because the business model platforms employee user's views and reaction to content is an important key performance indicator, misleading content with

---

[287] Kettemann/Fertmann, 'Viral Information: How States and Platforms Deal with Covid-19-related Disinformation: an Exploratory Study of 18 Countries' (Hamburg: Verlag Hans-Bredow-Institut, 2021), GDHRNet Working Paper #1, 126.

[288] OECD, 'Combatting COVID-19 Disinformation on Online Platforms' (OECD, 3 July 2020) <https://www.oecd.org/coronavirus/policy-responses/combatting-covid-19-disinformation-on-online-platforms-d854ec48/> accessed 15 January 2021.

[289] European Commission, 'Disinformation: EU Assesses the Code of Practice' (European Commission - European Commission, 10 September 2020) <https://ec.europa.eu/commission/presscorner/detail/en/ip_20_1568> accessed 15 January 2021.

[290] ibid.

[291] Kouzy and others (n 75).

[292] ibid.

[293] Gudivada, Rao and Paris (n 93).

high engagement and visibility can increase the company's profit.[294] This relation between profit and polarizing content can also explain why YouTube is monetising COVID-19 content after it has banned it only a month before. Trusted sources are still under-represented and should be promoted even more on the platforms. It is important to give authoritative sources and trusted healthcare content a loud voice in the pandemic to keep misinformation at bay.

Outlook

Platforms are here to stay. Their communicative role is likely to remain influential and to even to grow, especially in developing states. Private ordering, that is the application of private norms in online spaces through which they are constituted as normative orders, will continue to be a useful concept to understand platform behaviour. States and platforms both have different duties and responsibilities vis-à-vis freedom of expression. As we have shown, private ordering has its limits: Public law is necessary in order to control public values. Privately constructed normative orders often lack a socially responsible finality. Even carefully constructed quasi-judicial entities, meant to increase legitimacy of platform law, suffer from flaws.

A basic problem of content moderation cannot be solved by even the most cleverly crafted law. It is this: While the primary responsibility for safeguarding individual spheres of freedom and social cohesion rests with states, it is platforms that have the primary power (in the sense of effective impact) to realize and influence rights and thereby cohesion. They set the rules, they design the automated tools, they delete and flag. Platforms have started to do better in terms of protecting rights, but they are still far off - in normative terms - when it comes to ensuring social cohesion.

Currently, all major platforms follow the approach of leaving as much "voice" online as possible (though overblocking happens), deleting only dangerous postings (e.g., death threats) and adding counter-statements (e.g. warnings) to problematic speech (e.g. disinformation). Covid-19 has gradually changed this, as we have seen above. For the first time, a cross-platform phenomenon became visible: the recognition that mostly lawful speech could be highly corrosive of societal values (like public health) and that platforms needed to use all tools in their normative arsenal, automatic filtering, downranking, deleting, counterinformation, flagging, to support efforts to fight Corona. If it worked overall rather well for fighting Corona, the one questions which remains is this: What about protecting other societal values against less-well designed threats? Here both more rights-conscious and more authoritarian futures are possible and continued engagement in critical platform research is essential.

[294] Svenja Boberg and others, 'Pandemic Populism: Facebook Pages of Alternative News Media and the Corona Crisis -- A Computational Content Analysis' [2020] arXiv:2004.02566 [cs] 21, 11.

Legal mechanisms for protecting freedom of expression on the internet – The Case of Serbia

Jelena Simic
FACULTY OF LAW, UNION UNIVERSITY IN BELGRADE

Introduction

We can say that the Internet is a public space, accessible to almost every individual, which allows us to participate directly in the public exchange of opinions, even in a global context, from the depths of our privacy, even conditional anonymity.[295] Participation in the exchange of information on the Internet is no longer conditioned by any intermediaries in the form of traditional, one-way media, and a high degree of interaction is achieved almost momentarily.[296] It was the U.S. Supreme Court in the 1997 case of Reno v. ACLU that singled out interactivity as one of the peculiarities of the Internet and based on it characterized the Internet as a unique and completely revolutionary medium that allows simultaneous communication, thus giving it the highest degree of protection in terms of freedom of expression.[297] That is why freedom of expression is a key element when we talk about rights and freedoms on the Internet. Compared to traditional media, it is clear that Internet users receive and communicate information without hindrance, thus blurring the line between speakers and audience. Social networks, platforms for information exchange and the Internet in general, act as a real arena of freedom almost without any restrictions such as laws, rules, regulations, which often hinder us in real life. However, just like in real life, setting boundaries is important and healing in building interpersonal relationships and respecting others. When it comes to the Internet, the question that has been in the focus of debate for a long time is, who should set those boundaries, who should regulate social networks, who should control the Internet? The "big tech" companies that own most of the social media, or the states and legislators that have set the framework for all other media so far?

Internationally, one of the triggers for this debate was certainly the events surrounding the US presidential election - from marking Trump's tweets as fake news due to non-recognition of election results, to the incursion of Trump supporters into the Capitol, as a result of which Trump's social network accounts were suspended in order to prevent him from calling for riots.[298] These events have sharpened the debate about the role of social networks in political life. When it comes to the local prism, more precisely the Republic

The author is an associate professor at the Faculty of Law, Union University in Belgrade, Serbia (e-mail: jelena.simic@pravnifakultet.rs)

[295] Martinoli, A., (ur.), Priručnik: Regulatorni okvir i poslovni modeli onlajn medija, Fondacija za otvoreno društvo, Beograd, 2019.

[296] Ibid.

[297] For more information about the Supreme Court verdict in Reno v. ACLU, 521 U.S. 844 (1997) see: (https://www.aclu.org/legal-document/reno-v-aclu-supreme-court-decision 02/01/2022).

[298] In the United States, the issue of regulation and accountability of social platforms is dealt with by the 1996 Communications Decency Act, or a part of it better known as "Section 230", which regulates the field of Internet communication. It directly releases ISPs and tech companies from liability for content posted by third parties using their communications network. For more see: Section 230: An Overview, Congressional Research Service, 7.04.2021., Available at: https://crsreports.congress.gov/product/pdf/R/R46751 12.12.2021.

of Serbia, which this paper will deal with, we can say that the whole series of events that took place in a relatively short period opened a debate on these same issues. I will list only a few of the most important events.

In March 2020, the news that the social network Twitter shut down 8,558 Twitter accounts in Serbia caused great attention of the domestic and world public. Twitter announced on its official Twitter Safety profile that their authentication teams had deleted more than 8,500 accounts used to promote the ruling party in Serbia and its leader.[299] It was a network of "bot" accounts that were engaged exclusively in promoting the ruling party and its leader during the 2017 presidential election campaign. Most of the "bots" actually retweeted and responded to the same tweet at the same time, which are clear features of an organized network of "bots" with a political plan. According to the report, these accounts tweeted more than 43 million times in four years and published more than 8 and a half million links that led to the websites of the ruling party and various pro-government media.[300]

Then, in mid-March 2021, Serbia found itself in the company of 32 countries in which, from mid-March 2021, Facebook began to apply a new advertising policy. As part of a campaign to reduce the spread of misinformation, the US company has introduced a rule that anyone who wants to advertise elections or politics in these countries must verify their identity through an identification mark issued by the country in which they want to publish ads and state "who bears liability for the ad".[301] On that occasion, the head of corporate communications at Facebook said that greater authenticity makes it harder for users to abuse the Facebook platform and increases accountability.[302]

Another controversial event regarding social networks and political parties took place in August 2021. In August 2020, the social network Twitter announced that it would start marking the accounts of media houses that are politically connected with the governments of certain countries. A year later, 11 media houses from Serbia, which have accounts on this social network, were marked as "cooperating with the Government of Serbia".[303] Among them are several of the largest media outlets in Serbia (both public and private). In a text related to the rules of use of that social network, Twitter defined state-related media as the ones "in which the state controls editorial content through financial means, direct or indirect political pressures and / or controls the production and distribution of content."[304] Labeling government-related accounts provides additional context for accounts controlled by certain government officials, government-related media outlets, and individuals who work closely with those entities. In practice, the consequence of this is that Twitter, as it is stated, will not recommend their accounts to people, nor will it "boost" the reach of their tweets. Twitter announced that these behaviors observed in Serbia violate the company's policy and represent a targeted attempt to undermine freedom of expression.[305]

---

[299] Radojević, V., 2020, Kako je radila srpska „bot" armija: 43 miliona tvitova podrške Vučiću, 4.04.2020., Raskrinkavanje, Text available here: (https://www.raskrikavanje.rs/page.php?id=Kako-je-radila-srpska-bot-armija-43-miliona-tvitova-podrske-Vucicu-642 12.12.2021.)

[300] Ibid.

[301] Komarčević D., Posebne izborne mere 'Fejsbuka', na listi i Srbija, Crna Gora i Severna Makedonija, Radio Slobodna Evropa online, 17.03.2020. Text available here: (https://www.slobodnaevropa.org/a/fejsbuk-izbori-mere-srbija-crna-gora-severna-makedonija/30492720.html 12/12/2021.)

[302] Ibid.

[303] Komčarević, D., Pet "Šta" o Twiteru u Srbiji, Radio Slobodna Evropa online, 18/08/2021. Text available here: (https://www.slobodnaevropa.org/a/srbija-twitter-laz-vesti-genocid/31416738.html 12/12/2021).

[304] Ibid.

[305] Ibid.

The last in a series of so-called controversial events took place on August 11, 2021, when Twitter confirmed that it intends to remove content denying the Srebrenica genocide from its social platforms.[306] The same intention was announced by another major Internet company - Google.[307] These decisions of Twitter to mark the media that cooperate with the state and to remove the announcements in which the Srebrenica genocide is denied, provoked violent reactions in the part of the right-wing Serbian public.

So, we can conclude that in all these events that took place in Serbia, social networks were used to help the ruling political parties or certain political structures increase their influence in creating the political will of their citizens and placing certain political ideas. At the same time, they opened the issue of the regulation and restrictions on freedom of expression on social networks, as well as the question of whether online sources of information are controlled or manipulated by the government or some other power player to emphasize certain political interests. Therefore, it seems important to become acquainted with how they work and see if it can be said that the legal mechanisms for the protection of freedom of expression on the Internet in the Republic of Serbia work. This is especially bearing in mind that the governments of the Western Balkan countries have been continuously breaching the right to freedom of expression for some time by shutting down media outlets and social movement sites, as well as intimidating certain activists on the Internet.[308] Such measures undermine the foundations of democracy and hamper civic activism, which is often on shaky ground in this part of the world. The following section will therefore review the regulations of the Republic of Serbia dealing with freedom of expression on the Internet, with special reference to the review of mechanisms for legal protection against hate speech on the Internet, as a particularly important topic in political and social education.

Challenges of determining legal framework

Numerous authors have written about the challenges of protecting human rights and freedom of expression in the new information environment. (Gregg, 2012; Hick, Halpin & Hoskins, 2016; Khor, 2011).[309] This topic is extremely important because the threat to the right to free expression on the Internet occurs in different forms and can be violated by different agents. Sometimes this violation is clear and unambiguous, such as when we talk about authoritarian governments, which can even block access to the Internet or certain content infrastructurally, or when, at the request of the government, private companies block or filter content.[310] However, violating this right can be even more sophisticated. For example, governments (often

[306] In fact, the Initiative to remove the content denying the murder of a large number (8000) of Bosniaks in Srebrenica in 1995 originated from the Genocide Research Institute of Canada (IGK). According to their data, most of the reports insulting Srebrenica victims come from Serbia, but there are also reports from Russia, France and other countries. On August 11, a Twitter spokesman said in a written response to a media outlet in Serbia that hate speech and messages "have absolutely no place" on the social network. Komčarević, D., Five "Whats" about Twitter in Serbia, Radio Free Europe online, 18/08/2021., Text available here: (https://www.slobodnaevropa.org/a/srbija-twitter-laz-vesti-genocid/31416738.html 12/12/2021).

[307] Ibid.

[308] For more information, see: Internet freedoms in the Western Balkans, Share Foundation, Civil Rights Defenders, (https://crd.org/wp-content/uploads/2020/04/SRB_Saz%CC%8Cetak_Slobode-na-internetu.pdf 01/02/2022) as well as the Share Foundation Report: The State of Digital Rights and Freedoms in Serbia: (https://resursi.sharefoundation.info/sr/resource/monitoring-digitalnih-prava-i-sloboda-na-kraju-godine-tehnicki-napadi-ponovo-u-fokusu/ 01/02/2022).

[309] Mitrović, M., Sloboda izražavanja i zaštita podataka o ličnosti na internetu: Perspektiva internet korisnika u Srbiji, CM Komunikacija i mediji 15 (47), 5-34, Beograd, 2020. (https://scindeks-clanci.ceon.rs/data/pdf/2466-541X/2020/2466-541X2047005M.pdf , 20/12/2021.)

[310] For example, one of the most controversial sets of laws - Yarovaya law came into force in 2018 in Russia. This set of legal changes is actually a set of amendments that change a dozen laws that significantly apply to freedom on the Internet. While the government sees the law as an anti-terrorism measure aimed at increasing user security, Internet activists have sharply criticized the law and called it Russia's Big Brother Law. See more: (Freedom House Report: Russia, 2018, Available at: https://freedomhouse.org/country/russia/freedom-net/2018 12.12.2021.). Also, in

even democratic ones) monitor the activities of citizens on the Internet, while learning about it leads citizens to self-censorship, i.e. the fear of Internet users to publish critical content about the work of their governments, for fear of retaliation and sanctions.[311] Or, the ability of private companies on the Internet, such as social networks, to manage information, or to favor certain content for commercial or other interests.[312]

In response to such growing, negative trends, and in order to combat terrorism and extremism of all kinds, including those promoted on social media, the EU is preparing a new set of laws called the "Digital Services Act"[313] – that is, an act on digital services, which envisages a number of legal acts that will regulate, among other things, speech on social networks. If this package of laws is adopted, all member states will have to apply the new rules. Among their provisions are penalties for social networks if they do not remove any content containing elements of extremism from their platforms within 60 minutes of its publication.[314] This could be implemented at the level of the whole EU in the same way as the GDPR brought about changes in the legislation for all member states in the field of data privacy protection. A model for such regulation specifically in the area of hate speech is the law passed by Germany - the Network Enforcement Act (NetzDG).[315] This law stipulates that social networks must establish a mechanism for quick and easy reporting of inappropriate or harassing posts, and that they must be removed as soon as possible, under the threat of high fines.[316] A similar law was passed by France in 2019, using growing hate speech and misinformation as an argument for tighter regulation of networks.[317] However, the law has been deemed unconstitutional in 2020.

2017, Germany introduced the NetzDG (Network Enforcement Act), the so-called "Law on Facebook", which is based on the fight against hate speech and false news, but critics consider it a law that seriously violates freedom of expression. Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz - NetzDG) /Act to Improve Enforcement of the Law in Social Networks. Available at: (https://perma.cc/RW47-95SR 10/12/2021)

311 Mitrović, M., (2020), pp.6-8.

312 In addition to the well-known techniques of favoring sponsored content on Internet platforms, sometimes content can be favored for political and even experimental reasons. Let's remember the affair that in 2012 brought into question the trust in Facebook, when a random sample of more than 600 thousand people was divided into two groups. For one group, NewsFeed was set to show only positive posts, while the other group was only exposed to negative posts (statuses, photos, news, etc.). At the same time, the activities of users who did not know or agree to be respondents in this psychological experiment were monitored to determine whether positive / negative posts will have an impact on their activities and behavior on this social network. Such easy-going approach to experimenting on humans has raised many questions, from the ethics of such a procedure to the need for explicit user consent, to the limitless ability of Facebook to manipulate the results presented in NewsFeed, which directly affects freedom of expression. For more info, see: Chambers, C. (2014). Facebook fiasco: Was Cornell's study of 'emotional contagion' an ethics breach? The Guardian, 30/06/2014 (https://www.theguardian.com/technology/2014/jun/30/facebook-emotion-study-breached-ethical-guidelines-researchers-say 12/12/2021) in: Mitrović, M., (2020), p.8.

313 Digital Service Act - The Digital Services Act significantly improves the mechanisms for the removal of illegal content and for the effective protection of users' fundamental rights online, including the freedom of speech. It also creates a stronger public oversight of online platforms, in particular for platforms that reach more than 10% of the EU's population. For more info visit European Parliament website: (https://www.europarl.europa.eu/news/en/press-room/20220114IPR21017/digital-services-act-regulating-platforms-for-a-safer-online-space-for-users 12/12/2021).

314 Popović Aleksandra, Ko uređuje društvene mreže – zakonodavstvo SAD, EU i Srbije, Talas.rs, 22/01/2021, (https://talas.rs/2021/01/22/ko-ureduje-drustvene-mreze-zakonodavstvo-sad-eu-i-srbije/ 12/12/2021).

315 Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz - NetzDG) /Act to Improve Enforcement of the Law in Social Networks. Available at: (https://perma.cc/RW47-95SR 10/12/2021)

316 Ibid.

317 France online hate speech law to force social media sites to act quickly, The Guardian, Agence France-Presse in Paris, 9/07/2019. Available at: (https://www.theguardian.com/world/2019/jul/09/france-online-hate-speech-law-social-media 10/02/2022).

Legal framework in the Republic of Serbia

The European Convention on Human Rights and Fundamental Freedoms, as well as the recommendations of the Council of Europe, emphasize that member states are responsible for implementing human rights standards and fundamental freedoms on the Internet.[318] In addition, the member states of the Council of Europe, including the Republic of Serbia (here and below: RS), have an obligation to their citizens to respect, protect and promote human rights and fundamental freedoms on the Internet. The main parameters of Internet freedom according to the Recommendations of the Council of Europe CM / Rec (2016) 5, refer to: (1) creating an environment in which freedom on the Internet is possible, (2) the right to freedom of expression, (3) the right to freedom of peaceful assembly and association, (4) the right to private and family life.[319] The legal framework of the Republic of Serbia when it comes to "freedom on the Internet" will be analyzed in the context of the first two mentioned parameters.

Creating an environment where freedom on the Internet is possible

Free access to the Internet or internet neutrality is one of the extremely important principles of the functioning of the Internet. We can say that internet neutrality in the narrow sense is the principle of Internet functioning, according to which internet service providers should treat all data on the network in the same way, without discrimination based on content, type of platform - website and application, communication methods, etc.[320] Specifically, they should not in any way favor a particular type of data in their content or source over other websites.[321]

If we take the countries of the Western Balkans as an example, we can see that although some countries guarantee the impartiality of the network by law, others do not deal with this topic so explicitly.[322] In Serbia, the Internet is relatively free and open, and the access rate to the Internet is high.[323] According to the Serbian Statistical Office data for 2020, 74.3% of households own a computer, while 94.1 own a mobile phone.[324] In 2020, 81% of Serbian households had an Internet connection, and Internet services are relatively affordable. According to the annual data of the Statistical Office of Serbia for 2020, there are significant differences in terms of Internet connectivity between urban centers, where 87.1% of the population has Internet access, and smaller settlements and villages, where 70.4% of the population has Internet access.[325] Although the Law on Electronic Communications prescribes that "a set of basic electronic communications services features a certain scope and quality, available to all in the territory of the Republic

[318] Internet freedoms report 2020, Civil Rights Defenders and Share Foundation, Mirkovic, N., and Merrell, F., (ed.), p. 4, Available here: (https://crd.org/wp-content/uploads/2020/04/200402_GRA_InternetFreedoms_Narativa_A4_Spreads.pdf 1/02/2022).

[319] Recommendation CM/Rec (2016)5[1] of the Committee of Ministers to member States on Internet freedom, Available here: https://mediainitiatives.am/wp-content/uploads/2017/03/Recommendation-of-the-Committee-of-Ministers-on-Internet-Freedom-in-English.pdf 1/02/2022).

[320] Milić, D., Neutralnost interneta u pravu Republike Srbije, MillicLawOffice, 8/05/2019, Full text available here: (https://www.milic.rs/blog/internet-pravo/neutralnost-interneta-u-pravu-republike-srbije/ 10/12/2021).

[321] Ibid.

[322] Internet freedoms report 2020, Civil Rights Defenders and Share Foundation, Mirkovic, N., and Merrell, F., (ed.), p. 7, Available at: (https://crd.org/wp-content/uploads/2020/04/200402_GRA_InternetFreedoms_Narativa_A4_Spreads.pdf 1/02/2022).

[323] Freedom of the Net, Serbia – 2021, Freedom House Report, Available online at: https://freedomhouse.org/sr/country/serbia/freedom-net/2021 01/2/2022).

[324] Ibid.

[325] Ibid.

of Serbia at affordable prices", the State Regulatory Agency for Electronic Communications and Postal Services (RATEL) in its Digital Inclusion Report for 2019 states that providers have not built the necessary infrastructure, because it would have limited economic sustainability in areas with less affluent populations".[326] Therefore, availability is something that needs to be improved.

The Constitution of the Republic of Serbia

The Constitution of the Republic of Serbia as the highest legal act does not contain direct guarantees of the neutrality of the Internet, but rather draws them from the freedom of the media and the right to information.[327] As in everyday life, proclaimed human rights apply in the Internet environment, so the Constitution guarantees freedom of thought, conscience, belief and religion, the right to persuasion, freedom of thought and expression, as well as the freedom to speak, write, paint or in any other way seek, receive and spread information and ideas.[328] It also guarantees that there is no censorship in the Republic of Serbia and that everyone is free to establish newspapers and other means of public information in accordance with the law without approval, in the manner prescribed by law.[329] Perhaps the most concrete contact of the Serbian constitutive act with the neutrality of the Internet can be related to the right of everyone to be truthfully, completely and timely informed about issues of public importance and the media are obliged to respect that right, as well as the right to access data in possession of state bodies and organizations entrusted with public authority, in accordance with the law.[330]

Law on Electronic Communications

The basic, general, regulation that indirectly prescribes the neutrality of the Internet in the Republic of Serbia is the Law on Electronic Communications[331] which also defines the Internet by determining it as a global electronic communication system composed of a large number of mutually connected computer networks and devices which exchange data using a common set of communication protocols.[332] The law does not precisely define the concept of neutrality of the Internet, but its concept and protection of access primarily through the principles and objectives of this act which are based, inter alia, on numerous principles, among which the issue of neutrality of the Internet can be recognized, particularly in the principle defined as follows - providing opportunities for end users to freely access and distribute information when using public communications networks and services, as well as to use applications and services of their choice.[333] Furthermore, the Law defines the service of providing Internet service as an electronic communication service which is generally provided for a fee, and consists entirely or mainly of signal transmission in electronic communication networks, including telecommunications services and

[326] Ibid.

[327] Ustava Republike Srbije ("Sl. glasnik RS", br. 98/2006), Articles 50 and 51 of the Constitution of the Republic of Serbia ("Official Gazette of the Republic of Serbia", no. 98/2006).

[328] Article 46 of the Constitution

[329] Article 50, par. 3, of the Constitution.

[330] Article 51, par. 1, of the Constitution.

[331] Zakon o elektronskim komunikacijama RS (The Law on Electronic Communications ("Official Gazette of the Republic of Serbia ", no. 44/2010, 60/2013 – Constitutional Court decision, 62/2014 and 95/2018 – state law).

[332] Article 4 par. 1 subpar.15, of the Law on Electronic Communications of the Republic of Serbia.

[333] Article 3 of the Law on Electronic Communications of the Republic of Serbia.

services of distribution and broadcasting of media content, but does not include services providing media content or performing editorial control over media content transmitted via electronic communications networks and services, nor does it include information society services that do not consist entirely or predominantly of the transmission of signals via electronic communications networks.[334]

By analyzing the aforementioned norms, we can conclude that service providers do not have the authority to engage in the evaluation of content that is transmitted or to affect the speed of its flow. The same act regulates, among other things, the position and work of the Regulatory Agency for Electronic Communications and Postal Services (RATEL)[335] which exercises public authority in order to effectively implement the established policy in the field of electronic communications, and among other things, decides on the rights and obligations of operators, i.e. postal operators and users, cooperates with bodies and organizations responsible for broadcasting, competition protection, consumer protection, protection of personal data and other bodies and organizations concerning issues important for the field of electronic communications and postal services.[336] The Ministry of Trade, Tourism and Telecommunications is in charge of supervising RATEL[337], as an executive authority which controls the application of the provisions guaranteeing the freedom of the Internet, and apart from it there is no other body in charge of supervising or regulating Internet content in Serbia. Therefore, although the Republic of Serbia, at the legal level, has a certain level of guarantees for the neutrality of the Internet, the problem is that the regulatory bodies that enforce the law cannot be said to be independent.

Another problem is the non-transparency of the work of both operators and operators, again due to the impossibility of separation from the executive branch. Nevertheless, according to the data stated in the report of the Share Foundation for 2020[338] there have been no serious cases of digital content restrictions in the country, nor do laws define restrictions on electronic or online communication, given that Serbia does not have a special law regulating online content issues, and general media laws, such as the Law on Public Information and Media and the Law on Electronic media, are not used to prohibit or restrict online speech.[339] Internet content is widely available and political, cultural or social content is not blocked.[340]

---

[334] Article 4, par. 1, subpar 10, of the Law on Electronic Communications of the Republic of Serbia

[335] RATEL has two main bodies - the Board of Directors and the Director of the Agency. The members of the Board of Directors are elected by the National Assembly, on the basis of a public competition conducted by the competent Ministry. RATEL is financially independent of the executive authority, as it is financed by various fees (for example, those for charging for the use of frequencies) paid by service providers. However, all surplus funds must be transferred to the state budget. See more about RATEL on the official website: (https://www.ratel.rs/cyr/ 02.01.2022).

[336] Article 2 of the Law on Electronic Communications of the Republic of Serbia.

[337] Articles 119 - 123 of the Law on Electronic Communications of the Republic of Serbia.

[338] Articles 126 and 127 of the Law on Electronic Communications of the Republic of Serbia.

[339] Freedom of the Net, Serbia – 2021, Freedom House Report, Available online at: https://freedomhouse.org/sr/country/serbia/freedom-net/2021 01.2.2022.).

[340] A report by the Share Foundation said the government had blocked a number of betting sites. In October 2020, users of certain Internet providers could not access a number of betting websites that had previously operated freely. As a justification for these bans, Internet service providers stated that they adhered to the new Law on Games of Chance passed in April 2020, which prohibits "participation in games of chance, which are organized abroad, for which bets are placed and paid on the territory Republic of Serbia". Ibid. Freedom House Report 2021. Ibid. Freedom House Report 2021.

Right to freedom of expression

Thanks to the ever-improving technical-technological infrastructure and the availability of the Internet, today, as users and consumers, we are constantly online, on the network. The Internet and digital space are increasingly becoming a place where freedom of expression and media freedom face many challenges and problems - algorithmic and platform-oriented informing leads to spread of information in closed circles of like-minded people or acquaintances (so-called news bubbles), the quantity and the speed of spreading misinformation is so great that it sometimes seriously endangers social dialogue, and the possibilities for spreading intolerant, discriminatory and hateful content have become even greater.[341]

There are numerous legal guarantees of freedom of expression and freedom of the media in Serbia. They are therefore primarily guaranteed by the Constitution[342], and then specified by the Law on Public Information [343] and the Law on Electronic Media.[344] Serbia has also ratified the most important international acts guaranteeing freedom of thought and freedom of the press, the International Covenant on Civil and Political Rights and the European Convention for the Protection of Human Rights and Fundamental Freedoms. The Serbian Constitution requires that the provisions on rights related to freedom of expression and media information be interpreted in favor of promoting the values of a democratic society, in accordance with applicable international standards and the practice of international institutions, including the European Court of Human Rights. All these guarantees apply to all citizens equally.

However, when it comes to freedom of expression on the Internet, the legal consequences of freedom of expression can be treated differently depending on whether the content published on the Internet is published on a registered or unregistered medium. The Law on Public Information and Media of the Republic of Serbia defines what is considered media and under what conditions. For the purposes of this law, the media are Internet portals of traditional media (press, agencies, radio and TV stations) and independent publications, i.e. editorially designed websites or Internet portals, and entry in the Media Register is specified as a mandatory condition for them.[345] The Law explicitly excludes internet forums, social networks and similar platforms from its definition, while other forms of production and distribution of informative content on the Internet (blogs, web presentations, aggregators) are not considered media if they are not registered in the Media Register. The legislator thus left the choice to the civil and online media to register as media, if they wish, and thus gain the appropriate status with all rights and obligations, while unregistered citizen and online media remain outside the scope of the law.

So, as the Law states, the media, in the sense of this law, are not: platforms, such as internet forums, social networks and other platforms that allow free exchange of information, ideas and opinions of its members, or any other independent electronic publication, such as blogs, web presentations and similar electronic

[341] Maksić, T., Mediji i nove politike upravljanja internetom - Sloboda izražavanja i medijske slobode u digitalnom okruženju, Fondacija za otvoreno društvo, BIRN, Beograd, 2020, p.5.

[342] Article 46 of the Constitution.

[343] Zakon o javnom informisanju i medijima RS (The Law on Public Information and Media, "Official Gazette of the Republic of Serbia ", no. 83/2014, 58/2015 and 12/2016 – authentic interpretation.

[344] Zakon o elektronski medijima RS (The Law on Electronic Media, " Official Gazette of the Republic of Serbia ", no. 83/2014 and 6/2016 – state law.

[345] Article 29-31. of the Law on Public Information of the Republic of Serbia

presentations, unless registered in the Media Register, in accordance with this law.[346] Whether someone decides to publish the content through the media or in some other form, directly affects the scope of rights and obligations, i.e. their liability for the published content. Thus, for example, when it comes to compensation for material and non-material damage, the general regime of liability in accordance with the Law on Contracts and Torts of the Republic of Serbia is applied for unregistered media and the Law on Public Information and Media for registered media. Or, for example, unregistered media, including unregistered online media, cannot count on financial support from the public budget. Such a principle is expressed through the Law on Public Information and Media. Or, when it comes to criminal-legal relations, criminal law still provides the broadest protection of the individual through the crime of insult, where priority is given to freedom of speech and opinion by establishing a special degree of guilt, necessary to establish liability.[347] An aggravating circumstance and more severe punishment is prescribed if the insult is pronounced "through the press, radio, television or similar means", which gives the possibility of creative interpretation regarding liability for the content of online media, because the court may opt for an interpretation according to which online media platforms are means “similar” to traditional media.

When it comes to hate speech, which is unfortunately very present on the Internet and social networks, a wide range of legal protection mechanisms is available, which we will present in the following section.

Hate speech on the Internet

Although there is no legally regulated control of content on social networks, in the Republic of Serbia there are various mechanisms that can be used in case of hate speech or violation of personal rights on social networks. According to the Constitution of the Republic of Serbia, at the national level, hate speech can be sanctioned through criminal, media or anti-discrimination laws. Laws dealing with the regulation of public communication and the media system are also relevant. Individuals, organizations, editors, but also internet service providers, hosting providers and content providers can be held responsible for hate speech.[348]

The Criminal Code of the Republic of Serbia[349] prohibits inciting or provoking national, racial or religious hatred, or intolerance among peoples or ethnic communities living in Serbia, and this crime is punishable by 6 months to five years in prison. The penalty of three months to three years is envisaged for anyone who disseminates or otherwise makes public texts, images or any other representation of ideas or theories that advocate or incite hatred, discrimination or violence against any person or group based on race, skin color, religion, nationality, ethnic origin or any other personal characteristic.[350]

According to the Law on Organization and Competences of State Bodies for Combating High-Tech Crime, hate speech and internet threats are placed under the jurisdiction of the Special Department of the Higher

---

[346] Krivokapić, N., Colić, O., Maksimović, M., Pravni položaj onlajn medija u Srbiji: vodič namenjen onlajn i građanskim medijima kao korisnicima, SHARE Fondacija, Novi Sad, 2015, p. 13. Abailable here: (https://resursi.sharefoundation.info/wp-content/uploads/2018/10/vodic-pravnipolozaj_onlajn_medija_u_srbiji_-_preview_.pdf 12/01/2022).

[347] According to the Criminal Code of the Republic of Serbia ("Official Gazette of the RS", no. 85/2005, 88/2005 - amended, 107/2005 - amended, 72/2009, 111/2009, 121/2012, 104/2013, 108 / 2014, 94/2016 and 35/2019), Article 170 of the Law, insult is a criminal offense. It is important to note that the prosecution for insult is undertaken upon a private lawsuit of a person who believes that their right has been violated.

[348] Predrag M. Nikolić, Govor mržnje u internet komunikaciji u Srbiji, doktorska disertacija, Beograd, 2018., p.107. (https://www.fpn.bg.ac.rs/wp-content/uploads/2018/07/Nikolic_Predrag_Disertacija_FPN.pdf 01/02/2022)

[349] Krivičnog zakonika RS, Article 317 of the Criminal Code of the Republic of Serbia

[350] Article 38. par. 4 of the Criminal Code of the Republic of Serbia.

Prosecutor's Office in Belgrade, the Special Department of the Higher Court in Belgrade and the Ministry of the Interior Department for High-Tech Crime.[351] As stated in this law, high-tech crime are criminal offenses in which computers, computer systems, computer networks, computer data, as well as their products in material or electronic form appear as an object or means of committing criminal offenses. The aim of this law and these institutions is to detect, prosecute and bring before court, inter alia, crimes against freedoms and rights of man and citizen, sexual freedom, public order and peace and constitutional order and security of the Republic of Serbia, which due to the manner of execution or used means can be considered high-tech criminal offenses.[352]

The Law on Prohibition of Discrimination is another law that bans hate speech by prohibiting the expression of ideas, information and opinions that incite discrimination, hatred or violence against a person or group of persons because of their personal characteristics, in public media and other publications, at gatherings and places available to the public, by printing and displaying messages or symbols, and otherwise.[353] This law treats hate speech as a severe form of discrimination and in this regard it should not be relevant at all whether that hate speech occurs in the online world or in the offline world. The application of the Law on Prohibition of Discrimination enables the conduct of court proceedings, either directly by persons affected by discrimination, or indirectly, through the Commissioner for Equality of Citizens who receives and considers complaints concerning violations of this Law, and gives opinions and recommendations in specific cases, as well as warnings.[354] The proceeding can be ended with a single determining, declaratory verdict whether certain conduct constitutes an act of discrimination or not.[355]

In the Law on Prohibition of Manifestation of Neo-Nazi and Fascist Organizations and Prohibition of the Use of Neo-Nazi and Fascist Symbols and Marks, Article 6 in paragraph 1 directly links "incitement, provoking and spread of hatred and intolerance" prohibited by Article 3 of the Law with “making symbols, marks or propaganda material containing neo-Nazi or fascist marks available to the public through computer systems”.[356]

An array of media laws also regulates hate speech, but the most important one for the Internet is the Law on Public Information and Media, which defines online publications of newspapers, radio and TV programs as well as news agency services as media, i.e. public information agents.[357] The law stipulates that upon the proposal of the competent public prosecutor, the competent court may prohibit the distribution of

[351] Zakon o organizaciji i nadležnosti državnih organa za borbu protiv visokotehnološkog kriminala, Article 4 and 5 of the Law on Organization and Competences of State Bodies for Combating High-Tech Crime ("Official Gazette of the Republic of Serbia ", no. 61/2005 and 104/2009).

[352] Article 3, par. 3 of the Law on Organization and Competences of State Bodies for Combating High-Tech Crime.

[353] Article 11 of the Law on Prohibition of Discrimination (“Official Gazette of the Republic of Serbia”, no.22/2009 and 52/2021)

[354] In case it notices “frequent, typical and severe cases of discrimination” this institution can issue a warning to the public. Due to the increasing hate speech in the media content, as well as in the comments on media portals, the Commissioner for the Protection of Equality, in 2018, sent a recommendation to internet portals in Serbia in order to prevent the publication of content and comments that may incite hatred or violence against persons or groups of persons. See the text of Recommendation No. 021-01-00327 / 2018-02 of 27 September 2018: (http://ravnopravnost.gov.rs/rs/preporuka-mera-za-ostvarivanje-ravnopravnosti-za-internet-portale/ 12/11/2021).

[355] For more information about the proceedings, see the Commissioner’s website: (http://ravnopravnost.gov.rs/ko-smo-i-sta-radimo/ 12/11/2021).

[356] Zakon o zabrani manifestacija neonacističkih ili fašističkih organizacija i udruženja i zabrani upotrebe neonacističkih ili fašističkih simbola i obeležja RS, The Law on Prohibition of Manifestation of Neo-Nazi and Fascist Organizations and Prohibition of the Use of Neo-Nazi and Fascist Symbols and Marks ("Official Gazette of the Republic of Serbia", no. 41/2009).

[357] The Law on Public Information and Media ("Official Gazette of the Republic of Serbia", no. 83/2014, 58/2015 and 12/2016 – authentic interpretation).

information or other media content if the information refers to an act of direct violent destruction of the constitutional order and an act of direct violence against a person or group based on race, nationality, political affiliation, religion , sexual orientation, disability or other personal characteristics, and the publication of information is directly threatened by a serious and irreparable consequence whose occurrence cannot be prevented in any other way.[358] In addition, this law explicitly prohibits hate speech by stipulating that ideas, opinions or information published in the media must not incite discrimination, hatred or violence against a person or group of persons because of their belonging or non-belonging to a race, religion, nation, sex, due to their sexual orientation or other personal characteristics, regardless of whether the crime was committed by publishing.[359] The law further stipulates that there is no violation of the prohibition of hate speech in cases where it is part of the text without the intention to discriminate and if in fact it is intended to critically point out discrimination, hatred or violence.[360] Also, caricature and satirical portrayal of a person is not considered a violation of the dignity of the person, i.e. of the right to authenticity. The content of sites registered with the Serbian Business Register Agency as online media falls under the jurisdiction of the Ministry of Culture and Information, which oversees the implementation of the Law on Public Information and Media.[361]

The law on electronic media is another law that explicitly prohibits hate speech. The regulatory body for electronic media ensures that the program content of the media service provider does not contain information that incites, openly or covertly, discrimination, hatred or violence due to race, skin color, ancestry, citizenship, nationality, language, religious or political beliefs, gender, gender identity, sexual orientation, property status, birth, genetic characteristics, health status, disability, marital and family status, conviction, age, appearance, membership in political, trade union and other organizations and other actual or presumed personal characteristics.[362] Violations of these prohibitions are subject to fines, and the Regulatory Body for Electronic Media is responsible for regulating the content.[363]

In addition to this series of legal solutions, there are also provisions through which the media self-regulate or co-regulate. The Code of Journalists of Serbia also deals with hate speech, instructing journalists to do everything in their power to avoid "discrimination based on race, gender, age, sexual orientation, language, religion, political and other opinion, national or social origin."[364] In case of violation of the provisions of the Code of Journalists, the Press Council, as a body that monitors ethical standards in the print media and responds to citizens' reports, makes a public decision, which informs the reported media to delete problematic content and publish apologies if necessary.[365]

---

[358] Article 59 of the Law on Public Information and Media of the Republic of Serbia

[359] Article 75 of the Law on Public Information and Media of the Republic of Serbia.

[360] Articles 76 and 77 of the Law on Public Information and Media of the Republic of Serbia

[361] Article 132 of the Law on Public Information and Media of the Republic of Serbia. In addition, numerous other institutions that deal with regulation of information and communication content must abide by the laws that regulate this field and issue warnings or sanction inappropriate content (such as Serbian Internet Domain Registry).

[362] Article 51, of the Law on Electronic Media ("Official Gazette of the Republic of Serbia," no. 83/2014, 6/2016 – state law and 129/2021).

[363] Article 5 of the Law on Electronic Media.

[364] Part V, par. 4 of the Code of Journalists of Serbia; the entire document is available here: (https://savetzastampu.rs/dokumenta/kodeks-novinara-srbije/ 12/11/2021).

[365] For more information on the competence of the Press Council and the procedures for appeals before the Press Council, see: (https://savetzastampu.rs/o-nama/sta-mozemo-da-uradimo-za-vas/ 11/12/2021).

When it comes to hate speech on social networks, since they are not registered as a "medium" in the Republic of Serbia, they are guided by their own rules, which differ from company to company, so for example, Twitter has stricter rules concerning hate speech than Facebook.[366]

Conclusions

The Internet has changed our attitude towards life because it has connected us on a global level and turned the former audience into new media.[367] The Internet has enabled us to follow the lives of millions in real time and convinced us to know everything about places we have never visited. That is why and because of that, it is necessary to constantly work on improving the mechanisms of legal regulation of the Internet.

As far as the legal regulation of social networks is concerned, the Republic of Serbia is very far from more concrete legal solutions on this topic, due to the low level of technological literacy of our population, as well as the economic underdevelopment related to it, but also the lack of interest of major platforms to deal with the Serbian market. The other side that could lead this debate is the state of Serbia itself, more precisely the legislator, but even on that side, it seems that regulating social networks is not a priority. Therefore, one of the recommendations would certainly be for the competent state bodies to encourage a broad public debate with several stakeholders on regulations concerning primarily respect for human rights on the Internet, as well as a debate on the legal regulation of content on social networks. In that sense, it would be good to adopt a new law on the media, which would recognize and thus regulate platforms for social networks, although they are not media in the traditional sense of the word.

It is also necessary to ensure that media and information literacy programs are implemented at the national level and that additional efforts are made to protect Internet neutrality, in line with EU Internet neutrality rules. The digital media sphere should provide free access to information and knowledge, equal opportunities for everyone to contribute to public debate and decentralization of power in the field of information and education. At the same time, public policies in the media sphere must make sure to respect the rules and protect basic human rights, such as privacy and security.

When it comes to hate speech on the Internet, although there is no doubt that there is room for improvement of the legal framework, it can be stated that the existing constitutional and provisions of certain laws we analyzed provide sufficient guarantees for protection against hate speech but unfortunately, insufficient application of existing legal frameworks and the lack of regulation for social networks still allow the digital space to be full of aggressive communication, threats and insults.[368] Therefore, it is important to encourage self-regulation of Internet portals that would make clear internal rules regarding the prohibition of hate speech in user-generated content and, more importantly, systematically improve preventive measures against hate speech, primarily in terms of educating citizens about the harmfulness of hate speech and its consequences.

---

[366] For Twitter's rules, see: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy 12/11/2021); For Facebook's rules, see: https://transparency.fb.com/policies/community-standards/hate-speech/ 01/02/2022)

[367] Martinoli, A., (ed.), (2019.), p.20.

[368] Krstić, I., Izveštaj o upotrebi govora mržnje u medijima u Srbiji, Savet Evrope i Poverenik za zaštitu ravnopravnposti, Beograd, 2020., pp.44-45; See also: Petrovski, A., Krivokapić, D., (ed.), GREŠKA 404: Digitalna prava u Srbiji 2014-2019, SHARE Foundation, Pres doo, Novi Sad, 2019., pp.56-57; 73-74.; Stojković, M., Pokuševski, D., Anonimna mržnja - Mehanizmi zaštite od govora mržnje na internetu, Beogradski centar za ljudska prava, Beograd, 2018, pp.38-41.

References

Literature

Gregg, B. (2012), Politics disembodied and deterritorialized: The Internet as human rights resource. In: Dahms, H. F. & Hazelrigg, L., (ed.), Theorizing Modern Society as a Dynamic Process (Current Perspectives in Social Theory, Vol. 30) (pp. 209–233). Bingley: Emerald Group Publishing Limited;

Hick, S., Halpin, E., & Hoskins, E. (2016), Human rights and the Internet. London: Palgrave Macmillan;

Krstić, I., (2020), Izveštaj o upotrebi govora mržnje u medijima u Srbiji, Savet Evrope i Poverenik za zaštitu ravnopravnposti, Beograd;

Khor, L. (2011), Human rights and network power. Human Rights Quarterly, 33(1), 105–127.

Krivokapić, N., Colić, O., Maksimović, M., (2015), Pravni položaj onlajn medija u Srbiji: vodič namenjen onlajn i građanskim medijima kao korisnicima, SHARE Fondacija, Novi Sad;

Martinoli, A., (ur.), (2019), Priručnik: Regulatorni okvir i poslovni modeli onlajn medija, Fondacija za otvoreno društvo, Beograd;

Maksić, T.,(2020), Mediji i nove politike upravljanja internetom - Sloboda izražavanja i medijske slobode u digitalnom okruženju, Fondacija za otvoreno društvo, BIRN, Beograd;

Mitrović, M., (2020), Sloboda izražavanja i zaštita podataka o ličnosti na internetu: Perspektiva internet korisnika u Srbiji, CM Komunikacija i mediji 15 (47), 5-34, Beograd;

Petrovski, A., Krivokapić, D., (ur.), (2019), GREŠKA 404: Digitalna prava u Srbiji 2014-2019, SHARE Fondacija, Pres doo, Novi Sad;

Stojković, M., Pokuševski, D., (2018), Anonimna mržnja - Mehanizmi zaštite od govora mržnje na internetu, Beogradski centar za ljudska prava, Beograd;

Internet Sources

The Commissioner for the Protection of Equality, Recommendation no: 021-01-00327/2018-02, 27.9.2018.: (http://ravnopravnost.gov.rs/rs/preporuka-mera-za-ostvarivanje-ravnopravnosti-za-internet-portale/ 12.11.2021.);

Chambers, C. (2014). Facebook fiasco: Was Cornell's study of 'emotional contagion' an ethics breach? The Guardian, (https://www.theguardian.com/technology/2014/jun/30/facebook-emotion-study-breached-ethical-guidelines-researchers-say 12.12.2021.);

Digital Service Act - The Digital Services Act, European Parliament: (https://www.europarl.europa.eu/news/en/press-room/20220114IPR21017/digital-services-act-regulating-platforms-for-a-safer-online-space-for-users);

Internet freedoms report 2020, Mirkovic, N., Merrell, F., (ed.), Civil Rights Defenders and Share Foundation, Available at: (https://crd.org/wp-content/uploads/2020/04/200402_GRA_InternetFreedoms_Narativa_A4_Spreads.pdf 10.12.2021.);

Freedom of the Net, Serbia – 2021, Freedom House Report, Available online at: (https://freedomhouse.org/sr/country/serbia/freedom-net/2021 12.12.2021.);

Freedom House Report: Russia, 2018, Available at: https://freedomhouse.org/country/russia/freedom-net/2018 12.12.2021.);

France online hate speech law to force social media sites to act quickly, The Gardian, *Agence France-Presse in Paris,* 9.07.2019. (https://www.theguardian.com/world/2019/jul/09/france-online-hate-speech-law-social-media 10.02.2022.);

Komčarević, D., Pet "Šta" o Twiteru u Srbiji, Radio Slobodna Evropa *online*, 18.08.2021. (https://www.slobodnaevropa.org/a/srbija-twitter-laz-vesti-genocid/31416738.html 12.12.2021.);

Komarčević D., Posebne izborne mere 'Fejsbuka', na listi i Srbija, Crna Gora i Severna Makedonija, Radio Slobodna Evropa *online*, 17.03.2020. (https://www.slobodnaevropa.org/a/fejsbuk-izbori-mere-srbija-crna-gora-severna-makedonija/30492720.html 12.12.2021.);

Milić, D., Neutralnost interneta u pravu Republike Srbije, MillicLawOffice, 8.05.2019., (https://www.milic.rs/blog/internet-pravo/neutralnost-interneta-u-pravu-republike-srbije/ 10.12.2021.);

Nikolić, P. M., (2018), Govor mržnje u internet komunikaciji u Srbiji, doktorska disertacija, Beograd, (https://www.fpn.bg.ac.rs/wp-content/uploads/2018/07/Nikolic_Predrag_Disertacija_FPN.pdf 10.12.2021.);

Popović Aleksandra, Ko uređuje društvene mreže – zakonodavstvo SAD, EU i Srbije, Talas.rs, 22.01.2021.,(https://talas.rs/2021/01/22/ko-ureduje-drustvene-mreze-zakonodavstvo-sad-eu-i-srbije/ 12.12.2021.);

Radojević, V., 2020, Kako je radila srpska „bot"armija: 43 miliona tvitova podrške Vučiću, 4.04.2020., Raskrinkavanje, (https://www.raskrikavanje.rs/page.php?id=Kako-je-radila-srpska-bot-armija-43-miliona-tvitova-podrske-Vucicu-642 12.12.2021.);

RATEL (https://www.ratel.rs/cyr);

Section 230: An Overview, Congressional Research Service, 7.04.2021., (https://crsreports.congress.gov/product/pdf/R/R46751 12.12.2021 12.12.2021);

Slobode na internetu u zemljama Zapadnog Balkana, Share fondacija, Civil Rights Defenders, (https://crd.org/wp-content/uploads/2020/04/SRB_Saz%CC%8Cetak_Slobode-na-internetu.pdf 01.02.2022.).

Legal Acts

Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz - NetzDG) /Act to Improve Enforcement of the Law in Social Networks. Available at: (https://perma.cc/RW47-95SR 10.12.2021.);

Zakon o elektronskim komunikacijama ("Sl. glasnik RS", br. 44/2010, 60/2013 - odluka US, 62/2014 i 95/2018 - dr. zakon;

Zakon o javnom informisanju i medijima *("Sl. glasnik RS", br. 83/2014, 58/2015 i 12/2016 - autentično tumačenje);*

Zakon o zabrani manifestacija neonacističkih ili fašističkih organizacija i udruženja i zabrani upotrebe neonacističkih ili fašističkih simbola i obeležja *("Sl. glasnik RS", br. 41/2009);*

Zakona o zabrani diskriminacije, ("Sl.glasnik RS", br.22/2009 i 52/2021.);

Krivični zakoniku Republike Srbije *("Sl. glasnik RS", br. 85/2005, 88/2005 - ispr., 107/2005 - ispr., 72/2009, 111/2009, 121/2012, 104/2013, 108/2014, 94/2016 i 35/2019)*;

Zakona o elektronskim medijima *("Sl. glasnik RS", br. 83/2014, 6/2016 - dr. zakon i 129/2021);*

Zakon o organizaciji i nadležnosti državnih organa za borbu protiv visokotehnološkog kriminala *("Sl. glasnik RS", br. 61/2005 i 104/2009);*

Ustav Republike Srbije ("Sl. glasnik RS", br. 98/2006);

Recommendation CM/Rec(2016)5[1] of the Committee of Ministers to member States on Internet freedom, Available at: (https://mediainitiatives.am/wp-content/uploads/2017/03/Recommendation-of-the-Committee-of-Ministers-on-Internet-Freedom-in-English.pdf 1.02.2022.).

Case Law

Supreme Court verdict in Reno v. ACLU, 521 U.S. 844 (1997), (https://www.aclu.org/legal-document/reno-v-aclu-supreme-court-decision 1.12.2021.)

Digital Rights of Platform Workers in Italian Jurisprudence

Federico Costantini and Alan Onesti
DEPARTMENT OF LAW, UNIVERSITY OF UDINE

Summary

This contribution provides an analysis of the decisions provided by Italian judicial courts and Data Protection Authority concerning the so called "platform workers". The relevance of this topic for the activities promoted by the GDHRNet Cost Action is due not only for the newness of the technologies deployed in such kind of platforms, but foremost by the fact that some of them involve, directly or indirectly, the matter of "Digital rights". Therefore, these decisions can be considered as written testimony of the first approach adopted by Italian jurisprudence to the challenges raised in terms of data protection and algorithmic discrimination, for example, in this field.

The report is composed of the following parts: Section I provides a general overview on the phenomenon of platform workers; Section II explains the methodology adopted in the research; Section III analyses each single case decided in Italy by judicial courts and by the Data Protection Supervisor; Section IV provides a synthesis of the digital rights emerged from the discussion; Section V offers some recommendations and a few final evaluations. At the end, references are provided.

"platform economy" and "platform workers", a general overview

The phenomenon of the 'collaborative economy', 'sharing economy' or 'gig economy' entered the EU a decade ago. Such kind of economic ecosystems, since then, have been defined in different ways: "*digital networks that coordinate labour service transactions in an algorithmic way*" (Pesole et al. 2018) and, more recently: "*organisations (that are most often, but not always, firms) that offer digital services that facilitate interactions via the Internet between two or more distinct but interdependent sets of users (whether organisations or individuals) and that generate and take advantage of network effects*" (Gawer and Srnicek 2021).

From a functional point of view, four types of digital platforms can be distinguished, depending on whether they concern (1) e-commerce, in a general sense, (2) the sharing of resources ("asset-based", as in the case of AirBnb), (3) the organisation of workforce ("digital labour platforms"), or (4) a way of effectively sharing of goods or services ("collaborative platforms" or "sharing platforms", e.g. BlaBlaCar, Kickstarter). Within the so called "digital labour platforms", a further distinction can be drawn depending on the fact that performance intermediated entail (1) manual or physical activities (e.g. Uber or TaskRabbit), (2) repetitive online tasks (e.g. Amazon Mechanical Turk) or (3) services involving high or specific skills (e.g. PeoplePerHour, Freelancer) (Pesole et al. 2019).

Due to the surge of "digital labour platforms", a new kind of workforce emerged, called "digital platform workers", which has spread worldwide over the last ten years. As we know, this is a category of workers whose activities are determined and controlled by means of continuous and pervasive interaction with sophisticated algorithms.

The novelty of the structural relations generated by digital platforms of this kind has brought to the attention of legislators the need of a stronger protection of workers, especially due to the vulnerability that afflicts it (Codagnone, Biagi, and Abadie 2016). Indeed, the working conditions in this field are usually severe, given the many risks undertaken (e.g. road incidents by riders), the low level of remuneration, and the length and distribution of working-shifts. In recent years we have witnessed the intervention of the EU legislation on the transparency of working conditions, with the recent Directive (EU) 1152/2019[369] and recent concerns expressed by the European Parliament[370]. On the other hand, as regards the Italian legal framework, it is also worthwhile remembering the extension of the discipline of the employment relationship provided for by Legislative Decree 81/2015[371] - and in particular the extension of the guarantees - which took place with Law Decree 101/2019[372] (Filì and Costantini 2019)[373].

From this perspective, in Italy it is noteworthy a recent wave of judicial decisions which strive in framing the pervasive and constant interaction between the individual worker and the online platform in the light of traditional labour law categories. Those decisions take into consideration also issues which recently have been included in the concept of "algorithmic discrimination"(Wachter, Mittelstadt, and Russell 2021; Gerards and Xenidis 2021; Mittelstadt et al. 2016; Floridi and Illari 2014), which occurs when a system automatically creates an unjustified imbalance between different categories of persons, thus affecting their rights under the legal system. Such issue is particularly sensitive, since in automatic reasoning it is physiological to find distortions - 'biases' - that depend on many factors - technological, social, human - and that are difficult to represent in a coherent way from a logical-informatic point of view, and therefore to eliminate or correct beforehand.

However, other elements depending on the context in which these technologies operate - the "complacency" that generally accompanies the outcome of digital automatisms, the lack of awareness of the intrinsic risks for operators, the extent and intensity of the potential harmful effects and their rate of propagation - can make the phenomenon particularly insidious, especially since normally - in almost all cases - the discrimination generated is indirect, since it can occur through the combination of factors that would be harmless if considered independently.

As regards these latter issues, the EU legal provisions sometimes can result inconsistent or inapt for many reasons: (1) the lack of coordination among the provisions included in the Fundamental Treaties (art. 2, art. 3§3 TEU, art. 19 TFEU, art. 21 of the Charter of Fundamental Rights), the secondary level of dispositions - mainly the notions of "direct discrimination" and "indirect discrimination" as defined by art. 2 § 2 of Directive 2000/43/CE[374] - and the general prohibition to process particular kind of data expressed by art. 9 §.1 GDPR; (2) the wide spectrum of interpretation of the "right of explanation", which still raises many

[369] https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:32019L1152

[370] See European Parliament Resolution adopted on 16 September 2021, Fair working conditions, rights and social protection for platform workers - New forms of employment related to digital development, P9_TA(2021)0385, https://www.europarl.europa.eu/doceo/document/TA-9-2021-0385_IT.html

[371] http://www.normattiva.it/eli/id/2015/06/24/15G00095/CONSOLIDATED.

[372] http://www.normattiva.it/eli/id/2019/09/04/19G00109/CONSOLIDATED/20220321

[373] Only recently in Italy platform workers have achieved legal protection regardless the status of self employed or employee. Indeed, Law Decree 101/2019 extended labour safety minimum obligations - especially in terms of health and physical integrity (Article 47 bis).

[374] Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin, OJ L 180, 19.7.2000, p. 22–26 , https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32000L0043.

discussions (Wachter, Mittelstadt, and Floridi 2017; Malgieri and Comandé 2017); (3) the unavoidable fact that interpretation of EU regulation may be quite different among member States.

Paradoxically, behind the apparent reassuringly sightseeing of thousands of cyclists, which swarm every day throughout our cities wearing cheerfully coloured thermal backpacks, lies a massive integration among different information technologies, which not only makes it difficult for the courts to identify and prosecute 'algorithmic discrimination' and violation of fundamental rights, but also to detect and prosecute possible abuses and identify suitable criteria for quantifying damage compensation.

Therefore, we can claim that "platform workers" are exposed to a twofold threat: the first relates to new forms of violation of labour-law-based rights; the second regards the emergence of specific threats involving "Digital rights", for which there aren't still effective remedies.

This document describes how Italian jurisprudence has addressed these issues, focusing on those specifically related to the latter topic.

Methodology

This report collects all the judicial decisions involving platform workers issued in the Italian legal system. Also included are the provisions of the Italian Data Supervisor Authority "Garante per la Protezione dei Dati Personali" as it is an independent administrative body that, according to Italian literature, embodies also jurisdictional powers.

A ID number has been assigned to each item. Each decision has been analysed according to a common pattern. The results are represented in the following tables, which include identifying data (date, authority, claimant, grade), a short description of the facts, the legal issues discussed, the party awarded by the decision, and also highlights the issue concerning "Digital Rights", if any. Whenever a case is related to others (e.g. in provisional proceedings), it is specified to whom. If a case is similar to a previous one, or it stems from the same proceeding, the connection is referenced. For clarity, the Italian names of the courts, proceedings and decisions are included between round brackets.

Analysis of Italian case-law (ALAN)

Case n.	1
Court / Authority	Court of Florence, labour law section (Tribunale di Firenze, sezione lavoro)
Date of the decision	01.04.2020
Type of proceeding	Provisional / Protective Proceeding (decreto ex art. 700 c.p.c.)
Claimant	Employee/Platform worker
Grade	First
Short description of facts	The claimant, a registered rider for Just Eat Italy s.r.l., requested a set of personal protective equipment against the risk of COVID-19 (gloves, sanitizing gels and cleaning products for his rucksack), the use of which (as regards the gloves and the mask) was recommended

	by the defendant itself for the purposes of carrying out his work in this period of epidemiological emergency. The defendant, despite the employee's requests to that effect, refused to provide such protective devices.
Legal issues raised	The worker took legal action to obtain an urgent measure. The claim is based on the Italian legislation for the protection of workers' health (Article 71, Legislative Decree No. 81/2008) which requires the employer to provide workers with equipment suitable for protecting their health and safety, suitable for the work to be carried out or adapted for such purposes, in compliance with and used in accordance with the legislative provisions transposing EU directives.
Decision	In favour of Employee/Platform worker
Issues concerning digital rights	-

Case n.	2
Court / Authority	Court of Florence, labour law section (Tribunale di Firenze, sezione lavoro)
Date of the decision	05.05.2020
Type of proceeding	Judgment of merit (ordinanza, consequential to the provisional decision dated 01.04.2020 - Cfr. case n. 1)
Claimant	Employee / platform worker
Grade	Second
Short description of facts	Cfr. Case n. 1
Legal issues raised	Cfr. Case n. 1
Decision	In favour of Employee/Platform worker: confirms the decision of the judgment of merit.
Issues concerning digital rights	-

Case n.	3
Court / Authority	Court of Bologna, labour law section (Tribunale di Bologna, sezione lavoro)
Date of the decision	14.04.2020
Type of proceeding	Provisional/Protective Proceeding (decreto ex art. 700 c.p.c.)

Claimant	Employee/Platform worker
Grade	First
Short description of facts	The applicant, a rider for Deliveroo Italy s.r.l., requested personal protective equipment against the risk of COVID-19 (gloves, sanitizing gels and cleaning products for his rucksack) to his employer. The request wasn't satisfied due to organizational problems (high number of requests, difficult of supplying said materials).
Legal issues raised	Cfr. Case n. 1
Decision	In favour of Employee/Platform worker
Issues concerning digital rights	-

Case n.	4
Court / Authority	Court of Bologna, labour law section (Tribunale di Bologna, sezione lavoro)
Date of the decision	01.07.2020
Type of proceeding	Judgment of merit (ordinanza, consequential to the provisional decision dated 14.04.2020 - Cfr. Case n. 3)
Claimant	Employee/Platform worker
Grade	First
Short description of facts	Cfr. Case n. 3
Legal issues raised	Cfr. Case n. 3
Decision	In favour of Employee/Platform worker: confirms the decision of the judgment of merit.
Issues concerning digital rights	-

Case n.	5
Court / Authority	Court of Bologna, labour law section (Tribunale di Bologna, sezione lavoro)
Date of the decision	11.08.2020
Type of proceeding	Judgment of merit (ordinanza ex art. 700 c.p.c.)
Claimant	Employer
Grade	Second (appeal of the decision of the General Court of Bologna dated 01.07.2020 - Cfr. Case n. 4)

Short description of facts	Cfr. Case n. 3
Legal issues raised	Cfr. Case n. 3
Decision	In favour of Employee/Platform worker: confirms the decision of the first grade of judgment.
Issues concerning digital rights	Cfr. Case n. 3

Case n.	6
Court / Authority	Court of Florence, labour law section (Tribunale di Firenze, sezione lavoro)
Date of the decision	09.02.2021
Type of proceeding	Provisional/Protective Proceeding (decreto ex art. 28 Statuto dei Lavoratori)
Claimant	Trade Unions of workers
Grade	First
Short description of facts	Deliveroo Italy s.r.l., with a communication to all its riders (dated 02.10.2020), forced them to accept a collective agreement signed with the trade union Ugl Rider, as a condition to continue their work. In case of refusal, the job contract would have been solved.
Legal issues raised	Others Trade Unions took a legal action, according to art. 28 of the Statute of Worker's rights[375], to ascertain the unfair labour practice by the employer. In particular, they claimed that: The trade union Ugl "Rider" that subscribed the "C.C.N.L. Rider" (collective worker's agreement for the Riders), together with Assodelivery (association of employers), was not qualified to do so, as it had obtained an illegitimate financial support by the employer; The sudden termination of the of the contract of more than 8.000 riders, without involving the trade-unions, could be qualified as a mass layoff in violation of the right of trade unions to be informed and involved in the management of collective measures and of collective dismissal. The aforementioned right is applicable to platform workers, according to art. 2 of the Legislative Decree no. 81/2008.

[375] LEGGE 20 maggio 1970, n. 300, Norme sulla tutela della liberta' e dignita' dei lavoratori, della liberta' sindacale e dell'attivita' sindacale, nei luoghi di lavoro e norme sul collocamento (GU n.131 del 27-05-1970) http://www.normattiva.it/eli/id/1970/05/27/070U0300/CONSOLIDATED

Decision	In favour of the Employer (Deliveroo Italy S.r.l.)
Issues concerning digital rights	-

Case n.	7
Court / Authority	Court of Palermo, labour law section (Tribunale di Palermo, sezione lavoro)
Date of the decision	12.04.2021
Type of proceeding	Judgment of merit (ordinanza)
Claimant	Trade Unions of workers
Grade	First
Short description of facts	The platform's owner withdrawn from the contract with the rider in advance of its natural end, as the worker refused to sign a new contract that would have been compliant with the collective contract signed between the employer's association and a trade union to which the rider didn't belong.
Legal issues raised	The worker took legal action to establish that the employer's withdrawal was illegal, as it discriminated the worker for his membership to a specific Trade Union.
Decision	In favour of Employee.
Issues concerning digital rights	The Court decided that the withdraw was not only illegal, but also null and void (with the result of being obliged to restore the worker contract), as it was against the prohibition of discrimination provided by the Italian Constitution and by the Article 14 of the CEDU[376] The Court has also condemned the employer to compensate the worker for damages for an amount of 5.000 €.

Case n.	8
Court / Authority	Court of Bologna, labour law section (Tribunale di Bologna, sezione lavoro)
Date of the decision	31.12.2020
Type of proceeding	Judgment of merit (ordinanza)

[376] Every worker has the right to partecipate to the Trade Union that is the most rappresentative for him, and cannot be obliged to accept a collective agreement that has been signed by a employee's Trade Union, which is different to the one he belong to. Moreover, he can't suffer retaliations or poor treatments, based on the membership to a Trade Union.

Claimant	Trade Unions of workers
Grade	First
Short description of facts	The workload division system between all the riders registered for Deliveroo Italy S.r.l. was entrusted to be based on an algorithm. Among the instructions, one established that riders who revoked their availability with a forewarning lower than 24 hours, would have been penalized with a lower possibility of booking a specifically time slot, having consequentially a lower possibility of work, thus a minor income.
Legal issues raised	The applicants took legal action to terminate and sanction Deliveroo Italy's behaviour, as it was discriminatory (art. 14 CEDU) and damaging to the freedom of assembly and association (art. 11 CEDU) since it did not took properly in consideration the cause of the revocation, especially the exercise of union rights (e.g. to assembly with others).
Decision	In favour of Trade Unions workers.
Issues concerning digital rights	The Court granted the application because the algorithm didn't distinguish between the reasons of the revocation, penalizing independently both the defaulting workers and those who acted in force of a recognized right, e.g. the freedom of assembly and association recognized by the article 11 of CEDU. It also caused a discrimination (prohibited by article 14 of CEDU) towards all workers that were part of a Trade Union instead of another. The Court also condemned the employer to compensate the worker for damages for an amount of € 50.000.

Case n.	9
Court / Authority	Court of Milan, labour law section (Tribunale di Milano, sezione lavoro)
Date of the decision	04.07.2018
Type of proceeding	Judgment of merit (sentenza)
Claimant	Employee
Grade	First
Short description of facts	The applicant was a registered rider for Foodinho S.r.l. for four months, when the employer withdrawn from the contract, imposing him – for the prosecution of the relation - to subscribe a new contract. It was never signed by the contrclaimants, however the old

	contract continued until the employer interrupted definitely the relation after an occupational injury happened to the worker.
Legal issues raised	The applicant took legal action to establish the existence of a salaried practice between him and Foodinho S.r.l., asking the reintegration as well to work and the establishment of the illegitimacy of the withdrawal or dismissal. The request was based on the fact that the rider worked continuously for four months, for eight hours a day and for seven days a week. He also claimed that he has respected directives and orders about the tasks and the delivery given by the employer.
Decision	In favour of Employer
Issues concerning digital rights	The Court rejected the claim because evidences showed that the rider could decide *if* and *when* to provide his performance. According to the decision, such an element is incompatible with a salaried practice, characterized by the obligation to work in a given timeslot. The circumstance that, after offering their availability, riders should respect specific rules about the execution of the job, did not mean that the riders couldn't choose whether to work or not.

Case n.	10
Court / Authority	Court of Palermo, labour law section (Tribunale di Palermo, sezione lavoro)
Date of the decision	24.11.2020
Type of proceeding	Judgment of merit (sentenza)
Claimant	Employee
Grade	First
Short description of facts	The applicant worked as a rider for Foodinho S.r.l. from 28.09.2018 until 03.03.2020, when he was disconnected from the platform and never connected again, without any justified reason.
Legal issues raised	The Court granted the application because it noticed that the freedom of the rider to decide if and when to work was only apparent: indeed, he could work only in the time slots made available by the platform, and moreover the deliveries were assigned by the platform through the algorithm, that uses criteria absolutely different from the interest or the preferences of the worker.
Decision	In favour of Employee
Issues concerning digital rights	-

Case n.	11
Court / Authority	Court of Turin, labour law section (Tribunale di Torino, sezione lavoro)
Date of the decision	11.04.2018
Type of proceeding	Judgement of merit (sentenza)
Claimant	Employees
Grade	First
Short description of facts	The applicants worked as riders for Foodinho S.r.l. on the base of multiple temporary contracts of collaboration, until the employer retired from the contract. The job involved the use of a system of remote surveillance of workers.
Legal issues raised	The applicants took legal action to establish the existence of a salaried practice between them and Foodinho S.r.l., asking as well the reintegration to work and the establishment of the illegitimacy of the withdrawal or dismissal. Alternatively, the applicants request the extension to their self of the salaried workers' legal framework, due to the application of art. 2 of the Legislative Decree no. 81/2008. Moreover, the applicants request a compensation for damages because of the data protection violation committed by the employer, which adopted a system of remote surveillance of workers.
Decision	In favour of the Employer.
Issues concerning digital rights	Data protection violation (remote surveillance of workers). The court finds that, despite the information notice provided is generic, there is no violation of data protection regulation and there is no evidence of suffered damage for which is requested a restoration.

Case n.	12
Court / Authority	Court of Appeal of Turin (Corte d'Appello di Torino)
Date of the decision	11.01.2019
Type of proceeding	Judgement of merit (sentenza)
Claimant	Employees
Grade	Second (Appeal of the judgment of the Court of Turin dated 11.04.2018 - cfr. case n. 11)

Short description of facts	Cfr case n. 11
Legal issues raised	The appellants insisted to the claims concerning the existence of a salaried practice and the extension of the salaried workers' legal framework, due to the application of art. 2 of the Legislative Decree no. 81/2008. The claimants didn't raise appeal for the profile of the decision regarding the compensation for damages because of the data protection violation.
Decision	In favour of the Employer
Issues concerning digital rights	About the data protection violation, the Court confirmed the decision of the first grade, because there was no evidence of damage. Moreover, the claimants didn't raise appeal for this profile.

Case n.	13
Court / Authority	Supreme Court (Corte suprema di Cassazione)
Date of the decision	24.01.2020
Type of proceeding	Judgment of legitimacy (sentenza)
Claimant	Employee
Grade	Third (Appeal of the judgment of the Court of Appeal of Turin dated 11.01.2019 - Cfr. case n. 12)
Short description of facts	Cfr. Cases n. 11 e 12
Legal issues raised	The appellants insisted to the claims concerning the extension of the salaried workers' legal framework, due to the application of art. 2 of the Legislative Decree no. 81/2008. The claimants didn't raise appeal for the profile of the existence of a salaried practice neither for the profile of the decision regarding the compensation for damages because of the data protection violation.
Decision	In favour of Employee
Issues concerning digital rights	The Supreme Court established that the job of the riders of Foodinho are due to the collaborations regulated by the art. 2 of the Legislative Decree no. 81/2008 and so that is applicable to the riders the salaried worker legal framework. On the contrary, the Court didn't pronounce about the data protection violation because the claimants didn't raise appeal for this profile.

Case n.	14
Court / Authority	Data Protection Authority (Garante per la protezione dei dati personali)
Date of the decision	10.06.2021
Type of proceeding	Administrative procedure
Claimant	Proceeding started by the Authority on its own initiative.
Grade	First
Short description of facts	From the investigations made by the Commissioner, resulted that Foodinho S.r.l., as the data controller, processed personal data of 18.686 riders in violation of the G.D.P.R. and the Italian legislative decree n. 196/2003.
Legal issues raised	The processing of the personal data made by the company was in violation of the articles 5, par. 1, lett. a), c) e e) (principle of lawfulness, correctness, limitation of conservation); 13 (information notice); 22, par. 3 (suitable tools for the automated treatment of data); 25 (data protection by design and data protection by default); 30, par. 1, lett. a), b), c), f) e g); 32 (preventive measure); 35 (impact evaluation); 37, par. 7 (communication to the control authority of the responsible of data protection); 88 (data protection during the employment relationship) of the GDPR; article 114 (warranties in matter of remote control) of the Italian legislative decree n. 196/2003. The Authority ordered to comply with data protection provisions and fined Foodinho S.r.l. for unlawful data processing with a € 2.600.000,00 sanction.
Decision	Against Foodinho S.r.l.
Issues concerning digital rights	The most important digital right that is treated in this case, is the right to privacy (art. 8 CEDU and GDPR).

Case n.	15
Court / Authority	Data Protection Authority (Garante per la protezione dei dati personali)
Date of the decision	22.07.2021
Type of proceeding	Administrative procedure
Claimant	Proceeding started by the Authority on its own initiative.
Grade	First
Short description of facts	From the investigations made by the Commissioner, it resulted that Deliveroo S.r.l., as data controller, processed personal data of 8.000

	riders in violation of the G.D.P.R. and the Italian legislative decree n. 196/2003.
Legal issues raised	The treatment of the personal data made by the company was in violation of the articles 5, par. 1, lett. a), c) e e) (principle of lawfulness, correctness, limitation of conservation); 13 (information notice); 22, par. 3 (suitable tools for the automated treatment of data); 25 (data protection by design and data protection by default); 30, par. 1, lett. c), f) e g); 32 (preventive measure); 35 (impact evaluation); 37, par. 7 (communication to the control authority of the responsible of data protection); 88 (data protection during the employment relationship) of the GDPR; article 114 (warranties in matter of remote control) of the italian legislative decree n. 196/2003. The Data Protection Authority ordered to comply with data protection provisions and fined Foodinho S.r.l. with a € 2.600.000,00 sanction.
Decision	Against Deliveroo Italy S.r.l.
Issues concerning digital rights	The most important digital right that is treated in this case, is the right to privacy (art. 8 CEDU and GDPR).

Challenges concerning "digital rights" for "platform workers"

Interestingly, the analysis of the Italian jurisprudence on platform workers sheds a peculiar light on the discussion on "Digital Rights". Indeed, from the comparison among them it emerges that technology creates an ecosystem in which are tightened the legal ties among parties, increasing the asymmetries between them, and strengthening the subordination between employer and employees. Moreover, the automation of internal processes and the virtualization of human resources, increases the efficiency of organizational processes, creating a unprecedent dependency by the infrastructure by humans: everyone - workers, customers, third parties - becomes a simple user of a given set of resources. Indeed, in some cases - specifically, those decided in Bologna[377] and Palermo[378] - the discussion is dominated not by theoretical arguments - for example, whether a rider should be considered a self-employed worker or an employee - but by very practical observations based on the analysis on how the management of human resources were controlled by the algorithms governing the platform. For example, in those cases it is thoroughly described the procedure of signing up of new riders, and the huge amount of personal data processed during remote monitoring of the activities performed: e.g. availability especially during peek-hours, efficiency in routing strategies for each delivery, feedback received from clients or customers, and so on.

In the following table it is offered a synthesis of the collected data, showing that the matter of "Digital rights" was discussed in one third of the cases.

[377] Tribunale Bologna sez. lav., 31 dicembre 2020.

[378] Sentenza dei Tribunale di Palermo 12 aprile 2021.

Description	ID	Number	Percentage
Cases analysed	1-15	15	100%
Cases resolved in favour of the platform workers	1, 2, 3, 4, 5, 7, 8, 10, 13	9	60,00
Cases decided in favour of the platform owner	6, 9, 11, 12	4	26,67
Platform owner fined	14, 15	2	13,33
Cases not involving digital rights	1, 2, 3, 4, 5, 6, 10, 12, 13	9	60,00
Cases involving digital rights	7, 8, 11, 14, 15	5	33,33
Cases involving privacy (art. 8 CEDU)	11, 14, 15	3	20,00
Cases involving discrimination (art. 14 CEDU)	7, 8	2	13,33

This is a relevant information showing that recent Italian jurisprudence seems keen - if not ready - for a full understanding of the meaning of "digital rights" in this context.

The main concern is related to data protection. In this sense it is interesting to note that discrimination is considered a specific risk included in this field, even if there is one decision which isolates discrimination as a separate issue.

Conclusions and policy recommendations

According to the results of this analysis, and following the arguments spent in the decisions above analysed, we can observe that current legislation and ordinary judicial remedies are insufficient to solve a complex matter such as the legal issues concerning "platform workers", especially those including "Digital rights". Of course, in this effort laymen cannot be left alone, being required the support from international institutions ('OECD Framework for the Classification of AI Systems' 2022), and from the community of experts (Schwartz et al. 2022) in order to assess the risks posed by artificial agents. In the EU context, for example, it is worthwhile to be mentioned an attempt to prepare the grounds for the future "Artificial Intelligence Law" (Floridi et al. 2022).

Besides ordinary legislation and judicial proceedings, a few alternative approaches can be envisioned, suitable to offer solutions that can bring a lasting benefit for workforce but also for platform providers and in general for the society.

Technological design.

The incorporation of ethical values in technological devices has been officially recognized by the EU legislator with article 25 of GDPR, regulating the privacy "by design" and "by default" approach. In this sense, it could be possible to incorporate the protection of Digital rights directly into the algorithm governing the platforms, in order to provide built-in operating mechanisms of trade union negotiation and assistance. International guidelines and collection of best practices could help.

Collective bargaining agreements.

Rather than an *a priori* legislation which, being general and abstract, leaves *per se* too wide margins of interpretation, or an *ex post* judicial proceeding which, being expensive and uncertain, cannot be pursued by many workers, it could be useful to include a more binding and specific regulation using collective

agreements between union workers and employers' associations. Of course, in this sense public institutions play a fundamental rule of intermediation in order to avoid abuses and "false flag" strategies as those adopted in one of the Italian cases.

Local arrangement and code of conducts.

Since the services provided by such kind of platforms are strongly territorial (e.g. delivery), it could be an opportunity for municipalities to step up and regulate some specific aspects that could improve significantly the quality of jobs of platform workers (e.g. creating stations or offering shed zones for riders waiting for a call) or assisting workers and employers in adopting voluntary codes of conduct that could not only improve the quality of jobs, or raise the productivity of platforms, but also provide benefit for the whole community[379].

Regulatory sandboxes and living labs.

The concept of "regulatory sandbox" is included in the EU proposal called "Artificial Intelligence Act"[380] (articles 53 and 54), which intend to provide regulation concerning the processing of data and the security measures to safeguard the deployment of artificial agents. In this sense, this tool could be used to establish provisional legal frameworks in order to experiment new forms of regulations and models of interaction suitable to protect the "Digital Rights" of "platform workers"[381].

Acknowledgements

This report addresses the topic discussed in the 11th Webinar included in the GDHRNet Highlights Lectures, Tuesday 1st Feb 2021 (13-14 CET), "The Platform Economy and Human Rights", Janneke Gerards (Utrecht University), Valeria Filì and Federico Costantini (University of Udine).

This document is the output of a research activity jointly done by the authors. Section I and II have to be attributed to Federico Costantini, Section III to Alan Onesti, Section IV, V, and VI to both.

References

The references are mentioned using Zotero and the citation style Chicago 17th.

Codagnone, Cristiano, Federico Biagi, and Fabienne Abadie. 2016. 'The Passions and the Interests: Unpacking the "Sharing Economy"'. JRC Science for Policy Report EUR 27914 EN. JRC Science for Policy Report. Institute for Prospective Technological Studies. https://doi.org/10.2791/474555.

Filì, Valeria, and Federico Costantini, eds. 2019. *Legal Issues in the Digital Economy. The Impact of Disruptive Technologies in the Labour Market*. ADAPT Labour Studies 15. Newcastle-upon-Tyne: Cambridge Scholars Publisher. http://public.eblib.com/choice/PublicFullRecord.aspx?p=5848611.

Floridi, Luciano, Matthias Holweg, Mariarosaria Taddeo, Javier Amaya Silva, Jakob Mökander, and Yuni Wen. 2022. 'CapAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act'. SSRN Scholarly Paper ID 4064091. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.4064091.

[379] Cfr. "Carta dei diritti fondamentali del lavoro digitale nel contesto urbano" (c.d. Carta di Bologna) of 2018.

[380] Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts, COM/2021/206 final, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

[381] It is noteworthy that a similar attempt is being made by an ongoing research project https://wetransform-project.eu/.

Floridi, Luciano, and Phyllis Illari. 2014. *The Philosophy of Information Quality*. Synthese Library. Berlin-Heidelberg: Springer.

Gawer, Annabelle, and Nick Srnicek. 2021. *Online Platforms: Economic and Societal Effects*. European Parliament - European Parliamentary Research Service - Scientific Foresight Unit. https://doi.org/10.2861/844602.

Gerards, Janneke, and Raphaële Xenidis. 2021. 'Algorithmic Discrimination in Europe'. 978-92-76-20746–7. Luxembourg: Publications Office of the European Union. https://doi.org/10.2838/544956.

Malgieri, Gianclaudio, and Giovanni Comandé. 2017. 'Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation'. *International Data Privacy Law* 7 (4): 243–65.

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. 'The Ethics of Algorithms: Mapping the Debate'. *Big Data & Society* 3 (2): 205395171667967. https://doi.org/10.1177/2053951716679679.

'OECD Framework for the Classification of AI Systems'. 2022. 323. OECD DIGITAL ECONOMY PAPERS. OECD.

Pesole, A., E. Fernández-Macías, C. Urzí Brancati, and E. Gómez Herrera. 2019. 'How to Quantify What Is Not Seen? Two Proposals for Measuring Platform Work'. Siviglia: European Commission.

Pesole, A., M.C. Urzí Brancati, E. Fernández-Macías, F. Biagi, and I. González Vázquez. 2018. *Platform Workers in Europe*. Vol. JRC112157. EUR 29275 EN. Luxembourg: Publications Office of the European Union. https://doi.org/10.2760/742789.

Schwartz, Reva, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. 2022. 'Towards a Standard for Identifying and Managing Bias in Artificial Intelligence'. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.1270.

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. 2017. 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation'. *International Data Privacy Law* 7 (2): 76–99.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2021. 'Why Fairness Cannot Be Automated: Bridging the Gap between EU Non-Discrimination Law and AI'. *Computer Law & Security Review* 41. https://doi.org/10.1016/j.clsr.2021.105567.

Part IV: Platforms and Elections

The Legal Framework of Online Parliamentary Election Campaigning - An Overview of the Legal Obligations of Parties and Platforms in Germany and the EU

Matthias C. Kettemann, Vincent Hofmann, Mara Barthelmes, Nicolas Koerrenz, Lena Marie Hinrichs and Linda Schleif[382]
LEIBNIZ INSTITUTE FOR MEDIA RESEARCH | HANS-BREDOW-INSTITUT AND ALEXANDER VON HUMBOLDT INSTITUTE FOR INTERNET AND SOCIETY

Introduction

The election campaign for the 2021 German federal election has taken place to a particularly high degree on the internet, especially on the pages of large online platforms such as Facebook, Twitter, YouTube or Telegram, mainly due to the Corona pandemic. This online election campaign is regulated by norms from various legal sources, which define various rights and obligations for online platforms as well as for content creators in general and for political parties in particular.

On the side of the sources of law, a distinction must be made between norms of private and state orders. The latter regulate online communication on different legal levels: Under international law, the European Convention on Human Rights, among others, grants freedom of expression, which could be interfered with if content or profiles were deleted. According to the UN Guiding Principles on Business and Human Rights (Ruggie Principles), companies must also respect human rights. Under European law, a number of legislative projects have been launched with the aim of regulating large online platforms. The draft Digital Services Act provides for particularly extensive obligations.[383] The German Basic Law protects in Article 5 (1), as does the ECHR in Article 10 (1), the freedom of expression of content authors, but also property (Article 14 (1) of the Basic Law, Article 1 of the 1st Additional Protocol to the ECHR) and the entrepreneurial freedom (Article 12 (1) of the Basic Law; Article 16 ECHR) of the platforms. The German Penal Code, which makes insulting content or content inciting to criminal offences a punishable offence, the German State Media Treaty (MStV), which obliges platforms to maintain a diverse public debate, especially through transparency obligations, and the German Network Enforcement Act (NetzDG), which provides regulations for the consistent deletion of illegal content, are of importance for the online election campaign.

Alongside these norms of state orders are the norms of private orders. These are, in particular, the terms of use of the online platforms, but also self-commitments of the parties to conduct themselves in the online election campaign. The terms of use of the major online platforms have become increasingly important for

[382] This article is a translated and updated version of Kettemann et al., Der rechtliche Rahmen des Online-Bundestagswahlkampfs. Ein Überblick über die rechtlichen Verpflichtungen von Parteien und Plattformen, Superwahljahr Paper 02/2021, https://leibniz-hbi.de/de/blog/rechtlicher-rahmen-des-online-bundestagswahlkampfs

[383] European Commission, "Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a Single Market for Digital Services (Digital Services Act) and Amending Directive 2000/31/EC", COM/2020/825 final.

online election campaigns as the platforms have become more relevant for shaping public opinion. If, for example, a candidate or elected representative is deprived of the possibility to spread their views and goals through the social network because of violations of the terms of use, this can have serious consequences for the electoral success of the blocked person. Facebook and Twitter, among others, have recognised their responsibility in the context of elections and have set specific rules for political content.

This interplay of norms of state and private orders reveals both problems of democratic theory, for example in the deletion of profiles (keyword: Donald Trump) by social networks, as well as practical deficits, for example in the containment of false news.

National framework

Legal framework of online party communication in election campaigns

Communication through government agencies

Political communication through the channels of government agencies is severely restricted. In the past, it was even assumed that the public opinion-forming process was completely free of the state, in which the state and its representatives were not active participants in the discourse. The only intermediaries between the state and the people were the media, protected under Article 5(1) sentence 2 of the Basic Law, and the political parties, privileged under Article 21(1) of the Basic Law. In the meantime, the assumption has solidified that state communication is also an integral part of public discourse (in the form of public relations, press work, warnings, etc.) and is sometimes even explicitly provided for in public law norms. Communication conducted by state agencies, as sovereign representatives, enjoys both special reach and a heightened level of trust.[384] State communication may therefore not be used for any purpose and is limited in questions concerning the "what" and "how" of communication: The statement of the governmental body must be related to the respective task assignment and responsibility, which, depending on the institution, can encompass a broad (e.g. government with state management authority) or narrow (e.g. specialised authority) field of topics and contents.

State communication must be factual, correct in content and disseminated with restraint (factuality and correctness requirement). In addition, state agencies must communicate clearly and completely, and the state must always be recognisable as the author of a statement (communicator clarity). These requirements are an expression of the principles of democracy and the rule of law. In the case of information provided by state authorities that interferes with fundamental rights (e.g. warnings about products such as e-cigarettes), the principle of proportionality must also be observed. Furthermore, the requirement of party-political neutrality applies to state information activities, which was elaborated in particular in the context of court proceedings regarding public criticism of the German parties AfD and NPD. However, state actors do not have to act in a completely neutral manner, as the respective office holders can also defend their office politically and stand up for free democratic basic values. [385]

[384] Wissenschaftlicher Dienst des Deutschen Bundestages, "Politische Äußerungen von Hoheitsträgern", 19 March 2018, https://www.bundestag.de/resource/blob/556768/776c7bb3e6cd1fd9ed85e539cca79b59/wd-3-074-18-pdf-data.pdf, p. 3.

[385] BVerfGE 44, 125.

These restrictions on communication only apply to state bodies if they also communicate as state bodies.[386] Therefore, there are no clear legal restrictions for opposition parties with regard to the use of communication channels, as their statements are not state communication. For public officials, an assessment of the overall circumstances of the communication must be made to determine whether it is to be classified as a state or private statement. In particular, the content, location and context of the message are relevant. A private account of a public official on a platform is not sufficient. If it emerges from the consideration of the overall circumstances that the public official is expressing himself/herself in his/her party-political function, this is protected by the extensive freedom of Art. 5, Art. 21 GG. If, on the other hand, it emerges that the statement was made in a state function, the person is bound by the rule of law under Articles 20 (3) and 1 (3) of the Basic Law.

Overall, the communication of public officials through the channels of the office is very limited and therefore not very suitable for election campaign purposes.

Communication through parties

The legal framework for the conduct of online election campaigning results for the parties from different norms that establish both rights and obligations of the parties.

German Basic Law

The right to freedom of expression also applies to political parties under Article 5 (1) of the Basic Law. In principle, this protects both the parties' online presence and their election advertising. This protection is also derived from Article 21 of the Basic Law, which constitutes the right of parties to participate in the formation of political will. However, according to the current interpretation of Article 21 of the Basic Law, the right to place election advertisements, which is concretised in the State Media Treaty (§ 68 para. 2 MStV), only applies to broadcasting organisations.[387] Other platforms are free in terms of content; here there is no entitlement of the parties to election advertising or presence.

The parties' freedom of expression under Article 5(1) of the Basic Law is limited, among other things, by the principles of electoral law under Article 38(1) of the Basic Law. All voters must be able to make their free choice of a political idea on an informed basis.[388] This is not guaranteed if a party spreads false information. This behaviour is then, as a rule, no longer covered by the freedom of opinion under Article 5(1) of the Basic Law because of the collision with Article 38(1) of the Basic Law. The same could apply to the use of social bots, because here it is a matter of deception about the support regarding a political view, which could also disturb the informative basis of the voter in the case of an increased extent of the use of such social bots.[389]

---

[386] Wissenschaftlicher Dienst des Deutschen Bundestages, "Politische Äußerungen von Hoheitsträgern", 19 March 2018, https://www.bundestag.de/resource/blob/556768/776c7bb3e6cd1fd9ed85e539cca79b59/wd-3-074-18-pdf-data.pdf, p. 3.

[387] Cf. Wissenschaftlicher Dienst des Deutschen Bundestages, "Parteienwerbung in privaten Medien", https://www.bundestag.de/resource/blob/651780/3fe16363e541588a2dcbdb3d8b851375/WD-10-044-19-pdf-data.pdf, 5 July 2019, p. 7.

[388] Klaas, Arne, "Demokratieprinzip im Spannungsfeld mit künstlicher Intelligenz", MMR 2019, 84, p. 88.

[389] Ibid.

The dissemination of untrue content about other politicians is also regularly no longer covered by freedom of expression: Even in a political context, the freedom of expression of the person spreading the statement is secondary to the protection of the honour of the person concerned in the case of untrue content.[390]

Simple legal regulations

At the level of simple legal regulations (laws ranking below the constitution), the German Criminal Code sets limits to online election campaigning. In the context of online communication, criminal offences such as insult (§§ 185 ff. StGB) or incitement of the people (§ 130 StGB) are particularly relevant. Especially in online election campaigns, the offences of voter coercion (section 108 StGB) or voter deception could be fulfilled.

With regard to the financing of parties in general and of the election campaign in particular, the German Political Parties Act (PartG) prescribes certain transparency obligations, such as the filing of an accountability report (§ 23 PartG), which must also contain the expenses related to the election campaign (§ 24 para. 5 no. 2 lit. c PartG).

The European General Data Protection Regulation (GDPR) and the German Federal Data Protection Act (BDSG) limit the powers and duties related to the collection of data from users. For example, the processing (definition in Art. 4 No. 2 of the GDPR) of data containing political opinions of the person is only permitted in rare exceptional cases (Art. 9 of the GDPR). However, such an exception may already be the consent of the data subject. The creation of a political personality profile is also limited by said norms.[391]

Legal framework for platforms in online election campaigns

NetzDG

On 1 October 2017, the German Network Enforcement Act (NetzDG) came into force. The law is intended to improve law enforcement in social networks and lead to the consistent deletion of criminal content. The background was also the increasing spread of such content in the digital space and the experiences from the 2016 US election campaign.[392]

The core element of the NetzDG is the obligation for operators of social networks to remove "obviously illegal content" within 24 hours after it has been reported by users and to remove illegal content from the platform after 7 days. Important for the moderation practice are the obligation to set up complaint possibilities and regular reporting on the deletion practice according to the NetzDG. However, this obligation only takes effect if the reported content fulfils the facts of a criminal norm mentioned in § 1 (3). § 1 (3) mentions, among other things, incitement to commit a criminal offence, insult or the dissemination of signs of anti-constitutional organisations. Disinformation in particular rarely meets the elements of such criminal offences.

In addition, the NetzDG is only applicable to platforms that are operated with the intention of making a profit (§ 1 (1) sentence 1 NetzDG) and, in principle, not to services that serve individual communication (§

[390] BVerfG, NJW 2000, 3485.

[391] Klaas, Arne, "Demokratieprinzip im Spannungsfeld mit künstlicher Intelligenz", MMR 2019, 84, p. 90.

[392] Cf. BT-Drs. 18/12356, p. 1, http://dipbt.bundestag.de/dip21/btd/18/123/1812356.pdf.

1 (1) sentence 3 alt. 1 NetzDG). With reference to these two exceptions, Telegram, for example, evades the rules of the NetzDG. Whether Telegram, with groups of up to 200,000 members and public channels with an unlimited number of possible subscribers[393] , actually differs so much from other platforms such as Facebook that a non-application of the NetzDG is justified, seems questionable. Also, according to its own statements, Telegram does not aim to make a profit.[394] On the other hand, the network also plans to monetise content according to its own statements, but without wanting to make a profit.[395]

State Media Treaty

The German Interstate Treaty on Broadcasting (RStV) was replaced by the German State Media Treaty (MStV) on 7 November 2020. This is intended to adapt the legal framework to the changed conditions in response to the progressive digitisation of the media landscape.[396] To this end, the media concept of the RStV, which focuses on broadcasting, has been replaced by the term media platform (§ 2 No. 14 MStV), which includes all services that process media content into an overall offer, regardless of the technology used to distribute the content.

In addition, the MStV also includes so-called media intermediaries (§ 2 No. 16 MStV), which also sort journalistic-editorial content and present it in a generally accessible way without combining it into an overall offer. These include, for example, search engines or social networks, which, as described above, have a strongly growing influence on the formation of public opinion.[397]

One aim of the MStV is to ensure diversity of opinion and communicative equality of opportunity in the media landscape.[398] This goal is to be achieved through comprehensive obligations for media intermediaries insofar as they have an influence on the formation of public opinion.[399] In this case, they are obliged, among other things, to disclose the functioning of their algorithms and may not determine them in such a way that individual journalistic-editorial content is systematically and without justification disadvantaged (§§ 93(1), 94(1)). Social bots must also be labelled as such (§ 18(3)). The state media authorities can enact statutes to concretise the requirements for media intermediaries. These will be of great importance for the further development of the effectiveness of the MstV.

The responsibility for checking infringements also lies with the state media authorities. They can receive notifications of suspected violations from media providers. How many such notifications the state media authorities have received in the context of the 2021 federal election and how quickly they can process them will significantly determine the success of the MStV in securing diversity of opinion in social networks.

---

[393] Telegram, Q&A, "What is the difference between groups and channels?", https://telegram.org/faq/de#f-was-ist-der-unterschied-zwischen-gruppen-und-kanalen.

[394] Telegram, Questions and Answers, "Why should I trust you?", https://telegram.org/faq/de#f-warum-sollte-ich-euch-vertrauen.

[395] Telegram, Q&A, "How will Telegram make money?", https://telegram.org/faq/de#f-wie-wird-telegram-geld-verdienen.

[396] Martini, BeckOK Informations- und Medienrecht, MStV Preamble, marginal no. 43.

[397] Cf. Die Medienanstalten, "Intermediäre", https://www.die-medienanstalten.de/themen/intermediaere.

[398] Cf. explanatory memorandum to the MStV, LT-Drs. N 18/6414, 89, https://www.landtag-niedersachsen.de/drucksachen/drucksachen_18_07500/06001-06500/18-06414.pdf.

[399] Martini, BeckOK Informations- und Medienrecht, MStV Preamble, marginal no. 45.

European legislative projects

Digital Services Act (DSA)

On 15 December 2020, the EU Commission presented its draft Digital Services Act (DSA). As the successor to the E-Commerce Directive, the DSA is intended to contribute to secure and trustworthy online communication in the form of an EU regulation (Art. 1 para. 2 lit. b).[400]

According to the draft, intermediaries remain liable for illegal content as soon as they become aware of it. In order to enable such knowledge to be gained, an efficient complaints management system must ensure that reports of suspected illegal content are processed quickly and reliably. Similarities to the German NetzDG can be seen here.[401] The transparency obligations for intermediaries are also to be expanded: Among other things, they must make their guidelines for moderating and restricting content publicly available.

Art. 24 of the draft contains the obligation of online platforms to make information about the displayed online advertising transparent. For users, it should be clear, unambiguous and recognisable in real time for each individual advertisement whether and whose advertisement is being displayed. In addition, the most important parameters for determining the target group must be visible. This should make personalised advertising and especially the controversial microtargeting used for election campaigns more transparent.[402]

In addition, platform providers are obliged to submit annual transparency reports, which must include information on deleted content as well as on recommendation algorithms.

Another important regulatory subject of the DSA are the "Very Large Online Platforms" (VLOPs). These are "systemically relevant" platforms with at least 45 million monthly active users in Europe. The Commission considers such platforms to have a particular influence on public debates and sees them as important sources of information in the context of public opinion-forming. The Commission also sees risks in the reach of these platforms, such as the dissemination of illegal content, manipulation or restrictions on the exercise of fundamental rights, especially freedom of expression.[403] The DSA provides for possible measures to minimise these risks, e.g. publicly accessible archiving of all advertisements placed (Art. 30), increased content moderation (Art. 27) and the obligation to conduct an annual independent audit to determine whether the requirements of the DSA have been met (Art. 28).[404]

The topics of fake news and disinformation, on the other hand, if they do not fall under illegal content anyway, are hardly regulated in the DSA.[405]

---

400 European Commission, 2020/0361 (COD), p. 49, https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:52020PC0825&from=de.

401 Berberich/Seip, "Der Entwurf des Digital Service Act", GRUR-Prax 2021, 4 (5), https://beck-online.beck.de/Bcid/Y-300-Z-GRURPRAX-B-2021-S-4-N-1.

402 Ibid. , 4 (5), https://beck-online.beck.de/Bcid/Y-300-Z-GRURPRAX-B-2021-S-4-N-1.

403 Ibid.

404 European Commission, 2020/0361 (COD), p. 68, https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:52020PC0825&from=de.

405 Berberich/Seip, Der Entwurf des Digital Service Act, GRUR-Prax 2021, 4 (7), https://beck-online.beck.de/Bcid/Y-300-Z-GRURPRAX-B-2021-S-4-N-1.

Digital Markets Act (DMA)

The draft Digital Markets Act (DMA), which was presented at the same time as the DSA, is directed against the business practices of so-called gatekeepers. These are companies that bring together a large number of customers with a large number of companies and exert a significant influence on the EU internal market over a certain period of time (Art. 3 para. 1 DMA-E). This also includes the social networks that are particularly relevant in the context of opinion-forming on the Bundestag elections. The large networks enjoy their supremacy in particular through the enormous amounts of data of their users, which enable them to offer an optimal range of products. In order to dissolve this data-based supremacy, gatekeepers should be restricted in their use of data and obliged to pass on the data. For example, data obtained from commercial users may not be used in competition with these users. There is also an obligation to pass on data on the behaviour of customers and persons confronted with advertisements to the companies that have placed the advertisement or sold the product.

Data Governance Act (DGA)

In addition to its advantages for consumers, data protection, which has been strengthened by the General Data Protection Regulation (GDPR) since 2018, has also brought significant obstacles for companies. In 2020, for example, in a survey by the industry association Bitkom, a good half of the companies surveyed stated that their innovative strength was limited by the GDPR.[406] The difficulties associated with the GDPR are to be reduced by the Data Governance Act (DGA), the draft of which was presented on 25 November 2020, in the form of a European regulation. While maintaining the standards of the GDPR, it is intended in particular to strengthen the confidence of users in the security of the processing of personal data and thus increase the possibility of data use by companies and research institutions.[407]

The core element of the draft is the requirement that personal data should not be anonymised and held in trust by the internet companies themselves, but by a neutral body. The supervision of this intermediary body is the responsibility of a supervisory authority created by the member state in the headquarters of the intermediary.[408]

The DGA also creates the legal basis for so-called data altruism. Through a release valid for the entire EU, all users can voluntarily make their data available for purposes of general interest. Monitoring of data use is also the responsibility of the national supervisory authorities.[409] By balancing data protection on the one hand and the interest of research and industry in the use of data on the other, the DGA aims to make Europe the "most data empowered continent", as EU Commissioner Thierry Breton said at the presentation of the draft.[410]

---

[406] Bitkom, "Every 2nd company forgoes innovation for data protection reasons", https://www.bitkom.org/Presse/Presseinformation/Jedes-2-Unternehmen-verzichtet-aus-Datenschutzgruenden-auf-Innovationen.

[407] European Commission, "Proposal for a Regulation on European Data Governance (Data Governance Act", https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-data-governance-data-governance-act.

[408] Rieke, EU Data Governance Act, bvdw.org, https://www.bvdw.org/der-bvdw/news/detail/artikel/eu-data-governance-act/.

[409] Ibid.

[410] Ibid.

Private ordering

Facebook

The Facebook platform has formulated community standards to protect democratic and liberal values. These prohibit the manipulation of elections. Spreading support for violence around elections and voter registration are also prohibited.[411] Content glorifying and supporting violent or terrorist acts will be removed from Facebook.[412] Furthermore, the misrepresentation of election data, especially regarding candidates, locations and eligibility to vote, is not permitted. Facebook also prohibits its users from influencing elections, especially in the form of intimidation and demonstrations of power.[413] Facebook does not allow political exclusion that denies someone the right to political participation.[414] Facebook also does not allow political influence through coordinated, non-authentic behaviour on behalf of a foreign actor or government agency.[415]

In addition, videos created or manipulated by artificial intelligence or deep learning technology (so-called "deep fakes") are inadmissible if they lead to a false assumption about what a person is saying. Parodies and satire are exempt from this.[416]

In the help section for businesses, Facebook informs about the requirements for elective advertising. Advertisers must complete an advertising authorisation process and the ad must include a disclaimer stating who is paying for the ad. It should also be easier to see who owns a Facebook page and how certain media may be politically influenced by the state.[417] In addition, the ads will be stored in an ad library to create more transparency.[418]

In general, advertisements must not contain lurid or discriminatory content. In addition, no misinformation should be displayed or controversial social and political issues used for commercial purposes.[419]

Facebook itself creates the problem that users are primarily shown information that is close to their own interests due to recommendation logic. Facebook's terms of use state that all available data on Facebook products and outside will be used for personalisation.[420] Based on the belief that stronger connections create better communities, Facebook uses this data to show targeted people, events and other content that is

[411] Facebook (ed.), Community Standards, I.1., https://de-de.facebook.com/communitystandards/credible_violence.

[412] Facebook, "Community Standards", I.2.

[413] Facebook, "Community Standards", I.3.

[414] Facebook, "Community Standards", III.12.

[415] Facebook, "Community Standards", IV. 20.

[416] Facebook, "Community Standards", IV.22.

[417] Kühl, Eike, "Wie groß ist Facebooks Macht im Wahlkampf", 22.09.20, ZeitOnline, https://www.zeit.de/digital/internet/2020-09/facebook-manipulation-waehler-us-wahl-regeln-soziale-medien, p. 2.

[418] Facebook, "Info on election advertising or advertising on politically or socially relevant topics", https://de-de.facebook.com/business/help/167836590566506?id=288762101909005.

[419] Facebook, "Advertising Policy Prohibited Content," 3, 11, 13, 14, https://www.facebook.com/policies/ads/prohibited_content.

[420] Facebook, "Terms of Use", 1. Services We Offer. We provide you with a personalised experience, https://de-de.facebook.com/terms.

related to its own content and interests.[421] Advertisements for services and products of companies and organisations are controlled by matching personal data about users' interests and activities with the target audience specified by the company.[422] By controlling the content presented in this way, the discourse can appear less diverse.[423]

Twitter

Twitter has specifically issued a policy on the integrity of civic processes. It prohibits the manipulation and interference with elections. This includes, in particular, misleading information about how to vote and election results, as well as attempts to suppress and intimidate voters. In the run-up to elections in the respective country, Twitter activates a special reporting function through which tweets from all users with a suspicion of a violation of these special rules can be reported. Reported tweets can be flagged or deleted if violations are found. In the case of serious or repeated violations, the user's entire account can be blocked. Polarising or controversial points of view as well as inaccurate statements about parties, representatives or candidates are explicitly not a violation.[424]

Accounts of governments or state-affiliated media companies are uniformly marked on Twitter by means of a small flag symbol in the status of the account. Twitter does not promote or recommend accounts or tweets marked in this way.[425] In addition, synthetic or manipulated content can be marked as such and its visibility restricted in order to avoid misleading users.[426]

Regardless of elections, harassment and intimidation are prohibited on Twitter. In the event of violations, the platform can ask users to delete the content in question and temporarily put the account into read-only mode or block it permanently.[427] Twitter provides even stricter measures against the threat and glorification of violence. Potential violations can be reported not only by users, but by anyone. "Any account that posts threats of violence will be immediately and permanently banned.[428]

YouTube

Similar to Facebook and Twitter, misinformation about elections and technically manipulated content that misleads users are prohibited on YouTube. YouTube also prohibits YouTube also prohibits inaccurate statements about the eligibility of political candidates or the legitimacy of incumbent government officials. Attempts to disrupt or obstruct an election are also prohibited.[429] To

---

[421] Facebook, "Terms of Use", 1. Services We Offer. We connect you with people and organisations you care about, https://de-de.facebook.com/terms.

[422] Facebook, "Terms of Use", 1. Services we offer. We help you discover content, products and services that may be of interest to you and 2. How our services are funded, https://de-de.facebook.com/terms.

[423] Bundestag Drucksache 19/24200, 11.11.2020, p. 58 f.; Schmidt, "Soziale Medien. Eine Gefahr für die Demokratie?", 11.05.2019, https://www.bmbf.de/de/soziale-medien-eine-gefahr-fuer-die-demokratie-8606.html.

[424] Twitter, "Civic Process Integrity Policy", as of 01.2021, https://help.twitter.com/de/rules-and-policies/election-integrity-policy.

[425] Twitter, "Information on labels on Twitter accounts of government officials and labels of state media", https://help.twitter.com/de/rules-and-policies/state-affiliated.

[426] Twitter, "Policy on Synthetic and Manipulated Media", https://help.twitter.com/de/rules-and-policies/manipulated-media.

[427] Twitter, "Abusive Behaviour", https://help.twitter.com/de/rules-and-policies/abusive-behavior.

[428] Twitter, "Threat of Violence Policy", https://help.twitter.com/de/rules-and-policies/violent-threats-glorification.

[429] YouTube, "Policy on Spam, Deceptive Practices and Fraud", https://support.google.com/youtube/answer/2801973.

prevent fraudulent interference in elections, YouTube works with Google's Threat Analysis Group, other technology companies and law enforcement agencies.[430] YouTube will remove film, image and audio material relating to the aftermath of terrorist attacks or other acts of violence.[431]

To support politically serious sources, the platform shows priority trustworthy content for news and info topics under "Next Videos". In addition, journalistically high-quality content can be highlighted.[432]

Similar to Twitter and its flag icons, details of funding sources are displayed below videos of public or government-funded sites.[433] For election ads, the funding of the ad must be disclosed in the public transparency report.[434] All rules set by YouTube explicitly apply regardless of the political orientation of the content or user.[435]

Telegram

Telegram attaches great importance to the privacy of its users. The platform categorically refuses to process requests regarding (illegal) non-public content. The content of (group) chats is "a private matter for the respective users".[436] Only publicly accessible content can be reported by email or directly via the user's profile.[437] According to Telegram, group chats are intended for families, friends and small teams. However, the permitted group size of up to 200,000 members and the possibility of making groups public,[438] , speaks in favour of a wide-ranging use. The platform makes it clear that it does not support political censorship and that statements critical of the government may therefore also be disseminated on Telegram. Although copyright-infringing, terrorist and pornographic content is to be blocked, the platform wants to give space for the dissemination of "alternative[r] opinions".[439]

In order to make it more difficult to oblige the handover of data, Telegram uses different data centres around the world for its cloud chats. Due to the different jurisdictions, several court orders from different countries would be necessary to oblige Telegram to hand over data.[440]

---

[430] YouTube, "Security and Election Policies, Foreign Influence", https://www.youtube.com/intl/ALL_de/howyoutubeworks/our-commitments/supporting-political-integrity/#foreign-interference.

[431] YouTube, "Violent or Cruel Content Policy", https://support.google.com/youtube/answer/2802008?hl=de&ref_topic=9282436.

[432] YouTube, "Security and Election Policies, News and Information on Elections", https://www.youtube.com/intl/ALL_de/howyoutubeworks/our-commitments/supporting-political-integrity/#election-news-and-information.

[433] Ibid.

[434] YouTube, "Security and Election Policies, Political Advertising", https://www.youtube.com/intl/ALL_de/howyoutubeworks/our-commitments/supporting-political-integrity/#political-advertising.

[435] YouTube, "Security and Election Policies, Remove Content", https://www.youtube.com/intl/ALL_de/howyoutubeworks/our-commitments/supporting-political-integrity/#removing-content.

[436] Telegram, "Questions and Answers, I found illegal content on Telegram. How can I have it deleted?", https://telegram.org/faq/de#f-ich-habe-illegale-inhalte-auf-telegram-gefunden-wie-kann-ich-d.

[437] Telegram, "Questions and Answers, I found illegal content on Telegram. How can I have it deleted?", https://telegram.org/faq/de#f-ich-habe-illegale-inhalte-auf-telegram-gefunden-wie-kann-ich-d.

[438] Telegram, "Q&A, What's the difference between groups and channels?", https://telegram.org/faq/de#f-was-ist-der-unterschied-zwischen-gruppen-und-kanalen.

[439] Telegram, "Questions and answers, wait a minute. 0_o You delete something at the request of a third party?", https://telegram.org/faq/de#f-ich-habe-illegale-inhalte-auf-telegram-gefunden-wie-kann-ich-d.

[440] Telegram, "Q&A, Do you respond to data requests?", https://telegram.org/faq/de#f-reagiert-ihr-auf-datenanfragen.

In addition, the platform offers the possibility to communicate via so-called secret chats. Due to the end-to-end encryption used only in this area, Telegram itself has no access to the distributed content, according to its own information. In these chats, there is the option to set a so-called "self-destruct timer", after which the messages are deleted from the sender's and recipient's device.[441] Telegram recommends the additional protection of (not end-to-end encrypted) cloud chats by means of a password, if the user has reasons to "doubt your mobile carrier or government"". [442]

Telegram does not take a position on dealing with discriminatory, defamatory or inflammatory content in its terms of use, privacy policy or FAQs. An exception to this is the propagation of violence, which is prohibited according to the published terms of use, but only for the public parts of the platform. However, consequences for violations are not threatened.[443] If Telegram receives information about accounts of terror suspects, the IP address and telephone number can be passed on to the authorities.[444] However, Telegram does not impose an obligation to this effect - for example, on the grounds of protecting the general public. According to Telegram, no such measures have yet been taken.[445]

Telegram also does not make any statements about possible false voting information.[446] In case of reported phishing, spam or other abuse, Telegram may block the responsible profiles or restrict their ability to contact strangers. Telegram also reserves the right to analyse cloud chat data for this purpose using algorithms.[447] Telegram's terms and conditions do not provide for a transparency report along the lines of the NetzDG.

Other orders

Parties can commit themselves to certain behaviour in online election campaigns in order to guarantee specific principles of conduct in addition to the legal regulations and conditions of the platforms.[448] This is what the Greens did in 2017, among others, committing not to use social bots and to publish more detailed information on party donations than required by the Political Parties Act. Disinformation campaigns were also prohibited in the self-imposed election campaign rules.[449]

Overview and comparison

Most of the big platforms have terms of use that are intended to prevent election manipulation. Although they differ in details, their general approach is similar. Only Telegram falls off the grid and also allows

[441] Telegram, "Questions and Answers, Who is Telegram for?", https://telegram.org/faq/de#f-fur-wen-ist-telegram-gedacht; "How do secret chats differ?", https://telegram.org/faq/de#geheime-chats; Telegram, "Privacy Policy, 3.3.2. Secret Chat", https://telegram.org/privacy.

[442] Telegram, "Questions and Answers, How does two-step confirmation work?", https://telegram.org/faq/de#f-wie-funktioniert-die-zweistufige-bestatigung.

[443] Telegram, "Terms of Service", https://telegram.org/tos?setln=de.

[444] Telegram, "Privacy Policy, 8.3. Law Enforcement Authorities", https://telegram.org/privacy#8-3-law-enforcement-authorities.

[445] Ibid.

[446] Telegram, "Questions and Answers", https://telegram.org/faq; "Terms of Service", https://telegram.org/tos?setln=de; "Privacy Policy", https://telegram.org/privacy.

[447] Telegram, "Privacy Policy, 5.3. Spam and Abuse", https://telegram.org/privacy?setln=de#5-3-spam-and-abuse.

[448] Künast, Renate, "Rules for election campaigns in the digital age", ZRP 2019, 62, p. 65.

[449] Bündnis 90 / Die Grünen, "Grüne Selbstverpflichtung für einen fairen Bundestagswahlkampf 2017", 13.02.2017, https://cms.gruene.de/uploads/documents/20170213_Beschluss_Selbstverpflichtung_Fairer_Bundestagswahlkampf.pdf.

potentially illegal, or at least problematic, content in public channels. A summary of the election-related platform rules of Facebook, YouTube, Twitter and Telegram is shown in the following figure.

Election-related platform rules of Facebook, YouTube, Twitter and Telegram

Platform	Facebook	Twitter	Youtube	Telegram
Prohibition of false election data	☑	☑	☑	☒
Labelling and transparency of election advertising	☑	☑	☑	☒
Marking of governmental organisations	☒	☑	☑	☒
Prohibition/marking/limitation of visibility of "social bots" or "deep fakes"	☑	☑	☑	☒
Prohibition of terrorist and violent content	☑	☑	☑	☑

Problem cases

In the run-up to the election, a number of problem areas were already identified which were to be observed during the election campaign. With regard to the State Media Treaty, the enactment, use and effectiveness of the statutes regulating media intermediaries by the state media authorities (§ 96 MStV) were of particular interest. If content or entire profiles are deleted, it must be analysed on the basis of which norm this is done. This could be both the private terms of use of the networks, state norms such as the NetzDG or an interplay of such norms. This goes hand in hand with the question of whether the networks also take the same measures such as deletion, commenting or profile blocking in comparable situations, or whether different decisions and measures are taken even in ideologically comparable groups. In this context, it could be of particular relevance whether the measure is directed against the profile of a candidate. Messenger services such as Telegram could be of particular importance for the dissemination of illegal content. The fact that these services started as private communication spaces and that a (large) part of their platform is also used for this purpose makes it more difficult to take action against illegal content on these services in view of the NetzDG and the draft DSA.

The sum of the measures against published content could also provide information on whether freedom of expression finds different, actual limits online than offline. The question also arises whether disinformation campaigns can be effectively prevented within the framework of existing legal regulations or whether new regulations are needed here.

Summary

The catalogue of norms for online election campaigns has expanded considerably. Not only have new norms of state order been enacted, which place particular responsibility on the operators of the large online platforms. Due to the enormous reach of these networks, their private norms in the form of terms of use, privacy settings or rules of conduct have also taken on a significance that could exceed the significance of state norms in terms of their actual impact. In the German super-election year 2021, all these norms regulate an essential part of the election campaign and thus an essential democratic process. Many of the norms discussed are still young (NetzDG of 2017, MStV of 2020) or awaiting adoption (DSA, DMA, DGA). The

novelty of the norms and the fact that communication on the internet has become enormously important for election campaigns in Corona times call for an analysis of the impact of said norms in order to be able to develop alternatives to the existing regulations on this basis. These effects on the democratic process of opinion-forming will become particularly evident in the German super-election year of 2021.

Part V: Improving Platform Rules

Platform-proofing Democracy - Social Media Councils as Tools to Increase the Public Accountability of Online Platforms

Martin Fertmann and Matthias C. Kettemann[450]
LEIBNIZ INSTITUTE FOR MEDIA RESEARCH | HANS-BREDOW-INSTITUT

Definition of objectives

It is common cause that the evolution of the internet has had an impact on private and public communication behaviour. The internet has become one of the most important tools we use to exercise our rights, especially the right to information and the right to freedom of expression. As the European Court of Human Rights put it in 2015, the internet provides "essential tools for participation in activities and discussions concerning political issues and issues of general interest".[451] The Committee of Ministers of the Council of Europe emphasised that "the internet plays a particularly important role with respect to the right to freedom of expression".[452]

But where exactly does "communication on the internet" take place? Very often on and via *platforms*. We understand platforms to mean service providers offering internet-based Web 2.0 applications, linking user-generated content by means of application-specific user profiles. Platforms regulate access to the online communication space; indeed, they help constitute it. Consequently, the companies providing these services play an important role in the communication framework. This gives rise to certain questions: Given that this form of communication, with its high relevance for democracy, is privately designed and managed, how can citizens influence the rules which determine the limits of what may be said online? How can platforms enable greater citizen involvement in norm-setting, and what would the consequences of doing so be for platforms, states, individuals and societies? In essence: Who is allowed to define the rules which regulate online spaces? Do rules formulated exclusively by platforms exercising their domiciliary right ipso facto suffer a legitimacy deficit? Can platforms be assigned institutions which help ensure greater accountability to society? Can such "social media councils" platform-proof democracy?

With the establishment of the Facebook Oversight Board[453] in 2020, there now exists an example of a social media council which can be analysed – in terms of both its strengths and its weaknesses – as a sample of the institutionalised expression of the desire to integrate external experts in content governance decisions. Do social media councils offer a silver bullet to the challenge of political participation in the digital age?

---

[450] This contribution is based on Kettemann/Fertmann, Platform-Proofing Democracy (Potsdam-Babelsberg: Friedrich-Naumann-Stiftung für die Freiheit 2021/05), https://shop.freiheit.org/#!/Publikation/1084).

[451] EGMR 01.02.2015, Nos 48226/10 and 14027/11, Cengiz and Others vs Turkey, section 49.

[452] Committee of Ministers of the Council of Europe (2018): Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries, 7 March 2018, https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680790e14, S. 2.

[453] Oversight Board, https://oversightboard.com.

Does establishing such councils offer a convincing way of platform-proofing democracy and making the platforms more democratic?

These are the questions that helped establish the scope of this study, which is made up of six parts. We start with an introduction to the challenges of political participation in the digital age, with a particular focus on the normative design of opportunities for participation in setting and enforcing platform norms. This is followed by an overview of the history and conceptualisation of social media councils and an analysis of the Facebook Oversight Board as a paradigmatic social media council . An overview of other social media councils and a concluding summary and appraisal round the study off.

Introduction

In 2014, a workshop on "Public International Law of the Internet", hosted in Berlin by the German Foreign Office, among others, concluded that all digital policy stakeholders without exception were dissatisfied with the status quo: "States are frustrated about being unable to enforce the law on the internet. In the absence of clear and applicable regulations, companies don't know how to deal with (state and private) requests; they are effectively given no choice but to administer justice. Users worry about their data and about violations of their fundamental rights."[454] These frustrations represent a considerable challenge both for the 4.4 billion people who have access to the internet and for the 3.3. billion who do not,[455] as internet governance and access to online content were recognised as being constitutionally relevant topics at an early stage. For example, the United Nations were quick to link democratic constitutionality and development, but also orient the internet towards human development based on constitutional principles. At the UN World Summit on the Information Society (WSIS) (2003, 2005), the states of the world committed themselves to "a people-centred, inclusive and development-oriented Information Society", to be based on the purposes and principles of the Charter of the United Nations, international law and multilateralism, and "respecting fully and upholding the Universal Declaration of Human Rights".[456]

In and of itself, internet access does not lead to more democracy, although the rule of law and high internet access levels are positively correlated. However, the internet can be deployed as an effective means to strengthen civil society engagement. At the same time, and in addition to protecting spaces where individual freedom can be exercised, we also have to secure the societal prerequisites of social cohesion, which represents a considerable challenge in the face of the privatisation of online communication spaces and the dynamisation of online communication (including the renegotiation of "truths", the questioning of shared information assets, the changing of communication practices, and the distribution of media portfolios).[457]

A diffuse sense of unease persists, be it regarding the sharing of disinformation related to Covid-19, or, more recently, when the account of a sitting US president was suspended: The measures taken by platform

[454] Alexander von Humboldt Institut für Internet und Gesellschaft (HIIG), Workshop zu "Völkerrecht des Netzes", 8 September 2014, 7.

[455] Kettemann, Die normative Ordnung der Cyber-Sicherheit. Zum Potenzial von Cyber-Sicherheitsnormen, Normative Orders Working Paper 01/2019; Kettemann, Ein Internet für alle Menschen, Tagesspiegel Background Digitalisierung und KI, 5 June 2019, https://background.tagesspiegel.de/ein-internet-fuer-alle-menschen.

[456] UN Doc. WSIS-05/TUNIS/DOC/7-E, https://www.un.org/depts/german/conf/wsis-05-tunis-doc7.pdf.

[457] See Kettemann, Menschenrechte und politische Teilhabe im digitalen Zeitalter. Expert opinion provided in response to a request by the Committee on Human Rights and Humanitarian Assistance of the German Bundestag (Arbeitspapiere des Hans-Bredow-Instituts, Works in Progress # 2), 17 June 2020, https://leibniz-hbi.de/de/publikationen/menschenrechte-und-politische-teilhabe-im-digitalen-zeitalter.

companies are often welcomed in substance, but their impact on democratic discourse processes and democracy per se is considered a challenge. A comment by German Chancellor Angela Merkel was emblematic of this unease. In a statement on the Trump matter, the chancellor said that it was problematic that important decisions regarding communication rules (and the presence of politicians) in communication spaces were no longer being made by "lawmakers", but by "the managers of social media platforms".[458]

It is not a new insight that all relevant forces should be involved in developing, adopting and enforcing rules in functioning democracies. It helps to counter the concentration of public opinion and the concentration of power in the (communicative) structures in which social innovation is generated. As Germany's Federal Constitutional Court emphasised in 1986 on the topic of the freedom of broadcasting, it is sufficient to transfer "all significant decisions to an external organ that is independent of the state and which is subject to the influence of the relevant social forces and trends" while putting in place effective legal provisions to prevent a concentration of the power to shape public opinion. [459]

Yet currently many decisions with a considerable impact on online communication are essentially being taken by platforms on their own. While it is true that platforms have increasingly constructed their own normative orders as coherently conceived rule sets equipped with narratives to establish legitimacy,[460] they are generally far removed from the demands of the Committee of Ministers of the Council of Europe in its recommendation on internet intermediaries: "The process of drafting and applying terms of service agreements, community standards and content-restriction policies should be transparent, accountable and *inclusive*. Intermediaries should seek to collaborate (…) organisations representing the interests of users and affected parties (…) before adopting and modifying their policies. Intermediaries should seek to empower their users to engage in processes of evaluating, reviewing and revising, where appropriate, intermediaries' policies and practices. (…) Internet intermediaries should make available – online and offline – effective remedies and dispute resolution systems that provide prompt and direct redress in cases of user, content provider and affected party grievances."[461]

It is without any doubt possible to align these requirements with the core responsibilities of states in this context, namely the protection of fundamental and human rights in the digital environment.[462] States not only have the negative obligation of not violating the right to freedom of expression and other human rights in the digital context, but also the positive obligation to protect human rights while creating a regulatory environment for all, in which everybody can exercise these rights.

As most communication spaces on the internet are privately owned, intermediaries, including social media companies, have become important normative actors. Network effects and acquisitions have led to a situation where a relatively small number of important platform companies dominates the market. These

---

[458] Tagesspiegel, 11 January 2021, https://www.tagesspiegel.de/politik/meinungsfreiheit-von-elementarer-bedeutung-merkel-kritisiert-twitter-sperre-fuer-trump/26786886.html.

[459] BVerfG, 4 November 1986, 1 BvF 1/84 (4. Rundfunkentscheidung), https://openjur.de/u/175210.html.

[460] Kettemann/Schulz, Setting Rules for 2.7 Billion. A (First) Look into Facebook's Norm-Making System: Results of a Pilot Study (Hamburg: Working Papers of the Hans-Bredow-Institut, Works in Progress # 1, January 2020), https://leibniz-hbi.de/uploads/media/Publikationen/cms/media/5pz9hwo_AP_WiP001InsideFacebook.pdf.

[461] Council of Europe, Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries, section 11; appendix 2, especially 2.2.2. on enabling users to participate in formulating rules and 2.5 regarding access to effective complaint mechanisms (our emphasis).

[462] Kettemann, The Normative Order of the Internet. A Theory of Online Rule and Regulation (Oxford: Oxford University Press, 2020).

companies have certain obligations under international and domestic law. In accordance with the UN Guiding Principles on Business and Human Rights and the embedded "Protect, Respect and Remedy" framework (UN Guiding Principles, "Ruggie Principles"),[463] intermediaries should respect the human rights of their users (and other affected parties) in all their activities (including in formulating and applying terms of use) and remedy any negative impacts on human rights directly linked to their business activities.

At the global level, rights-based entitlements of individuals to participate in internet governance are being incorporated by means of the increased inclusion of individuals in governance decisions related to the internet.[464] Everybody – and especially citizens – has a democratic interest in participating in the internet and its regulation, in other words a stake, a value-based interest in the process and outcome of regulation, the operationalisation of which requires involving all stakeholders in all phases and normative processes.[465]

But how can the participation entitlements of individuals be realised at a smaller scale, too – by platforms? How can these be made "more democratic"?

The concept of social media councils is a valuable starting point in this debate. In Germany, it can be embedded in the fertile legal context of decades of experience with "council-based" governance in the media sector, including through the broadcasting and television councils of public broadcasters, the media councils of the state media authorities responsible for broadcast and telemedia,[466] and sectoral self-regulating bodies such as press[467] and advertising[468] councils.

Properly understood, social media councils are no utopia of self-regulation in the sense of John Perry Barlow's famous "Declaration of Independence" of cyberspace.[469] They are not meant to *replace* existing models of private and state regulation, but rather *complement* them *"to create an independent, accountable, and transparent mechanism that can cooperate with platforms to improve their own systems and eliminate the need for some regulation."*[470]

Accepting that there are no silver bullets and that incremental improvements are the best one might hope for in the complex regulatory triangle between states, companies and civil society[471] – not least because of

[463] See "Ruggie Principles": Report of the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises, John Ruggie, Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework, UN Doc A/HRC/17/31 dated 21 March 2011 (German version).

[464] German Foreign Office, Recommendation 5A/B, Options for the Future of Global Digital Cooperation, https://www.global-cooperation.digital/GCD/Redaktion/EN/Downloads/options-for-the-future-of-global-digital-cooperation.pdf?__blob=publicationFile&v=2. See Kettemann/Kleinwächter/Senges/Schweiger, Comments on Recommendation 5A/B of the UN High Level Panel on Digital Cooperation, How to Build an Enhanced Mechanism for Digital Cooperation. A Multistakeholder Statement from Germany, 27 April 2020, https://www.global-cooperation.digital/GCD/Redaktion/EN/Downloads/kleinwaechter-kettemann.pdf?__blob=publicationFile&v=2.

[465] Kettemann, Internet Governance, in Jahnel/Mader/Staudegger (eds.), Internetrecht, 4th edition (Vienna: Verlag Österreich, 2020), 47-73.

[466] This refers specifically to the decision-making bodies of the state media authorities and commissions, which are referred to as media councils ("Medienräte") in Baden-Württemberg, Bavaria, Berlin/Brandenburg, Hamburg/Schleswig-Holstein, Saarland and Saxony; for example, see sections 39-47 of the State Media Treaty for Hamburg and Schleswig-Holstein, the "Medienstaatsvertrag HSH".

[467] Deutscher Presserat (German Press Council), https://www.presserat.de.

[468] Deutscher Werberat (German Advertising Council), https://www.werberat.de.

[469] See Barlow (1996): "A Declaration of the Independence of Cyberspace" stating that "We believe that from ethics, enlightened self-interest, and the commonweal, our governance will emerge.", https://www.eff.org/de/cyberspace-independence.

[470] Donahoe/Hughes/Kaye (2019): "Social Media Councils: From Concept to Reality." https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/gdpiart_19_smc_conference_report_wip_2019-05-12_final_1.pdf, p. 8.

[471] Gorwa (2019). The platform governance triangle: conceptualizing the informal regulation of online content. Internet Policy Review, 8(2). https://doi.org/10.14763/2019.2.1407.

the challenging diversity of regulatory objectives – the question arises: How can social media councils contribute to such improvements?

History and design of social media councils

Overview and definition

We use "social media councils" to mean external governance structures tasked either with *formulating and/or applying rules or determining the discoverability or visibility of content* on social networks in addition to or instead of the platforms; or tasked with *monitoring* the platform's activities relating thereto. This implies that the membership of "social media councils" can include civil society representatives[472] and/or experts with the aim of creating multi-stakeholder governance, although this is not strictly necessary for them to be referred to by that name. The name should therefore not pre-empt the complex question of whether such an institute effectively constrains the influence of a company or – going even further – legitimises the social accountability of the governance system.

Origins of the concept

Proposals for governance mechanisms that provide affected parties with an independent channel for complaints and/or involve civil society or user representatives in formulating the private rules of platform companies are nothing new. Quasi-judicial private institutions for reviewing company decisions started appearing in the German debate on regulation almost a decade ago, for instance with reference to "Cyber Courts"[473] or in the form of alternative dispute resolution mechanisms.[474] In the US, participatory approaches to formulating the private rules of companies have been proposed, such as a "Content Congress"[475] or external advisory bodies.[476]

[472] For an overview of the definition and potential of this approach to expression on the internet, see Strickling/Hill (2018): „Multi-stakeholder Governance Innovations to Protect Free Expression, Diversity and Civility Online", in: Donahoe/Hampson: "Governance Innovation for a Connected World. Protecting Free Expression, Diversity and Civic Engagement in the Global Digital Ecosystem" (pp. 45-52), https://www.cigionline.org/sites/default/files/documents/Stanford%20Special%20Report%20web.pdf.

[473] Ladeur/Gostomzyk: "Der Schutz von Persönlichkeitsrechten gegen Blogs", NJW 2012, 710 (pp. 713); Ladeur: "Neue Institutionen für den Daten- und Persönlichkeitsschutz im Internet: "Cyber-Courts" für die Blogosphere", DUD 2012, 711 (pp. 712); also see Vesting (2015): Die Medien des Rechts. Bd. 4: Computernetzwerke. Weilerswist: Velbrück Wissenschaft, p. 205.

[474] Spindler: "Persönlichkeitsschutz im Internet – Anforderungen und Grenzen einer Regulierung" Gutachten F on the occasion of the 69th German Jurists' Conference, 2012, p. 133; a similar proposal was made by Wagner: "Haftung von Plattformen für Rechtsverletzungen (Teil 2)", GRUR 2020, 447; such procedures appear to hold promise regarding justice in individual cases, but seem less suited to triggering general improvements in the governance systems of the platforms; see the comparison by Brown (2020): "Models of Governance for Online Hate Speech", Council of Europe https://rm.coe.int/models-of-governance-of-online-hate-speech/16809e671d (p. 84).

[475] Tomson, D., Morar, D. (2018). A Better Way to Regulate Social Media. Wall Street Journal, https://www.wsj.com/articles/a-better-way-to-regulate-social-media-1534707906.

[476] Ash, Timothy Garton, Robert Gorwa and Danaë Metaxa. 2019. Glasnost! Nine Ways Facebook Can Make Itself a Better Forum for Free Speech and Democracy. Reuters Institute for the Study of Journalism, https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-01/Garton_Ash_et_al_Facebook_report_FINAL_0.pdf, S. 19 – 20.

The current debate around social media councils was advanced by proposals from NGOs, such as Global Partners Digital[477] and ARTICLE 19[478]. ARTICLE 19's concept of "Social Media Councils" was mentioned in the 2018 annual report by the then UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye.[479] His recommendation for platform companies was that they

> "must open themselves up to public accountability. Effective and rights-respecting press councils worldwide provide a model for imposing minimum levels of consistency, transparency and accountability to commercial content moderation. (…) All [companies] that moderate content or act as gatekeepers should make the development of industry-wide accountability mechanisms (such as a social media council) a top priority."[480]

The debate around social media councils is closely related to demands that platforms should align their private rules, which are often international in scope, with international human rights standards. In this context, social media councils as potential external supervisory bodies would play a role by publicly criticising violations, thereby in a sense acting as an institutionalised "trigger" to create societal and political pressure on companies.

In addition, social media councils could also be used to verify that national agencies' commands and requests to platforms are in compliance with international human rights standards. In cases where such actions were in violation of applicable human rights standards, the councils could publicly back companies in rejecting the requests; but this is only being proposed in isolated instances.[481] The debate – and, therefore, this study – is primarily focused on ways of boosting the democratic legitimacy of the platforms' private orders.

Requirements

Social media councils as a concept do not have a very long history. Insofar as they are used to supervise the discretionary powers of platforms and act in areas in which the platforms are not restricted by applicable national legislation, national law does not provide much in the way of substantive or procedural criteria for them. Human rights are a more important source of law in this context, not least because of the implied

---

[477] With a proposal that concentrates on non-binding notes regarding the formulation of private rules by an "Independent Online Platform Standards Oversight Body", see Global Partners Digital (2018): "A Rights-Respecting Model of Online Content Regulation by Platforms", https://www.gp-digital.org/wp-content/uploads/2018/05/A-rights-respecting-model-of-online-content-regulation-by-platforms.pdf, pp. 26-28.

[478] Article 19: "Self-regulation and 'hate speech' on social media platforms" (2018), https://www.article19.org/wp-content/uploads/2018/03/Self-regulation-and-'hate-speech'-on-social-media-platforms_March2018.pdf, pp. 20-22.

[479] UN General Assembly (2018): "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression", UN A/HRC/38/35, https://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/38/35, paragraphs 58, 59, 63, 72.

[480] Ibid.

[481] The idea was favourably received by some companies in 2019, but was not pursued further, see Donahoe/Hughes/Kaye (2019): "Social Media Councils: From Concept to Reality." https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/gdpiart_19_smc_conference_report_wip_2019-05-12_final_1.pdf at the bottom of p. 13; the most powerful illustration of the need for robust, international mechanisms to protect human rights on digital platforms – even against infringements by states – is the Rohingya genocide in Myanmar, which was partly stoked by appeals on social media platforms, see Irving (2019): "Suppressing Atrocity Speech on Social Media", in: AJIL Unbound 113: 256-261, https://www.cambridge.org/core/journals/american-journal-of-international-law/article/suppressing-atrocity-speech-on-social-media/494334D2936A6A6E7C547C70816714D4; the most recent example of the potential of such an institution is provided by directives issued by India against Twitter, which likely contravened the guarantees by the Indian Constitution, Mahapatra/Fertmann/Kettemann (2021): Twitter's Modi Operandi: Lessons from India on Social Media's Challenges in Reconciling Terms of Service, National Law and Human Rights Law, Verfassungsblog, https://verfassungsblog.de/twitters-modi-operandi.

supervisory function for private platform law; this applies even more for quasi-judicial social media councils.

The United Nations Guiding Principles for Business and Human Rights (UNGP) represent the most important benchmark for quasi-judicial social media councils. Their "soft law" defines a corporate responsibility to respect human rights. Under Principles 29, 30 and 31, the UNGP also formulate guidelines for creating non-governmental, especially corporate or independent, complaints institutions and procedures. [482]

In this sense, social media councils (also) have to ensure that their decisions comply with international human rights norms. Regarding the institutional design and the procedural practices of social media councils, the UNGP require institutions to be suitable to justify the trust of those who use them, based on transparent and independent membership and transparent processes. Furthermore, they have to ensure that proceedings are fair and accessible to all who may potentially be affected. They have to warrant that affected parties dispose of all the information needed to present their cases, and their decision-making methods must be transparent.

In addition to the UNGP, the "Santa Clara Principles", an industry norm,[483] can also be used to develop minimum criteria for social media councils. The principles require that "companies should provide a meaningful opportunity for timely appeal of any content removal or account suspension." The accompanying list of minimum standards for such appeals mechanisms lists some of the key components of a due proecess, e.g., "Human review by a person or panel of persons that was not involved in the initial decision. An opportunity to present additional information that will be considered in the review. Notification of the results of the review, and a statement of the reasoning sufficient to allow the user to understand the decision. In the long term, independent external review processes may also be an important component. (….)"

Even in the absence of national regulations for social media councils, there are therefore already guidelines that can be used to assess and shape the design and decision-making practice of such institutions.

Design decisions

Areas of activity

As comprehensive lists of possible areas of activity of social media councils already exist,[484] we will direct our attention to some core questions regarding the design of such institutions here. A transatlantic working group on platform governance accurately described the many possible factors affecting the design of social media councils: "Policy makers and multi-stakeholder groups might consider a wide range of organizational structures and precedents to choose from, with format, purpose, jurisdiction, makeup, member selection,

[482] See also Council of Europe: Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries, section 11; appendix 2, especially 2.2.2. on enabling users to participate in formulating rules and 2.5 regarding access to effective complaint mechanisms.

[483] "The Santa Clara Principles: On Transparency and Accountability in Content Moderation" are an industry norm developed through collaboration by civil society and academia. They have been adopted by many companies on a voluntary basis; see https://santaclaraprinciples.org/; and Crocker et al: Who Has Your Back? Censorship Edition 2019, https://www.eff.org/de/wp/who-has-your-back-2019.

[484] See the functions listed by Tworek, (2019) Social Media Councils, pp. 99, https://www.cigionline.org/sites/default/files/documents/Platform-gov-WEB_VERSION.pdf#page=100.

standards, scope of work, and scalability to be determined in line with the underlying mission of the council."[485]

A council's underlying mission can be manifold. Obvious candidates include:

- preventive protection against unjustified measures against content; remediation after such measures have been imposed;
- systematic impulses to improve the governance systems of companies beyond individual cases;
- enabling an access to justice for as many affected parties as possible beyond automated and/or internal platform mechanisms;
- greater transparency;
- diversity-oriented supervision of content curation;
- securing the discoverability of certain content in the public interest;
- specific supervision regarding political campaign advertising and political communication; and
- supervision of basic design decisions and potential influences that guide users (so-called persuasive design and dark patterns).

Regulatory implementation: self-regulation or co-regulation

Social media councils can be implemented on the basis of voluntary cooperation between companies and experts and/or civil society (self-regulation), or alternatively in the form of models where social media councils are embedded in a framework defined by the state (co-regulation, regulated self-regulation[486]).

Statutory implementations of co-regulation are conceivable. But it is unclear to which extent lawmakers can, within a constitutional framework, prescribe procedures and institutions that lie outside the domain of government to platforms that make decisions and formulate rules which lawmakers cannot make and formulate precisely because the opinion-forming process lies outside the jurisdiction of government.

If social media councils are to improve company decision making in areas where lawmakers are unable or unwilling to formulate guidelines, it is apparent that social media councils will – at least initially[487] – be implemented by means of self-regulating initiatives. But even such self-regulation could be politically encouraged and collaboratively shaped, for instance when concrete, but non-binding proposals for an institution are formulated and political pressure is exerted on companies to participate in implementing them effectively ("quasi-regulation").[488] Examples include initiatives by expert NGOs such as ARTICLE 19, which exerted considerable influence on the process of developing the social media council concept and which is currently involved in introducing a national social media council in Ireland, as well as Ranking Digital Rights, which is already measuring the transparency of large technology companies by means an

[485] Transatlantic High Level Working Group (2020), Freedom and Accountability A Transatlantic Framework for Moderating Speech Online, https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/07/Freedom_and_Accountability_TWG_Final_Report.pdf, p. 26.

[486] An early contribution: Wolfgang Schulz / Thorsten Held: Regulierte Selbstregulierung als Form modernen Regierens. Commissioned by the Federal Commissioner for Cultural and Media Affairs of Germany. Final report. Hamburg: Publisher: Hans-Bredow-Institut, May 2002, p. 5, https://www.hans-bredow-institut.de/uploaWds/media/Publikationen/cms/media/a80e5e6dbc2427639ca0f437fe76d3c4c95634ac.pdf.

[487] A different view is put forward by Jarren/Gostomzyk (2020): Facebook's Hausgericht, https://www.medienpolitik.net/2020/04/facebooks-hausgericht, according to which co-regulation could already be considered at this stage.

[488] Tworek, Heidi (2019) Social Media Councils, (p. 100) https://www.cigionline.org/sites/default/files/documents/Platform-gov-WEB_VERSION.pdf#page=100

annual "Corporate Responsibility Index".[489] The decisions made by such a social media council, which had its origins in voluntary self-regulation, could later be taken into consideration by agencies and courts when interpreting existing obligations, thereby helping to solidify them.[490] It is also conceivable that incentives for participating in such an institution could be set at a later stage, for example by pointing out that the alternative is more stringent regulation.[491]

Mission: advisory, quasi-legislative, quasi-judicial

A further fundamental design decision is whether social media councils should be involved at the level of formulating rules and designing enforcement practices or whether they should only check individual judgments after the fact, in response to user complaints.

Involvement in rule-setting can in principle be designed as a quasi-legislative user parliament,[492] but is limited to an advisory role in norm-setting within many approaches.[493] Limiting a social media council to a purely advisory function risks restricting its influence. Conversely, binding rule-setting by a social media council creates the risk that a company could lose control over its platforms, which would likely disincentivise participation from a business perspective.

Apart from allowing involvement in developing rules and practices, creating opportunities for involvement in individual decisions regarding actions against user content is also conceivable. Taking into account the considerable volume of decisions that have to be made, a social media council would not be suitable as the first decision level for moderator decisions or even for initial appeals, but only as a later or higher-level review authority.[494]

Here, one needs to take into account that restitution (restoring the previous state) is only possible within certain limits because unjustified measures taken against content imply negative impacts in the form of missed communication opportunities which cannot be restored when such measures are lifted days, weeks or even months later (vice-versa, content removal represents also a very limited restitution when measures against the content have initially been erroneously rejected).[495]

---

[489] The initiative "Who targets me?" in support of transparent political advertising has also expressed an interest in participating in social media councils, see https://whotargets.me/en/oversight-boards-for-everything.

[490] For example, see the reference to the press codex, initially developed by the press in a self-regulating process, in interpreting statutory media obligations: Begr. zum Medienstaatsvertrag, LT-Drs. NRW 17/9052, 135; Lent, ZUM 2020, 593 (599); Heins/Lefeldt MMR 2021, 126.

[491] For certain journalistic/editorial online media, the new German State Media Treaty ("Medienstaatsvertrag") also takes this step in section 19, see Klausa: "Staatlicher Zahnersatz für den Presserat: Der Medienstaatsvertrag macht die Selbstregulierung der Presse zum Auslaufmodell", VerfBlog, 29 March 2021, https://verfassungsblog.de/staatlicher-zahnersatz-fur-den-presserat.

[492] Tomson, D., Morar, D. (2018). A Better Way to Regulate Social Media. Wall Street Journal, https://www.wsj.com/articles/a-better-way-to-regulate-social-media-1534707906.

[493] See the proposal by Bradley/Wingfield (2018): "A Rights-Respecting Model of Online Content Regulation by Platforms", Global Partners Digital, www.gp-digital.org/content-regulation-laws-threaten-our-freedom-of-expression-we-need-a-new-approach.

[494] For an overview f such tasks, see Brown (2020): "Models of Governance for Online Hate Speech", Council of Europe, https://rm.coe.int/models-of-governance-of-online-hate-speech/16809e671d (78-83).

[495] One of the first cases handled by the FOB (2020-004-IG-UA) serves as an illustrative example. At issue was the removal of content because of nudity; the content had been posted in October 2020 in connection with "Pink October", an international campaign to generate awareness for breast cancer. The FOB arrived at its decision four months later, in January 2021, and emphasised that the impossibility of restitution in the face of the expired campaign made it clear that its decisions needed to aim at transcending individual cases: https://oversightboard.com/decision/IG-7THR3SI1.

As the vast majority of such measures are by now automated, social media councils have an opportunity to exert influence by contributing to the design of such tools. Beyond that, it is precisely the design of platforms' feed and recommendation algorithms that represents a potential source of power which has to be supervised, implying that a complaints-based model cannot control platform measures which users *don't notice* (so-called shadowbanning). The potential benefits of a complaints-based social media council therefore are to be found primarily in potential systemic improvements which such a council could initiate based on individual cases.[496]

Existing social media council concepts, such as those proposed by ARTICLE 19[497] or the Stanford Global Digital Policy Incubator,[498] agree that a *combination* of a complaints-based institution (quasi-judiciary) and involvement in designing rules (quasi-legislative) is required. In any case, involvement in analysing content governance techniques which users are unaware of is critical.

Membership: councils of experts or citizens

Social media councils can be composed of experts on technology governance and freedom of expression, representatives of civil society groups or even randomly selected citizens.[499]

Such approaches to deliberative democracy in randomly selected small groups are discussed under the term "mini publics" and are not without controversy.[500] Keeping in mind current challenges to democracy, such as political polarisation and the spreading of disinformation, there is however something to be said for the development of "new forms of deliberative, collaborative and participative decision making that are evolving worldwide." [501]

On the other hand, formulating recommendations, defining binding rules or adjudicating complaints regarding expression on the net requires a certain level of expert knowledge, among other things to avoid unintended consequences. In this sense, models that combine representation with expert knowledge would seem advisable.

Geographic jurisdiction: national, regional or global

Social media councils can have national, regional or global jurisdiction. The Facebook Oversight Board operates at a global level. This offers certain benefits, but conceptualising councils at the national or regional level is also conceivable to help ensure that cultural and language contexts are appropriately reflected. Unified regional jurisdictions (such as Europe with EU law and the European Human Rights Convention) also suggest that unified social media councils are feasible, which could be designed in a way

---

[496] Brown (2020): "Models of Governance for Online Hate Speech", Council of Europe, https://rm.coe.int/models-of-governance-of-online-hate-speech/16809e671d, S. 133.

[497] Donahoe/Hughes/Kaye (2019): "Social Media Councils: From Concept to Reality." https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/gdpiart_19_smc_conference_report_wip_2019-05-12_final_1.pdf, S. 30-32.

[498] Ibid., pp. 26-29.

[499] For example, supported by MEP Geese (2021): Social Media Councils: Power to the people, https://alexandrageese.eu/der-dsa-teil-05-social-media-councils-power-to-the-people.

[500] Bächtiger, André, et al., (eds.). The Oxford Handbook of Deliberative Democracy. Oxford University Press, 2018, p. 1.

[501] OECD (2017): Recommendation of the Council on Open Government, available at: https://www.oecd.org/gov/Recommendation-Open-Government-Approved-Council-141217.pdf; the OECD also maintains a database of representative, deliberative institutions: https://airtable.com/shrRYPpTSs9NskHbv/tblfOHuQuKuOpPnHh.

that covers multiple platforms. Creating social media councils at the national level is also an option. They could function as alternative dispute resolution mechanisms for content moderation decisions made by platforms.[502]

Again, the design choices made here are not necessarily mutually exclusive, as national councils could be connected through a global association that defines best practices regarding the councils' work and principles.

"Material" jurisdiction: platform-specific or industry-wide

The jurisdiction of social media councils can be limited to a specific platform (platform-specific) or extended to cover many or all platforms or a specified type of platform (industry-wide). A specialised social media council would appear to be easier to implement by comparison, as only one company would need to support the concept if it is implemented through voluntary self-regulation. An industry-wide social media council would encounter greater challenges, not least in interacting with a multitude of different platforms and their diverse governance systems. A challenge for building such industry-wide mechanisms may also lie in applicable national anti-trust law that may restrict such forms of product (policy) related cooperation.

On the other hand, an industry-wide approach is especially promising because it could contribute to the independence of the institution. With industry-wide jurisdiction, such a council would not depend on its relationship with one or just a few companies for its existence and acceptance. As with other design dimensions, an iterative approach might be advisable, i.e., a social media council could be launched as an initiative by one or two companies and evolve over time into an industry-wide institution. It cannot be ruled out that other companies might join the platform-specific social media council of a competitor at a later stage. In the discussion to date, the predominant call has been for the establishment of industry-wide social media councils. [503]

Sources of inspiration of existing institutions of self-regulation

At this stage, the discussion around the design of social media councils is very much at the initial stages. Nonetheless, there are existing models of private institutions of self-regulation which are mentioned as examples or role models in the context of the social media council debate. They include press councils,[504] the US Financial Industry Regulatory Authority (FINRA),[505] the Canadian Broadcast Standards Council (CBSC),[506] the Internet Corporation for Assigned Names and Numbers (ICANN)[507] and various

[502] Donahoe/Hughes/Kaye (2019): "Social Media Councils: From Concept to Reality." https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/gdpiart_19_smc_conference_report_wip_2019-05-12_final_1.pdf, p. 30.

[503] Ibid., proposals by GDPi (from p. 26) and ARTICLE 19 (from p. 30); also in favour Kaye, UN General Assembly (2018): "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression", UN A/HRC/38/35, https://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/38/35 , section 72;

[504] UN General Assembly (2018): "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression", UN A/HRC/38/35, https://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/38/35, section 58.

[505] Transatlantic High Level Working Group (2020), "Freedom and Accountability: A Transatlantic Framework for Moderating Speech Online", https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/07/Freedom_and_Accountability_TWG_Final_Report.pdf, pp. 26-27.

[506] See the proposal for a national (Canadian) co-regulation "Council for Moderation Standards," based on the example of the CBSC, in Tenove, Tworek, McKelvey (2018): "Poisoning Democracy: How Canada Can Address Harmful Speech online", https://ppforum.ca/wp-content/uploads/2018/11/PoisoningDemocracy-PPF-1.pdf, pp. 27-28.

[507] Tenove/Tworek/McKelvey (2018): "Poisoning Democracy: How Canada Can Address Harmful Speech online", https://ppforum.ca/wp-content/uploads/2018/11/PoisoningDemocracy-PPF-1.pdf, pp. 27-28.

institutions of self-regulation established in other industries to ensure that practices comply with human rights, for instance in the resource extraction and manufacturing industries.[508]

From a German perspective, an approach to define the design and composition of a social media council could take its guidance from the broadcasting councils of the German public broadcasters. These councils include socially relevant groups, such as unions, employer associations, churches, environmental groups etc., which are considered "trustees of the interests of the general public"[509] and monitor compliance with statutory duties in this capacity.[510] There is no constitutional requirement for representatives to be affiliated with such associations. Instead, lawmakers can also require unaffiliated or weakly organised groups to be represented in some other way.[511] This appears increasingly advisable in light of the constitutionally required[512] consideration of the equality provisions of Article 3 of Germany's Basic Law in determining the composition of the councils. A promising route to pursue is to develop a model for social platform councils based on the extensive literature and constitutional jurisprudence on broadcasting councils. Such a model should at the same time reduce existing representation deficits, for instance through (partly) random selection of council members.[513]

Pursuing this route would also be helpful in balancing the need for the social media council's independence and the large number of checks required with the cooperative relationship it needs to have with the company. If the chosen approach is that of a national social media council designed with reference to broadcasting councils, this could also help to create the required high levels of acceptance. Such a national institution could also be integrated into an international network of social media platforms, as indicated above.

The Facebook Oversight Board

Evolution

In 2018, Mark Zuckerberg announced that he wished to create "an independent appeal process" that would function "almost like a Supreme Court".[514] Following Zuckerberg's announcement, a global consultation process was set in motion, involving stakeholders from civil society, academia and politics. It lasted for over a year and was intended to establish the jurisdiction of such an institution and its requirements in terms

[508] Gorwa, R. (2019). The platform governance triangle: conceptualising the informal regulation of online content. Internet Policy Review, 8(2). https://doi.org/10.14763/2019.2.1407.

[509] Federal Constitutional Court judgment of 25 March 2014, 1 BvF 1/11, available at https://www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/DE/2014/03/fs20140325_1bvf000111.html (margin no. 40).

[510] For a short overview see Schulz/Held/Dreyer/Wind (2008): Regulation of Broadcasting and Internet Services in Germany: a brief overview, available at https://doi.org/10.21241/ssoar.71697 , pp. 11-12.

[511] BVerfGE 83, 238 – 6. Rundfunkentscheidung (p. 335), available at https://www.servat.unibe.ch/Dfr/bv083238.html.

[512] See BVerfGE 83, 238 - 6. Rundfunkentscheidung (p. 336 et seq), available at https://www.servat.unibe.ch/Dfr/bv083238.html.

[513] This is also suggested for broadcasting bodies, see Dobusch (2019) for example: Zusammensetzung der Rundfunkgremien, https://www.deutschlandfunk.de/zusammensetzung-der-rundfunkgremien-schoeffen-fuer-mehr.2907.de.html?dram:article_id=448617.

[514] Klein (2018): "Mark Zuckerberg on Facebook's hardest year, and what comes next." Vox, 2 April, https://www.vox.com/2018/4/2/17185052/mark-zuckerberg-facebook-interview-fake-news-bots-cambridge ; Tworek, Social Media Councils, https://www.cigionline.org/articles/social-media-councils .

of staffing, organisational, legal, financial etc. resources.[515] The process of establishing the Facebook Oversight Board (FOB) was subject to close academic scrutiny, especially by evelyn douek[516] and Kate Klonick.[517]

The FOB is not an isolated initiative. It is embedded in the context of Facebook's ongoing slow-but-steady development of the private order system of its platforms towards greater transparency and justification of their private decision-making powers. Measures introduced include global transparency reports,[518] a formalised system for amending community standards,[519] informal methods to enable civil society participation in formulating the standards,[520] and commissioning external researchers to publicly assess the company's decision-making processes.[521] All of these initiatives are to be welcomed, but in the absence of robust supervisory and enforcement structures, they have not imposed any substantial limits on the company's power.

Operating principle

The FOB is a global, currently platform-specific, quasi-judicial and advisory social media council. Its members are representatives of civil society and academia[522] who act as an appeals body that reviews Facebook's decisions to delete user-generated content on its platforms Facebook and Instagram. Decisions are based on the respective community standards and take into account "international human rights standards."[523] For any given case, the FOB's decisions are binding on the company in terms of its voluntary commitment and are intended to be transferable to "identical content with parallel context".[524] Furthermore, the FOB acts as an advisory body that issues public, non-binding "recommendations" regarding Facebook's general rules and practices.[525] The company is required to respond publicly within

---

[515] See Facebook (2019): "Global Feedback & Input on the Facebook Oversight Board for Content Decisions", https://about.fb.com/wp-content/uploads/2019/06/oversight-board-consultation-report-2.pdf.

[516] douek, evelyn: "Facebook's 'Oversight Board:' Move Fast with Stable Infrastructure and Humility" (2019). 21 N.C. J. L. & Tech. 1 (2019), https://ssrn.com/abstract=3365358.

[517] Klonick (2020): "The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression" (30 June 2020). Yale Law Journal, Vol. 129, no. 2418, https://ssrn.com/abstract=3639234; dies. (2021): "Inside the Making Of Facebook's Supreme Court", New Yorker, https://www.newyorker.com/tech/annals-of-technology/inside-the-making-of-facebooks-supreme-court.

[518] Facebook: Transparency reports, https://transparency.facebook.com.

[519] Facebook: Community standards, https://www.facebook.com/communitystandards/, German version https://de-de.facebook.com/communitystandards.

[520] For greater detail on stakeholder engagement, see https://www.facebook.com/communitystandards/stakeholder_engagement; for information on how it unfolded in practice, see the observations provided in Kettemann/Schulz: "Setting Rules for 2.7 Billion", https://www.hans-bredow-institut.de/uploads/media/Publikationen/cms/media/5pz9hwo_AP_WiP001InsideFacebook.pdf, pp. 15; pp. 23.

[521] See Reports by the Data Transparency Advisory Group, most recently Bradford, Ben et al.: "Report of the Data Transparency Advisory Group April 2019", https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf.

[522] The initial 20 members announced on 6 May 2020 included academics, political (internet) activists, a former prime minister of Denmark and a former judge of the ECHR, see https://www.oversightboard.com/news/ announcing-the-first-members-of-the-oversight-board/; but some commentators have criticised the fact that US citizens and people with a legal background are overrepresented if one considers the global Facebook user base, see Article 19 (2020) The Facebook Oversight Board: A significant step for Facebook and a small step for freedom of expression, https://www.article19.org/resources/facebook-oversight-board-freedom-of-expression.

[523] Facebook: Community Standards, available at https://de-de.facebook.com/communitystandards/ (Introduction.

[524] Facebook (2019): Oversight Board Charter, Article 4, https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf.

[525] Facebook (2019): Oversight Board Charter, Article 3 paragraph 1, https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf.

specified time frames.[526] Facebook is funding the legally independent body through a $130m trust for the period 2020-2026.[527] The FOB Charter was completed and published in November 2019. It provides the framework for the establishment of the FOB and is complemented by a set of bylaws[528] and a rulebook[529] adopted by its members.

The FOB has pointedly been given a structure that allows other platform companies to join it, opening the way for it to become an industry-wide institution in time. But whether joining would be attractive for other companies is questionable, given the close ties between the FOB and Facebook.[530]

First decisions

By March 2021, within its first six months of operation, the FOB had published decisions on a dozen cases from a range of countries. That is not even a drop in the ocean; it is a water molecule, considering that Facebook makes around three million[531] decisions regarding content removal every single day.

A closer look at the FOB's decisions shows, however, that it is serious about defining its own position in Facebook's system of norms. In doing so, it refers more to international human rights standards than to the Facebook terms of use.[532] Even though there is no "hard" mechanism to force Facebook to implement the FOB's decisions, the public pressure generated by its published decisions does seem to have an effect: Facebook has responded in a cooperative way and implemented the recommendations, as far as one can tell. However, there are still significant problems with accessing the data needed to establish to which degree Facebook has really modified its practices.[533] The initial decisions indicate that many problems will remain, but that the FOB will gradually improve the platforms' governance systems.

---

[526] Facebook (2021): Oversight Board Bylaws, 2.3.2, https://www.oversightboard.com/sr/governance/bylaws (S. 25).

[527] Harris: "An Update on Building a Global Oversight Board", available at https://about.fb.com/news/2019/12/oversight-board-update.

[528] Facebook: Oversight Board Charter ,https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf; Oversight Board Bylaws, https://about.fb.com/wp-content/uploads/2020/01/Bylaws_v6.pdf.

[529] Oversight Board (2020): Rulebook for Case Review and Policy Guidance, https://oversightboard.com/sr/rulebook-for-case-review-and-policy-guidance.

[530] Lapowsky (2020): How Facebook's oversight board could rewrite the rules of the entire internet, https://www.protocol.com/facebook-oversight-board-rules-of-the-internet.

[531] Calculation based on global statistics provided by the company on measures taken against content (posts, comments etc.) for "substantive" violations of its community standards, see "Community Standards Enforcement Report" for Q3 2020. "Substantive" means that content belonging to the categories Fake Accounts (1.3 billion items of content) and Spam (1.9 billion items of content) was not counted; if those categories were included, the figure would reach 36 million pieces of content per day. The following categories were included: "Adult Nudity & Sexual Activity: 36,700,00; Bullying & Harassment: 3,500,000, Child Nudity & Sexual Exploitation: 12,400,000, Dangerous Organizations: Organized Hate, 4,000,000, Terrorism: 9,700,000, Hate Speech 22,100,000, Regulated Goods: Drugs, 4,730,000, Regulated Goods: Firearms 1,050,000, Suicide and Self-Injury 1,300,000, Violent & Graphic Content 19,200,000 – totalling 116.700.000 pieces of content for the period July-September 2020. The data can be downloaded from https://transparency.facebook.com/community-standards-enforcement.

[532] Gradoni, Lorenzo: Constitutional Review via Facebook's Oversight Board: How platform governance had its Marbury v Madison, VerfBlog, 2021/2/10, https://verfassungsblog.de/fob-marbury-v-madison.

[533] Douek (2021): The Oversight Board Moment You Should've Been Waiting For: Facebook Responds to the First Set of Decisions, Lawfare Blog, https://www.lawfareblog.com/oversight-board-moment-you-shouldve-been-waiting-facebook-responds-first-set-decisions.

Further examples of social media councils

Other company initiatives

In addition to the FOB, there are several platform-specific advisory bodies which the respective companies use to a greater or lesser extent to market the idea of legitimisation through participation, but which are not designed to be effective "social media councils," as defined here. The bodies in question represent a form of informal cooperation rather than an institution. For instance, the activities of these bodies and/or their influence on the company's practices are opaque, while their members are not protected against being recalled without good cause by the companies in question.

Such initiatives include the "Councils" established by TikTok for the EU,[534] the Asia-Pacific region[535] and the US,[536] with a membership constituted in each case by individuals including (former) politicians and/or civil society representatives. As far as can be established with any degree of certainty, these institutions do not communicate independently with the public, do not report on their activities and are not independent, neither legally nor by means of a voluntary commitment by the company.

The "Trust and Safety Council"[537] operated by Twitter has a different kind of membership, where the members are NGOs rather than individuals. But this "Council" also serves as little more than a company forum for obtaining non-binding opinions on the company's actions. Because of the absence of an independent organisational structure or public image, it is more akin to an informal cooperation.

Beyond that, there is a range of industry-wide cooperation mechanisms focusing on appropriate ways of dealing with content. The Global Network Initiative is a well-known initiative in this category. It has brought together companies, NGOs and research institutes to develop codes of conduct in support of corporate practices that respect human rights and to which companies which are members commit themselves.[538] Such initiatives are criticised for not going far enough and not holding companies to their voluntary commitments sufficiently,[539] but they could represent a promising launchpad to start building social media councils at the international level.

Planned "Social Media Council" in Ireland

Regarding projects which have already been launched, the most promising initiative apart from the FOB is the establishment of a national social media council with broad jurisdiction in Ireland. ARTICLE 19, a non-governmental organisation, has proposed the creation of an Irish "Social Media Council" and promoted

---

[534] "European Safety Advisory Council", see Tiktok (2021): Meet TikTok's European Safety Advisory Council, https://newsroom.tiktok.com/en-be/meet-tiktoks-european-safety-advisory-council.

[535] "TikTok Asia Pacific Safety Advisory Council", see TikTok (2020): Introducing the TikTok Asia Pacific Safety Advisory Council, https://newsroom.tiktok.com/en-sg/tiktok-apac-safety-advisory-council.

[536] "TikTok Content Advisory Council", see TikTok (2020): Introducing the TikTok Content Advisory Council, https://newsroom.tiktok.com/en-us/introducing-the-tiktok-content-advisory-council.

[537] Twitter (2021): https://about.twitter.com/en/our-priorities/healthy-conversations/trust-and-safety-council.

[538] See https://globalnetworkinitiative.org/gni-principles, members include Facebook and Google.

[539] See Labowitz/Meyer (2016), for example: Why We're Leaving the Global Network Initiative, https://bhr.stern.nyu.edu/blogs/why-were-leaving-the-gni.

the idea that future Irish legislation[540] should support this endeavour. The NGO also offered to manage the institution in the context of a pilot project. At this stage it is not yet clear if this will in fact happen; however, one of the organisers expressed cautious optimism at a presentation in November 2020.[541]

"Institutions of regulated self-regulation" under the German Network Enforcement Act (Netzwerkdurchsetzungsgesetz, NetzDG)

In Germany, the recognised institutions of regulated self-regulation provided for under Article 3.2.3b) of the Network Enforcement Act provide a legal framework for quasi-judiciary institutions which make decisions about removing certain types of content covered by the Act. Social networks which are covered by the Act because of their size can opt to join these institutions to submit content for legal review. Currently, only one such institution is active: the NetzDG review panel of the German Association for Voluntary Self-Regulation of Digital Media Service Providers (*Freiwillige Selbstkontrolle Multimedia-Diensteanbieter*, FSM).[542]

The conceptualisation of these institutions falls short of the potential of social media councils. This is not so much owing to the way it has been implemented in practice by the FSM (the only example to date), but is a consequence of the NetzDG regulatory context. The NetzDG application framework limits the jurisdiction of institutions to the tightly circumscribed area of the (omitted) *deletion* of supposedly *unlawful* content. It does not address questions of content aggregation or restrictions on the visibility of content, nor does it cover the much broader field of removing content that only violates the terms of use of the platform providers.

Within this narrowly limited scope of the law, the NetzDG also fails to leverage these institutions to encourage platforms to implement systemic improvements of their internal procedures and decision-making structures. The institutions do not develop any form of "case law" from their regular reviews of individual cases and do not assess the process leading up to the submission of content to them. Within the NetzDG, any systemic improvements are to be triggered through a traditional approach of imposing fines for regulatory non-compliance (so-called "command and control regulation").

Even in terms of the extremely limited scope of the NetzDG, this regulatory approach is already highly controversial from a constitutional law perspective. Broadening it to other areas, such as measures against "legal, but harmful" content within the (at least *de facto)* existing discretion of the platforms would likely overstep these boundaries. This traces back to constitutional requirements relating to the independence of the process of opinion formation from the state and the very limited legislative power of the *Bund* (opposed to the *Länder*) in this context.[543] The institutions provided for under the Network Enforcement Act therefore do not offer a suitable framework for the development of social media councils.

---

[540] Online Safety and Media Regulation Bill, regarding the current legislative process see https://www.gov.ie/en/publication/d8e4c-online-safety-and-media-regulation-bill.

[541] See the recording of a lecture by Pierre François Docquir (ARTICLE 19), https://www.youtube.com/watch?v=zYYw7bBg88o&t=1530s, from time stamp 25:30.

[542] Freiwillige Selbstkontrolle Multimedia-Diensteanbieter: NetzDG-Prüfausschuss, https://www.fsm.de/de/netzdg.

[543] The legislative competence of the federal government regarding the Network Enforcement Act in its current form is mostly rejected in the literature, see Liesching in: Spindler/Schmitz/Liesching, 2. ed. 2018, NetzDG § 1 margin no. 10-12 and also Hoven/Gersdorf in: BeckOK InfoMedienR/ 31. Ed. 1 May 2019, NetzDG § 1 margin no. 5-8 with numerous further citations; in support, see Schwartmann (2020), Stellungnahme im Rahmen des NetzDG-Änderungsgesetzes (BT-Drucksache 19/18792), available at https://www.bundestag.de/resource/blob/700958/c2132ca5cbf50c600d04a0df0058c1b8/schwartmann-data.pdf; as well as, by the same author:

Summary and conclusion: the potential of social media councils

The concept of social media councils and especially the Facebook Oversight Board (FOB) is not without controversy. In response to the establishment of the FOB, for example, a "Real Facebook Oversight Board" was created, an initiative by critics of Facebook who call for government regulation while criticising the lacking effectiveness and slow establishment of the FOB, which – contrary to earlier announcements – only commenced work after the US presidential elections.[544]

With specific reference to the FOB, its structure[545] is seen to be insufficiently independent while its mandate[546] is considered to be too limited, among other things because practices not related to deletions can hardly be monitored and assessed.[547] A further criticism is that the FOB does not have the structural capacity to improve the situation because it can only review a very small number of individual cases.[548]

Some commentators also express a concern that quasi-judicial social media councils could erode the effectiveness of government legal protections,[549] impede the international standardisation of guarantees such as freedom of expression[550] or hamper the ability to use national courts in other ways.[551] These concerns cannot be dismissed entirely if quasi-judicial social media councils were in fact to be established and if they made a great number of decisions. However, considering that the approach to designing such social media councils – presumably the only feasible way – is to focus on reviewing a very small number[552] of "leading cases" to improve general systems, this does not appear to be a particularly significant risk.

A further concern regarding social media councils is that they effectively stabilise private orders without giving rise to real changes.[553] A counterargument is that (even minimal) iterative improvements through stronger social accountability for decisions, including those made by private companies, can only be a good thing and that – arguing from the point of view of the separation of powers – even a slightly broader

---

(2017) Stellungnahme zum NetzDG-Entwurf (BT-Drucksache 18/12356), available at https://www.bundestag.de/resource/blob/510886/002a8ce4b15005b96318abacee89199d/schwartmann-data.pdf.

[544] Butcher (2020): 'The Real Facebook Oversight Board' launches to counter Facebook's 'Oversight Board', https://techcrunch.com/2020/09/30/the-real-facebook-oversight-board-launches-to-counter-facebooks-oversight-board/; https://the-citizens.com/real-facebook-oversight.

[545] Morar (2019): Facebook's Oversight Board: A toothless Supreme Court?, https://www.internetgovernance.org/2019/10/02/facebooks-oversight-board-a-judiciary-with-no-constitution/

[546] Reed (2019): Facebook's Oversight Board needs a broader mandate that integrates human rights principles, https://rankingdigitalrights.org/2019/05/22/facebooks-oversight-board-needs-broader-mandate-that-integrates-human-rights-principles.

[547] Weinzierl (2019), Difficult Times Ahead for the Facebook "Supreme Court", VerfBlog, 2019/9/21

https://verfassungsblog.de/difficult-times-ahead-for-the-facebook-supreme-cour.

[548] McNamee, Roger; Ressa, Maria: Facebook's "Oversight Board" Is a Sham. The Answer to the Capitol Riot Is Regulating Social Media (2021), https://time.com/5933989/facebook-oversight-regulating-social-media/; with similar comment regarding its limited capacity see Ghosh, Dipayan; Hendrix, Justin: Facebook's Oversight Board Just Announced Its First Cases, but it Already Needs an Overhaul, VerfBlog, 2020/12/19, https://verfassungsblog.de/fob-first-cases.

[549] Jarren/Gostomzyk (2020): Facebooks Hausgericht, https://www.medienpolitik.net/2020/04/facebooks-hausgericht.

[550] Wagner, Haftung von Plattformen für Rechtsverletzungen (Teil 2), GRUR 2020, 329 (332).

[551] Weinzierl (2019).

[552] For every single FOB decision, there are millions of measures taken by Facebook (see above) that are not subject to review; considering the very low probability of the FOB's accepting a given case for review, it is unlikely that national courts would regard this as an alternative remedy in civil cases, for example.

[553] McSherry (2019) "Social Media Councils: A Better Way Forward, Window Dressing, or Global Speech Police?", https://www.eff.org/de/deeplinks/2019/05/social-media-councils-better-way-firward-lipstick-pig-or-global-speech-police.

distribution of the decision-making power of platforms should not be rejected as a matter of principle ("more could be done"). In this sense, social media councils could resemble institutional role models like press and broadcasting councils, which have existed in Germany for decades. As compromise solutions, they also face inevitable criticism, but have never been replaced due to the absence constitutionally, politically and/or practically viable alternatives.[554]

A common perspective on social media councils rooted in legal theory is that such councils form part of the emergent phenomenon of private "constitutions".[555] This viewpoint illustrates the considerable potential of the internal, functional differentiation of companies towards democratisation – or, at least, a partial implementation of rule of law-principles by the platforms and the creation of internal checks and balances.[556] In this way, social media councils can contribute to generating internal transparency and discussions about rules.[557]

On the precision scales that measure the balance of power in the field of digital expression, the relationship between platforms and states is far from being in balance. This may be lamentable, but it offers the advantage that it makes even bold institutional experiments worthwhile. In this sense, social media councils cannot replace private or government regulation. They can only partly relieve the regulatory pressure resulting from the manifold challenges of online communication; but they enable interesting models for the necessary redemocratisation of the normative orders of the hybrid[558] communication spaces of the public sphere in the present era.

For the further debate around social media councils, it is important to note that the attempt by Facebook remains only *one* version of such an institution. The "Oversight Board" should not be elevated to an archetype of a social media council, and the concept of social media councils should not be monopolised by Facebook.[559] Still, the Oversight Board provides a useful opportunity to discuss many of the core challenges and design decisions.

---

[554] Most recently, for some useful insights on the criticisms levelled against the German Press Council in the context of the new State Media Treaty, see Klausa: "Staatlicher Zahnersatz für den Presserat: Der Medienstaatsvertrag macht die Selbstregulierung der Presse zum Auslaufmodell", VerfBlog, 2021/3/29, https://verfassungsblog.de/staatlicher-zahnersatz-fur-den-presserat/, DOI: 10.17176/20210329-195047-0; for general observations on the criticism of press councils, see Pöttker, in: Baum/Langenbucher/Pöttker/Schicha: "Handbuch Medienselbstkontrolle", https://link.springer.com/chapter/10.1007/978-3-322-80808-0_12, pp. 125-131; similarly, regarding broadcasting councils, see Hahn: "Der Rundfunkrat — ein verzichtbares Kontrollinstrument?" in the same volume, https://link.springer.com/chapter/10.1007/978-3-322-80808-0_18, pp. 159-174.

[555] As a representative example of many similar contributions, see douek: "Facebook's "Oversight Board": Move Fast with Stable Infrastructure and Humility", North Carolina Journal of Law & Technology Volume 21, Issue 1 (2019), p. 4, which describes the social media council of Facebook as "one of the most ambitious constitutional projects of the modern era"; Pozen: "Authoritarian Constitutionalism in Facebookland" (2018), who argued (before the establishment of the Oversight Board) that Facebook was at best an authoritarian constitutional state, https://balkin.blogspot.com/2018/10/authoritarian-constitutionalism-in.html.

[556] Maroni, Marta: "Some reflections on the announced Facebook Oversight Board", available at https://cmpf.eui.eu/some-reflections-on-the-announced-facebook-oversight-board.

[557] Jarren/Gostomzyk (2020): Facebook's Hausgericht, https://www.medienpolitik.net/2020/04/facebooks-hausgericht.

[558] For an early example, see Ladeur, Neue Institutionen für den Daten- und Persönlichkeitsschutz im Internet: "Cyber-Courts" für die Blogosphere, DUD 2012, 711 (pp. 713): "It is conceivable that one could respond to the hybrid [authors' note: private-public] nature of the new media (…) by outlining a model of a legal order that is also hybrid in nature. It could start with an attempt to encourage the self-organisation of social rules which form the infrastructure of traditional media law (…) through national law, especially through jurisprudence as a kind of 'regulator' of the new media, but without replacing such self-organisation."

[559] In this context, we should also avoid using "oversight boards" as a generic term for social media councils, which is already happening in some instances. For example, see Who Targets Me, Oversight boards for everything (2021), https://whotargets.me/en/oversight-boards-for-everything.

An open marketplace of ideas requires competition. Now, it is becoming increasingly clear, that competition is also needed among different concepts to institutionally protect this market on the internet.

Further discussions on this topic could be based on the following considerations:

- Despite current uncertainty around their exact design, social media councils represent a good opportunity to increase the legitimacy of the normative orders of platforms, strengthen the protection of individual rights, and promote social cohesion.
- At the same time, it must be recognised that social media councils are no silver bullet. As independent mechanisms, they can complement existing models of private and state regulation of social networks and initiate improvements, but they cannot replace such models. In this context, the main emphasis is on improving company rules and the corresponding enforcement approaches.
- If not properly configured, social media councils may conceal actual power structures and fail to initiate real change. Therefore, such councils have to meet the highest standards of transparency regarding their own operations, while being equipped with appropriate rights to information and data access to ensure that diverse stakeholders can verify which systemic improvements they initiated, if any.
- In the absence of the national regulation of social media councils, there are already guidelines in the field of international human rights standards and industry agreements which would allow the configuration and decision-making practice of social media councils to be designed and measured.
- Politics, civil society, companies and multi-stakeholder groups provide a broad palette of configuration options for social media councils, depending on their underlying goals.
- A promising objective for social media councils is to initiate systemic improvements in the governance systems of companies beyond just individual cases. On the other hand, social media councils do not appear suited to contribute to remedy violations that have already occurred or to provide a type of legal hearing for as many affected parties as possible.
- The most important configuration dimensions of social media councils include their jurisdiction in various respects (industry-wide or platform-specific; national, regional or global; quasi-legislative or quasi-judicial design), their composition and the selection of members, their decision-making tools and the question of how to implement them, either through self-regulation or co-regulation.
- With specific reference to the idea of pursuing a national social media council for Germany, a promising approach would be to develop a model for social platform councils based on the extensive literature and constitutional jurisprudence on broadcasting councils. Such a basis in constitutional law could also contribute to creating the greater acceptance for social media councils needed in the absence of "hard" enforcement mechanisms.

The debate around the potential of social media councils to reimport democratic values into the private orders of public communication has only just begun. Private orders can (and should be) oriented towards public values via laws or legal judgments.[560]

[560] Kettemann/Tiedeke, Back up: can users sue platforms to reinstate deleted content? Internet Policy Review 9 (2020) 2, https://policyreview.info/articles/analysis/back-can-users-sue-platforms-reinstate-deleted-content, DOI: 10.14763/2020.2.1484

However, in light of the increase of multiple public spheres ("many publics")[561] and the growing recognition of the inner complexity or multi-faceted nature of the platforms themselves,[562] the time has come to reconceptualise the democratisation of platforms. In this context, social media councils could be the starting signal in the necessary race to establish a fairer configuration of the normative order of the digital world.[563]

[561] Kettemann/Tiedeke, Online order in the age of many publics, Kybernetes 13 (2020), https://www.emerald.com/insight/content/doi/10.1108/K-07-2020-0423/full/html?skipTracking=true

[562] Arun (2021): Facebook's Faces, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3805210.

[563] Kettemann, Deontology of the Digital: The Normative Order of the Internet, in Kettemann (ed.), Navigating Normative Orders. Interdisciplinary Perspectives (Frankfurt/New York: Campus, 2020), 76-91.

Internet and Self-Regulation: Media Councils as Models for Social Media Councils?

Riku Neuvonen
UNIVERSITY OF HELSINKI AND TAMPERE UNIVERSITY

Introduction

The big tech giants have faced increased public scrutiny and public hearings during the last years. The reason for this scrutiny is their transformative impacts on businesses, social connections and media use in everyday life. Social media platforms have been especially criticised for not taking seriously their position as crucial nodes of public life. Privately owned platforms' growing role and wide range of social functions have created regulatory problems. The business goals of such platforms have been colliding with common values underlying the public good. However, responses to these global problems have either been regional or national in scope.

Social Media Councils (SMCs) are one of the proposals for solving the problems caused by social media. The definition of SMCs is quite vague, however. In this paper, SMCs are defined the same way as in the proposal made by ARTICLE 19. SMCs are compared to contemporary media/press councils, which are part of the self-regulation of heritage media. The media (press) councils are analysed at a general level and focus is placed on how media councils in three selected countries (the UK, Sweden and Finland) differ at a more historical and detailed level.

First, I describe briefly why a growing need exists for solutions like SMCs to (self-)regulate platforms. Second, I briefly introduce the proposal for SMCs made by ARTICLE 19. Third, I provide a summary of media councils and focus in more detail on three selected countries to highlight how various media council models differ and to compare such media councils to SMC proposals.

Why Are SMCs Needed?

Problems with Platform Moderation

The concept of moderation is old. Its former definition stressed the need to eliminate or at least reduce the extremes between ancient philosophy and early Christianity. In the context of the internet, moderation means removing unsuitable content that violates rules. In the internet's early days, a bulletin board System (BBS) and discussion boards were moderated by administrators and trusted users. Gradually, media companies established discussion forums, and moderation became more professional in some cases. A platform's business model involves gathering information, which requires keeping users engaged with the platform as long as possible. Therefore, platforms have an incentive to curate content to lure and keep users, whereas harsh moderation measures would drive away such users. From the viewpoint of public good, moderation is needed to protect human rights and minimise harms.

In the beginning of the 1990s, most major internet companies were US-based companies. Kate Klonick (2018) has thus argued that the rules for these types of proto-platforms and moderators were strongly influenced by and sometimes even actively guided by US lawyers specialising in free speech. The First Amendment of the US constitution is more liberal in its guarantees and individualistic in nature than equivalent European doctrines or other free speech theories. Hence, the ethos and practices of many internet companies are heavily influenced by strict US interpretations of free speech and the liberal, individual vision of a free internet (Suzor, 2019).

Moderation practices vary. The first approach is manual content moderation. Tarleton Gillespie (2018) distinguishes between the work of internal teams, crowd workers, community managers, flaggers, super flaggers, peer support, external efforts and everyone else. Moderation is one aspect of platform administration, wherein the platform administrator (owner) can select users and edit content. Currently, large platforms are professionally moderated, with the work often outsourced and delegated to algorithms. Additionally, platforms can allow users to flag content and inform administrators and moderators.

Automated or algorithmic content moderation entails the use of automated detection, filtering and moderation tools to flag, separate and remove content. Hybrid content moderation is another approach, which incorporates elements of both manual and automated approaches (Singh, 2019, p. 7). Typically, this approach involves using automated tools to flag and prioritise specific content cases for human reviewers, who then make final judgements on a case-by-case basis. Both smaller and larger platforms are increasingly adopting hybrid content moderation because it helps reduce the initial workload for human reviewers.

Content can be moderated before or after publication. Ex ante content moderation means that content is screened before it is published. Ex post reactive content moderation takes place after a post has been published on a platform and subsequently flagged or reported for review by a user or entity, such as an internet referral unit or trusted flagger (Singh, 2019, pp. 7–8).

Despite recent pressures, the present state of moderation on social media platforms is largely inadequate. The billions of messages spreading globally create a huge challenge for reviewers. Automated, algorithmic tools have created problems because of their inability to detect and identify illegal or harmful content and/or expressions based on the specific context. The pressure to remove illegal or harmful content has compelled platforms to remove even slightly suspicious content that might harm rule-abiding users' freedom of speech. Users also constantly criticise the fact that appeal processes are slow and haphazard, if they are available at all.

On the other hand, some harmful or even illegal content has been maintained on platforms. A recent report by the Soufan Center (2021) elucidated the present state of content moderation problems. Though social media platforms promised to adopt measures that prevent false information during the 2020 US presidential election, the report claims that the Russia-driven amplification of QAnon-related narratives predominated in 2020 and that China continued to promote QAnon conspiracies toward the end of 2020, outpacing Russia-related amplifications in 2021.

Moreover, moderation is only part of a platform's content management operations – curation as plays a role (Klonick, 2018; Gillespie, 2018). Curated content helps keep users engaged on platforms. Platforms are, thus, incentivised to keep as much content available as possible. Especially controversial content, such as QAnon-related narratives, create debates and keep users engaged. Essentially, the incentive for platforms to manage content differs radically from that of users or the requirements of the public good. Therefore, moderation must be monitored.

The Problems with Regulation

In the beginning, the internet was considered a vehicle for new technological inventions and a place for freedom of speech. Developments can be divided into two periods. First, regulators as well as existing literature focused on the liability of intermediaries. Second, human rights were introduced in the form of different principles and declarations. Contemporary research increasingly focuses on how to protect human rights on the internet. Meanwhile, NGOs and other groups are presenting models for ensuring digital rights, and digital constitutionalism is a frequently used catchphrase when introducing a more rights-based approach to regulations (Pollicino, 2021). Altogether, a shift has occurred from promoting freedom to creating regulatory models based on a notion of rights that originate from basic human rights.

The ethos of freedom in the early days of the internet has also affected regulations. In the United States, Section 230 of the Information Technology Act of 2000 guarantees a liability exemption to intermediaries for third-party content if the intermediary has acted in good faith. In the European Union, Article 14 of the E-Commerce Directive of 2000 (ECD) states that intermediaries, or digital or online platforms, are not legally responsible for hosting illegal content – although they are required to remove such material once it has been flagged. This obligation only applies to certain content, though. Article 15 of the ECD prohibits the general monitoring of content, but it allows the monitoring of specific content as well as voluntary monitoring by platforms. Only courts can order intermediaries to remove illegal content; however, the circumstances under which social media platforms are considered intermediaries are unclear.

Similarly, social media is also on the fringes of media regulation. The word 'media' is a questionable component of the social media concept. From the perspective of traditional media regulation, social media is not media because social media companies have no editorial responsibility. Phil M. Napoli (2019) has argued that in some situations, social media companies have even tried to take advantage of media privileges without committing to any obligations. Social media platforms are too active to count as intermediaries, but they do not operate like traditional media.

The growing significance of platforms has increased the pressure to address unclear regulations. Solutions in Europe have been national and regional in scope. One of the first national laws regarding such platforms was Sweden's Electronic Bulletin Boards Responsibility Act of 1998, which applies to platforms other than internet bulletin boards (BBs). A more recent example of such national laws is Germany's Network Enforcement Act (Netzwerkdurchsetzungsgesetz) of 2018, which targets especially large social media platforms with German users. Both national laws reference criminal law and oblige platforms to moderate and remove content.

The European Commission has provided the impetus for the voluntary removal of content through, for example, the 2016 Code of Conduct on Hate Speech, the 2017 Communication on Tackling Illegal Content and the 2018 Recommendation on Measures to Effectively Tackle Illegal Content Online. In 2019, the European Parliament also adopted a report pushing for content monitoring to be outsourced to hosting services under the pretext of the fight against terrorism. EU initiatives have included a desire for legal safeguards that do not require practical measures. In 2018, the renewed Audiovisual Media Service Directive (AVMSD) began requiring video-sharing platforms to take appropriate measures to protect minors and the public from specific harmful content. These new obligations have been effectively executed by monitoring and removing content. As a directive, the AVMSD is more binding than previous recommendations, representing a change from the previous soft regulation. The Digital Service Act and the Digital Market

Act are the next steps in the move from soft regulation via communication or codes to harder regulation via more binding measures.

At the global level, human rights treaties by the UN and regional entities guarantee freedom of speech and privacy, which are the most crucial human rights for platform users. The UN human right treaties have steered development of a regulatory framework for the internet. The special advisors in particular have given statements on how the internet should be governed. Human rights are not only global but also regional. The European Convention on Human Rights (ECHR, drafted in 1950) is one of oldest regional treaties, and the European Court of Human Rights (ECtHR) interprets its provisions in various cases. All members of the European Council are committed to following the rules of the ECHR and practice of the ECtHR. Other regional human rights organisations and treaties also exist. The Organization of American States has established the Inter-American Court of Human Rights to interpret the provisions of the American Convention on Human Rights (ACHR). The role of the Inter-American court is more adjudicatory and advisory. The US and Canada are not members of the ACHR or the Inter-American court. The third regional human rights body to be mentioned is the African Court on Human and Peoples' Rights, which complements and reinforces the functions of the African Commission on Human and Peoples' Rights. Additionally, the EU Charter of Fundamental Rights of the European Union and the European Court of Justice (ECJ) have played a role in setting human rights standards in Europe, with both the EU and ECJ being quite active in cases related to the internet and platforms.

Another issue concerns the concurrence (or competing/conflicting role) of freedom of speech and other communication-related rights (freedom of art, freedom of assembly, freedom of science and access to information). The traditional conflict has been between freedom of speech and the right to privacy, but the current situation is more complex. Various treaties, courts and entities are creating multipolar fundamental/human rights situations. This problem is more severe in a digital environment because different actors regulate different factors, such as global self-regulation (e.g. ICANN), the domestic regulation of companies, the regulation of supranational competition, administrative regulation and the limitations laid down in law, both in a company's home country and in its country of operation. Human rights, fundamental rights and different principles are all sources of tension that steer regulatory choices in different directions.

However, these regional human rights treaties have little effect on the global internet. Human rights should be protected on the internet, but the UN represents soft law and more effective human rights watchdogs are in scope regionally. All global regulatory bodies, such as the ICANN, are focused on infrastructure, and expanding their competence to regulate content is politically impossible. Therefore, forms of regulation other than traditional state-based or organisation-based regulation are needed.

Social Media Councils in Nutshell

Altogether, there is a need for a global mechanism to make platform moderation more transparent and more accountable. There is not much academic discussion or other sources on social media councils. Therefore, this paper focuses on the ARTICLE 19 (from 12.10.2021) proposal as a model for such transparency and accountability. The initial proposal was made in 2018, and it has subsequently been updated based on consultations and different situations. Soon after, the idea for SMCs was endorsed by then UN Special Rapporteur on freedom of expression David Kaye. The current plan is to establish a pilot SMC in Ireland in the near future.

According to the proposal, the key objectives for the SMCs include reviewing individual content moderation decisions made by social media platforms on the basis of international standards on freedom of expression and other fundamental rights, providing general guidance on content moderation practices to ensure they follow international standards on freedom of expression and other fundamental rights, acting as a forum where stakeholders can discuss recommendations and using a voluntary-compliance approach to the oversight of content moderation that does not create legal obligations.

The model is therefore self-regulatory and contains very little legal precedent or other hard normative power. While SMCs' sources of power vary, they are based mostly on international standards on freedom of expression and other fundamental rights. It is interesting whether that will entail interpreting UN treaties or some regional human rights instruments or even fundamental rights guaranteed in national constitutions (or, e.g. the EU Charter) as the basis for SMC decisions. References to international standard are problematic because of multipolar fundamental rights situations and the co-existence of different rights or normative documents. The lack of consensus on human rights is one of the key reasons why the SMCs are needed, but at the moment the sources buttressing the decision-making power of SMCs are still quite fragmented.

The actions of SMCs are based on the Charter of Ethics. They gather information from social media platforms; the platforms are committed to following the decisions made by the SMCs. The main difficulty with SMCs is funding and the commitments required from social media companies. Interestingly, the Article19 proposal promotes freedom of speech more than other rights.

One section of the proposal describes the finds and prior experiences of SMCs. On this basis, it suggests that SMCs should be independent of government, commercial and special interests. The SMCs should be established via a fully inclusive and transparent process that includes broad and inclusive representation and they must be transparent. While these are all quite commendable goals, SMCs have existed for some time and were established in various ways, as we see in the next section. Based on workshops conducted in 2019, experts proposed creating SMCs at the national level and that they should resemble media councils.

The Oversight Board of Facebook

The Oversight Board (OB) is an articulation of Facebook's intentions to meet its claims of accountability through self-regulation. It was launched with a relatively high level of media attention in 2020, and it includes several well-known politicians, journalists and experts in international law, such as former Prime Minister of Denmark Helle Thorning-Schmidt, former *Guardian* Editor-in-Chief Alan Rusbridger, PEN America Chief Executive Officer Suzanne Nossel and Nobel Peace Prize Laureate Tawakkol Karman.

The rules of Facebook – 'Lex Facebook', as Lee A. Bygrave (2015) has termed them – form a hierarchically structured normative system. Their principles comprise Facebook Values at the top tier, Community Standards at the second tier and Internal Implementation Standards and AI Protocols as the interpretation of these abstract rules. The Standards and Protocols are specific, non-public instructions for moderators and protocols of algorithms. Facebook Values and Community Standards are public documents.

The Oversight Board's core function is to review content enforcement decisions, determine whether they were consistent with Facebook's content policies and values, and interpret decisions vis-à-vis Facebook's articulated values (OB Charter, 1.4.2). Therefore, all decisions must be based on Lex Facebook. The board has the option of offering advice on how Lex Facebook should be developed, but the company has no obligation to follow this advice. Lex Facebook constitutes the core structure of the norms upon which the

board bases its decisions. After its establishment, the board announced that it also considered international human rights norms and standards (Gradoni, 2021). The Rulebook for Case Review and Policy Guidance is a framework for the board, and it reportedly is in alignment with the UN Guiding Principles on Business and Human Rights (UNGP). The UNGP guidelines address how companies and states should prevent and remedy human rights abuses in a business context. It comprises three pillars and 31 principles. The UNGP pillars include the stated duty to protect human rights, corporate responsibility to respect human rights and individual access to a remedy if human rights are not respected or protected. However, the UNGP is a soft law by nature, lacking an enforcement mechanism.

Lex Facebook is an idiosyncratic system of standards, some of which are not public. One criterion for appeal to the OB is when content is not removed on the basis of its illegality. The OB decides solely on the basis of Lex Facebook and the board's bylaws. In its first decisions, the OB referred to the ICCPR and the UNGP, but what is the added value of these references? The OB does not interpret laws or human rights treaties; rather, it picks one guideline (UNGP) and one treaty (ICCPR) to give its decisions juridical camouflage. Its decisions thus far represent a very liberal and US-based interpretation of freedom of speech, with a hint of the European preference for proportionality (Pollicino et al., 2021). However, its fundamental problem is not whether platforms remove certain content, but rather, that users depend on these platforms and it is almost impossible for them to appeal the removal of such content (Ghosh & Hendrix, 2020).

The OB is not an SMC in the sense of the ARTICLE 19 proposal. It is more like a platform council, i.e. a council that monitors only one platform (or platforms owned by the same company) (Fertmann & Kettemann, 2021). The OB is an interesting experiment in that sense, and only the future will show whether Facebook and Instagram follow its guidance or whether the OB only creates conflict with Meta, which will lead to the bitter end of this experiment. One key test is the recent hearing for the whistleblower and leaker of the so-called Facebook Papers, Frances Haugen. The revelations regarding how the moderation efforts of Facebook focus only on the English-speaking world are especially significant. The controversy highlights the fact that such issues are global and that even global companies are not able to moderate content in every country and in different languages. Therefore, there is a need for regional or national SMCs, as ARTICLE 19 proposes.

Lessons from Media Councils

Media Councils in a Nutshell

The Alliance of Independent Press Councils of Europe (AIPCE) describes press (media) council functions as monitoring codes of ethics/practice and defending press freedom. The AIPCE defines councils as self-regulating bodies set up by the media. AIPCE is the most important union for press councils. For example, the World Association of Press Councils (WAPC) describes itself as a defender of free speech, but none of the members of the AIPCE members are members of the WAPC; instead, members include, for example, Turkey, Kenya, Zimbabwe and other councils from countries not known for free media. Therefore, a media council can also be part of an authoritarian state.

The media councils can serve as potential benchmarks for social media councils (Tambini, Danilo, & Marsden 2008). Therefore, it is essential to focus on the media councils in democratic states, i.e. members of the AIPCE. Most media councils are only established as a result of regulatory pressure from regulators (state) and public. The pressure is often driven by the growing importance of privacy and questionable

practices by the press. When media councils first adopted self-regulatory measures, the key issue was the publication of a person's name.

Recent research (Neuvonen, 2005) has claimed that self-regulation began in the year 1900, when Svensk Publicisklubben made recommendations regarding the publication of names. The impact of the recommendations remained limited, though (Weibull & Börjesson, 1995, pp. 58–60). However, the first ethical code of rules were published in the US by the Kansas Editorial Association in 1910 and in France by the Syndicat National des Journalistes in 1918. The problem with the first codes is that their scope and efficiency varied. Still, more codes were issued after the First World War, and after the Second World War almost all Western European countries had a code of practices/ethics (Laitila, 1995; Thorgeirsdottir, 2002, p. 460).

The first media council, the Swedish Pressens Opinionsnämd, was established in 1916. Some sources claim that some kind of council was already established in Norway in 1915, but this claim remains questionable (Heinonen, 1995, p. 72). Claims have likewise been made that a council was established in Croatia in 1910; this may have been the year when the Croatian Journalists' Association was founded, which is similar to ones founded in Slovenia and Norway. Therefore, in some cases it is difficult to distinguish self-regulation from general journalistic associations. It should be mentioned that in the US, the National News Council was active in the years 1973–1984 but was dissolved due to resistance from journalists who opposed any kind of media regulation. However, in 1971 the Minnesota News Council was still active. Nonetheless, the US is almost the only Western nation that actually opposes self-regulation.

The AIPCE has compiled a data bank on media councils and this introduction is based on that data. It collected this data via interviews mostly with members of European councils as well as Canadian councils. Seventy-five per cent of the councils focus on television and radio as well. The interviews also accounted for weblogs produced by a media outlet. It is interesting that 50% of the council reportedly did not include media outlets owned by individuals as members and 53% did not include media outlets owned by organizations as members. That indicates that structures and the level of inclusiveness vary. In some countries, council is a synonym for an association of journalists while in others the competence of the council is based on law.

From the viewpoint of SMCs, it is interesting that 66% of councils reportedly cover user comments on media outlet websites, while 47% of councils cover user comments on the social media page of media outlets. Furthermore, 91% of councils cover posts by media outlets on social media platforms, while 47% cover posts by individual journalists on social media platforms. Therefore, the media councils already have in some countries the competence to monitor third-party posts and posts by journalists. However, this also depends on how media outlets and journalists are defined.

Altogether, this databank, developed by Vereniging van de Raad voor de Journalistiek Belgium, is impressive. However, the interview approach has given rise to a few discrepancies in relation to academic studies on media self-regulation, but not to any significant degree. The main issue from the standpoint of SMCs is that media councils are products of the media environment of each country. Therefore, media councils can be models for SMCs, with the caveat that some media councils are controlled by governments.

Tale of the Three Councils

The Finnish Council for Mass Media (FCMM) is one of the few councils that has not undergone a major reorganization and that still plays a significant role in the media sphere. The FCMM was founded 1968 and

has existed largely unchanged since then. The biggest change is that since 2016, the position of chairperson has been made a full-time position. In comparison with some councils, the FCMM regulates practically all media from newspapers to celebrity (gossip) magazines, radio, children's magazines, political party newspapers, television and internet publications.

In addition, the FCMM was founded by both publishers and journalists. It consists of a chairperson and thirteen members. Eight members represent the media and eight members represent the general public. Therefore, the FCMM is not just a press club or a society merely owned by journalists.

Studies have shown that Finnish journalists are extremely committed to their ethical guidelines and to decisions made by the FCMM. Similarly, different studies have shown that most people in Finland have a high level of trust in the country's traditional news media. The FCMM was founded at a time when there were no restrictions on publishing news about a person's private life, aside from libel (Neuvonen, 2005). This led to several instances of journalistic overreach and prompted parliament to consider instituting regulations. The idea of self-regulation was to keep the state at arm's length from directly regulating content. Despite the establishment of the FCMM, new laws were enacted in 1973 even as self-regulation was widely adopted by the Finnish media.

The FCMM interprets the guidelines for journalists. It can also issue policy statements regarding questions of professional ethics and handles complaint investigations free of charge. The framework for the FCMM's operations are stipulated in its charter, which all media organizations have signed, committing themselves to self-regulation and accepting its objectives.

The key word accounting for the longevity of the FCMM is trust: journalists trust self-regulation, the public trusts journalism, media companies are committed to self-regulation and the state trusts self-regulation. Another key issue is coverage, with the FCMM covering practically all Finnish media and including members from outside the media.

However, the history of Finnish self-regulation is not just a straightforward success story. The first council, Suomen Sanomalehdistön Kunniaoikeusto, was established 1927 to settle disputes between journalists. It was a complete failure (Vuortama, 1984). The first code was drawn up in 1958 by Finnish media associations. The effort did not prevent the press from publishing scandalous news and, after the death of the famous writer Timo K. Mukka, attitudes towards the press became more judgmental. The public blamed a few scandalous publications for contributing to the writer's death. The JSN was established out of fear of government regulation, but it did not prevent stricter laws from being passed in 1973 (Neuvonen, 2005).

In April of 1916, the Swedish newspaper *Ny Dagligt* published a private letter, causing a scandal in that country. The scandal led to calls for greater regulation of the press. As a result, the first modern press council, Pressens Opinionsnämnd (PON), was established in Sweden in 1916. It only covered the press and suffered from a lack of funding in the early years, leaving 70% of complaints unresolved (Weibull & Börjessön, 1995, pp. 67–70). Sweden is in many ways similar to Finland, but the press council was in crisis in the 1960s due to more scandalous press reporting. The solution was to remodel the system and establish a new position of authority called the Pressombudsmannen in 1969.

The PON adheres to a strict publication code (publitetsregel). In addition, professional issues are regulated via a professional code (yrkesregel), which is interpreted by Council of Professional Ethics, the Yrkesetiska nämden, part of the Swedish Union of Journalists (Journalistförbundet). Television and radio are self-

regulated by the Swedish Press and Broadcasting Authority (Granskningsnämnden för radio och tv), while magazines are regulated by the Swedish Media Publishers' Association (Tidningsutgivarna).

The English Press Council (PC) was long considered an example of media councils. The first Royal Commission on the Press recommended in 1949 that a General Council of the Press should be formed to govern the behaviour of print media publishers (Frost, 2000, p. 177). The PC was criticized during its first decades of operation. The Second Royal Commission on the Press recommended that some council members should be members of the general public without any connections to the press. The press and PC were extensively criticized in the 1973 Younger Committee Report on Privacy, prompting the third Royal Commission on the Press to suggest the adoption of a written code of conduct for the press.

During the 1980s, the PC lost the confidence of journalists and publishers. As a result, in 1991 the Press Complaints Commission (PCC) replaced the PC based on findings reported in the Calcutt Report (Shannon, 2001). The PCC also drafted a code of conduct. However, the PCC was later dissolved after a phone hacking scandal in the early 2010s, uncovered by the Leveson Inquiry, which was a judicial public inquiry into the culture, practices and ethics of the British press. The resulting Leveson Report argued that the PCC could not efficiently regulate the press and recommended a new voluntary regulatory body. The new system does not involve self-regulation in its purest form, though it is based on the 2013 Royal Charter on self-regulation of the press (hereinafter the Charter). The Charter created a Press Recognition Panel to ensure that regulators of the press and other news publishers are independent, properly funded and able to protect the public. The first regulator recognized by the PRB was the Independent Monitor for the Press (IMPRESS). However, none of the large national publishers are not member of IMPRESS. Instead, most national publications are members of the Independent Press Standards Organisation (IPSO), which the PRB does not recognise. In addition, several prestigious newspapers (e.g. *The Guardian*, *Financial Times*) have established their own independent complaint systems. Therefore, the current status of media self-regulation in the UK seems quite complicated.

What Can Social Media Councils Learn from Media Councils?

The focus of the article has been on three media councils. Two of the three councils have been considered positive examples for decades, but both have had difficult times. In contrast, the Finnish Council for Mass Media has been quite stable for more than 50 years. The same applies to a few other councils. However, the basis for a council's competence varies, as does other aspects of how councils operate. Media councils have largely followed the example of other councils, but still no single prototype exists for a media council. In addition, in some countries the council is controlled by the government or else it is completely neglected.

As we have seen especially with respect to the UK regulators, the key issue is trust between stakeholders. The regulator (the state) must trust that publishers and journalists follow the rules set by the council. The public must trust that publishers and journalists will act ethically and follow the rules. Therefore, all participants must trust that the council tries its best and is impartial. At the beginning of media self-regulation, neither the state nor the general public trusted journalists and publishers. As a result, demands for stricter media regulation intensified. In Sweden in the 1960s, journalists and publishers lost faith that the council benefited the entire press. In the UK, the press has always been sceptical of self-regulation because self-regulatory bodies have appeared biased. Thus, the system has faltered and stumbled from crisis to crisis.

The case of the FCMM indicates that a few key elements are essential to earning the trust of the state and general public. First, the scope of the council must be as wide as possible. Second, regulation must be based on rules. Third, the council must represent all sectors, not just publishers or journalists. Fourth, some members of the council must be members of the general public. Fifth, self-regulation must be open and transparent.

However, the Nordic countries rank highly in their guarantees of freedom of speech, and each country has also developed an exemplary self-regulatory system, the so-called Nordic welfare state model (Syvertsen et al., 2014). It should be noted that all the Nordic countries have strong public service media outlets as well as strong national and independent, family-owned media. In comparison to the UK, the Nordic countries do not contain global media giants on the same scale. Therefore, the media owners tend to know their country and its culture quite well, and usually family-owned businesses have some added value other than just making a profit, especially in the media business, even though they are listed companies. The situation with social media is completely different. Almost all major social media companies are in the US and are listed on the stock exchange. The major competitors are found in China, and, for example, one of the most noteworthy regional companies is the Russian firm vKontakte. If it is difficult to regulate American companies, it is much harder to influence Chinese companies. Therefore, two of the examples were from Nordic countries, which represent exceptions in many ways in comparison to larger, more global social media companies. So, the SMCs could follow the path of self-regulators in Sweden and Finland and have a significant impact on UK regulators, which may cause problems and result in the need for constant remodifications. But numerous alternatives exist.

How can the lessons be applied to social media? First, SMCs cannot focus on just one platform; they must regulate all platforms that meet certain criteria. Second, SMCs must have their own ethical rules for regulating the activities of platforms at a higher level of scrutiny. Platforms have their own rules, both general public rules and secret detailed rules, but these rules need to be based on rules set by SMCs. Third, SMCs must include members who are not connected to the platforms. Fourth, the main asset of SMCs is the level of trust between different stakeholders, and that trust must be earned; there are no instant recipe for success. Fifth, the whole idea of SMCs is utopian, so the first step must be both realistic and credible.

References

Bygrave, L. A. (2015). *Internet governance by contract*. Oxford University Press.

Fertmann, M., & Kettemann, M. C. (2021): One Council to Rule Them All: Can Social Media Become More Democratic? Digital Society Blog, 22 June 2021, doi: 10.5281/zenodo.4773130.

Frost, Chris: Media Ethics and Self-Regulation. London 2000.

Ghosh, D., & Hendrix, J. (2021, December 19). *Facebook's Oversight Board just announced its first cases, but it already needs an overhaul*. VerfBlog. https://verfassungsblog.de/fob-first-cases/

Gillespie, T. (2018). *Custodians of the internet. Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gradoni, L. (2021, February 10). *Constitutional review via Facebook's Oversight Board: How platform governance had its Marbury v Madison.* VerfBlog. https://verfassungsblog.de/fob-marbury-v-madison/

Heinonen, Ari: Vahtikoiran omatunto. Tampere 1995.

Klonick, K. (2018). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, 131, 1598

Laitila, Tiina: Codes of Ethics in Europe in Reports on Media Ethics in Europe (Kaarle Nordenstreng, eds), 23-80. Tampere 1995.

Napoli, P. (2019). *Social media and the public interest: Media regulation in the disinformation age*. Columbia University Press.

Neuvonen, Riku: Sananvapaus, joukkoviestintä ja sääntely. Helsinki 2005.

Pollicino, O., De Gregorio, G., & Bassini, M. (2021, May 11). *Trump's indefinite ban: Shifting the Facebook Oversight Board away from the First Amendment doctrine.* VerfBlog. https://verfassungsblog.de/fob-trump-2/

Pollicino, Oreste. Judicial Protection of Fundamental Rights on the Internet: a Road Towards Digital Constitutionalism? Oxford, UK: Bloomsbury Publishing Plc, 2021. Print.

Shannon, Richard: A Press Free and Responsible: Self-regulation and the Press Complaints Commission, 1991-2001. John Murray 2001.

Singh, S. (2019, July 22). Everything in moderation. An analysis of how internet platforms are using artificial intelligence to moderate user-generated content. New America. https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/

The Soufan Center. (2021, April). Quantifying the Q conspiracy: A data-driven approach to understanding the threat posed by QAnon. https://thesoufancenter.org/wp-content/uploads/2021/04/TSC-White-Paper_QAnon_16April2021-final-1.pdf

Suzor, Nicholas P. Lawless: the Secret Rules That Govern Our Digital Lives. Cambridge, United Kingdom: Cambridge University Press, 2019. Print.

Syvertsen, T., Gunn, E., Mjøs, O., and Moe, H. (2014), The Media Welfare State: Nordic Media in the Digital Era, Ann Arbor, MI: The University of Michigan Press.

Tambini, D., & Marsden, C. T. (2007). *Codifying cyberspace: Communications self-regulation in the age of internet convergence*. Routledge.

Thorgeirsdottir, Herdis: Journalism Worthy of the Name. Lund 2002.

Weibull, Lennart, & Börjesson, Brit: Publicitiska seder. Falun 1995.

Vuortama, Timo: Hyvä lehtimiestapa. Rauma 1984.

EU COST Action – CA19143: Global Digital Human Rights Network

The GDHRNet COST Action will systematically explore the theoretical and practical challenges posed by the online context to the protection of human rights. The network will address whether international human rights law is sufficiently detailed to enable governments and private online companies to understand their respective obligations vis-à-vis human rights protection online. It will evaluate how national governments have responded to the task of providing a regulatory framework for online companies and how these companies have transposed the obligation to protect human rights and combat hate speech online into their community standards. The matters of transparency and accountability will be explored, through the lens of corporate social responsibility.

The Action will propose a comprehensive system of human rights protection online, in the form of recommendations of the content assessment obligation by online companies, directed to the companies themselves, European and international policy organs, governments and the general public. The Action will also develop a model which minimises the risk of arbitrary assessment of online content and instead solidifies standards which are used during content assessment; and maximises the transparency of the outcome.

The Action will achieve scientific breakthroughs (a) by means of a quantitative and qualitative assessment of whether private Internet companies' provide comparable protection of human rights online in comparison with judicial institutions, and (b) in the form of a novel holistic theoretical approach to the potential role of artificial intelligence in protecting human rights online, and (c) by providing policy suggestions for private balancing of fundamental rights online.

Contact: Dr Mart Susi, Action Chair, mart.susi@tlu.ee

Dr Vygantė Milašiūtė, Action Vice Chair, vygante.milasiute@tf.vu.lt

Mr Gregor Fischer-Lessiak, Science Communications Manager, gregor.fischer@uni-graz.at

WG 1 - Fundamental and vertical dimension of human rights online
Dr Tiina Pajuste, WG leader, tiina.pajuste@tlu.ee

WG 2 - Practical dimension of human rights online
Professor Dr Matthias C. Kettemann, WG leader, m.kettemann@leibniz-hbi.de

WG 3 - Dissemination, Exploitation and Sustainability
Dr Giovanni De Gregorio, WG leader, giovanni.degregorio@csls.ox.ac.uk

Action Management Committee:
https://www.cost.eu/actions/CA19143/#tabs|Name:management-committee