KU Leuven
Faculty of Law

European Master's Programme in Human Rights and Democratisation
A.Y. 2021/2022

The Power of Social Media (Regulation)

The Invisible Effect of Platform Governance on Freedom of Expression

Author: Wiebke Groth
Supervisor: Prof. Dr. Koen de Lemmens

Table of Abbreviations

AI	Artificial Intelligence
DSA	Digital Services Act
ECHR	European Convention of Human Rights
ECJ	Court of Justice of the European Union
ECtHR	European Court of Human Rights
FoE	Freedom of Expression
ICCPR	International Covenant on Civil and Political Rights
SRFoE	United Nations Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression
ToS	Terms of Service
UDHR	Universal Declaration of Human Rights
UGC	User-generated content
VLOPs	Very Large Online Platforms

Table of Definitions

Artificial Intelligence	"'Constellation' of processes and technologies enabling computers to complement or replace specific tasks otherwise performed by humans"[1]
Algorithm	"Computer code that translates data into outputs, conclusions, and information"[2]
Content Moderation	Social media companies' enforcement of their community guidelines[3] (here: Synonym for self-governance and -regulation)
Content Regulation	State initiatives aiming to remove content from the online domain[4]
(Internet) Intermediaries	"Give access to, host, transmit and index content […] originated by third parties on the internet"[5]
Internet Service Provider	"Subtype of intermediaries that connect devices with the internet."[6]
Platformization	"The penetration of economic, governmental, and infrastructural extensions of digital platforms into [cultural systems]"[7].
Platform Governance	"Covers both concepts of internal management and control by the platform operators themselves, and regulation or influencing from outside."[8]
Social Media Company	Subtype of internet intermediaries (e.g., Meta, Twitter, or Google)
Social Media Platform	"Sites and services that host public expression, store it on and serve it up from the cloud, organize access to it through search and recommendation, or install it onto mobile devices"[9]. They host and organize content from third parties for public circulation. (e.g., Facebook, YouTube, Twitter, Instagram or TikTok). (also: online platform)

[1] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/73/348' (UN General Assembly, 2018), 3.
[2] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 4.
[3] Rikke Frank Jørgensen and Lumi Zuleta, 'Private Governance of Freedom of Expression on Social Media Platforms: EU Content Regulation through the Lens of Human Rights Standards', *Nordicom Review* 41, no. 1 (2020): 57.
[4] Jørgensen and Zuleta, 57.
[5] Karine Perset, 'The Economic and Social Role of Internet Intermediaries', OECD Digital Economy Papers, vol. 171, OECD Digital Economy Papers (OECD, 2010), 9.
[6] Rebecca MacKinnon et al., 'Fostering Freedom Online. The Role of Internet Intermediaries', Serien on Internet Freedom (UNESCO Publishing, 2014), 22.
[7] David B Nieborg and Thomas Poell, 'The Platformization of Cultural Production: Theorizing the Contingent Cultural Commodity', *New Media & Society* 20, no. 11 (November 2018): 4276.
[8] Matthias Cornils, 'Designing Platform Governance', Governing Platforms (Algorithm Watch, May 2020):12.
[9] Tarleton Gillespie, 'Governance of and by Platforms', in *Handbook of Social Media*, ed. Jean Burgess, Thomas Poell, and Alice Marwick (SAGE, 2017), 1.

Abstract

Cambridge Analytica, the Brexit referendum, and the election of Donald Trump; These three famous cases remind every social media user that they need to be sceptical about the information they are shown on their feeds. After the initial shock of these cases, awareness about the potentially detrimental interference of social media grew. Consequently, the question of effective platform governance through enhanced regulation became more widespread. To contribute to this debate, this thesis will explore current social media regulations with a focus on Europe and the EU and discuss their implementation, strengths, weaknesses, and effects on Freedom of Expression. Through painting a comprehensive picture of current platform governance, this thesis was dedicated to the question of how the regulation and moderation of social media platforms affect users' right to Freedom of Expression. In this field, the investigation of law, and content moderation techniques in combination with an exploratory study to research their adverse effects are a rare find. Thus, the thesis first outlines how Freedom of Expression is protected and can legitimately be restricted. Secondly, it discusses how platforms are governed through policies, their terms of service and the efficacy of their mechanisms. Thirdly, it explores cases of when Freedom of Expression was restricted through methods of digital repression and consequently, how this affects social movements. Lastly, possible solutions and the new Digital Services Act of the EU will be explored. Overall, this thesis found that the current situation is a significant hazard to Freedom of Expression on the internet. The research reflects that overblocking happens regularly and thus, legal speech in our current online environment can be suppressed.

Table of Contents

Introduction

"In a lot of ways
Facebook is more like a government than a traditional company.
We have this large community of people,
and more than other technology companies
we're really setting policies."
- *Mark Zuckerberg*[10]

Today, society is very much aware of the potential (negative) power of social media. Through famous cases such as Cambridge Analytica, when profiles of Facebook accounts were used to influence political campaigns[11], or the thousands of Russian bots that were spreading disinformation about the US national election in 2016[12], people became more aware of this.

The quote above by Mark Zuckerberg, one of Facebook's founders, highlights the influence of social media and the astonishing similarities of platform company practices with governments. In other words, Facebook's policies seemingly have a similar effect on peoples' lives as governments' policies. Yet, the owners of Facebook are not elected leaders. As a matter of fact, users have absolutely no influence on who creates the rules of the social media platforms where they post their content. The immense power of social media platforms on Freedom of Expression is thus a very current concern that is on the minds of individuals and policymakers alike.

In April 2022, more than 4.65 billion people, which is 59 % of the world's population, used social media.[13] Social media platforms have fundamentally transformed the way we communicate and receive information. For a long time, the possibility for everyone with access to the internet to post content online was seen as a reinforcement of the information environment and therefore, democratic processes. Individuals, especially those who are part of marginalized groups, now had the chance to gain attention from new audiences and found new ways of organizing themselves to voice their grievances. Social media users can post about their experiences, thoughts, and opinions online. These posts are the "lifeblood of social media platforms"[14], they are social media's currency but also a great liability. Thus, social

[10] David Kirkpatrick, *The Facebook Effect: The Inside Story of the Company That Is Connecting the World*, 1st ed. (New York: Simon & Schuster Paperbacks, 2011), 254.
[11] Nicholas Confessore, 'Cambridge Analytica and Facebook: The Scandal and the Fallout So Far', *The New York Times*, April 2018.
[12] Scott Shane, 'The Fake Americans Russia Created to Influence the Election', September 2017.
[13] Simon Kemp, 'Digital 2022: April Global Statshot Report', DataReportal, April 2022.
[14] Sarah T. Roberts, 'Digital Detritus: "Error" and the Logic of Opacity in Social Media Content Moderation', *First Monday* 3, no. 23 (March 2018).

media has become increasingly important in our daily lives in the last decades. They function as *social infrastructures*[15]. Everyone can post on these social media platforms, and consequently, illegal content – from hate speech to sexual exploitation of children - is unavoidable. Content as such and its dissemination needs to be limited and regulated in order to protect users.

However, regulation of online content is challenging and needs to be balanced with Freedom of Expression. Freedom of Expression is at the heart of functioning democracies and a fundamental human right. Any limitation of it needs to be thoroughly thought through and taken seriously. Thus, this thesis attends this topic and tries to discuss the idea of social media regulations and their effects on our online communication. Specifically, the following will explore the challenge of the platform governance and examine its effects on Freedom of Expression and social movements worldwide.

[15] They are "embedded, taken for granted, rules by unquestioned standards and largely unseen services that govern public action". Rikke Frank Jørgensen, *Human Rights in the Age of Platforms*, Information Policy Series (Cambridge, MA: MIT Press, 2019), 166.

The Power of Social Media

In 2011, Facebook pages such as "Egyptian supporting the Tunisian revolution" and "We are all Khalid Said[16]" were created and the hashtag #Jan25 - calling people to go on the streets to protest the regime on National Police Day - quickly reached two million tweets on Twitter.[17] The subsequent protest occupied Cairo's Tahrir Square for 18 days and succeeded in forcing President Mubarak's resignation.[18]

This movement was part of the Arab Spring Revolution around 2011. It resulted in numerous discussions about the power of social media in mobilizing political dissident. The protests have even been called the "Twitter" or "Facebook Revolution"[19]. Social media platforms were also found to be a substantial factor in the organization and the distribution of information about the Turkish and Ukrainian protest movements in 2013 and 2014.[20] Since then, there have been extensive discussions on the effect of social media platforms on political and social movements.

The power of media was found to be a two-step process: First, opinions are transmitted by the media and secondly, they get echoed by people's social environment.[21] As a result, social media has the potential to lower the costs of protesting by facilitating access to information about, for instance, the time and place of protests and when a protest becomes dangerous (e.g., police presence or violence).[22] Furthermore, social media potentially fosters motivation for activism due to heightened frustration at perceived injustice and strengthens group identity and empowerment.[23] These positive aspects led to many realizing that social media can facilitate change and democratization processes and give voice to more people.

Yet, the power of social media as a force for positive change is contentious. It can harm movements as much as it can strengthen them. As much as social media can strengthen democracies, it can also reinforce authoritarian regimes. For instance, the control of the Chinese Communist Party (CCP) over social media has seemingly strengthened the regime.[24]

---

[16] Said was beaten to death by Egyptian police officers, pictures of his dead body were spread quickly on social media. See 'We Are All Khaled Said', https://www.facebook.com/elshaheeed.co.uk/.
[17] Halim Rane and Sumra Salem, 'Social Media, Social Movements and the Diffusion of Ideas in the Arab Uprisings', *Journal of International Communication* 18, no. 1 (April 2012): 104.
[18] 'Egypt Revolution: 18 Days of People Power', Aljazeera, January 2016.
[19] Rane and Salem, 'Social Media, Social Movements, and the Diffusion of Ideas in the Arab Uprisings', 97.
[20] Joshua A. Tucker et al., 'Big Data, Social Media, and Protest', in *Computation Social Science* (Cambridge University Press, 2016), 212.
[21] Clay Shirky, 'The Political Power of Social Media: Technology, the Public Sphere, and Political Change', *Council of Foreign Relations* 90, no. 1 (2011): 34.
[22] John T. Jost et al., 'How Social Media Facilitates Political Protest: Information, Motivation, and Social Networks: Social Media and Political Protest', *Political Psychology* 39 (February 2018): 88.
[23] Jost et al., 94.
[24] Shirky, 'The Political Power of Social Media: Technology, the Public Sphere, and Political Change', 39.

The CCP's Golden Shield Project effectively blocks content which is against Chinese interests through keyword filters leading to censorship, surveillance, propaganda, and self-censorship.[25] As a result, protesting and political expression are significantly suppressed.[26] Thus, social media can be a medium for censorship as well. But it is not just autocracies that rely on social media to further their agenda, democracies rely more and more on digital censorship activities.[27] For instance, democratic states which are particularly interested in the information economy tend to restrict citizens' access to information more.[28]

In general, social media gives every user the ability to communicate his or her thoughts to potentially every social media user around the world. It significantly transformed what we understand as Freedom of Expression (FoE) which is about societies that are enabling inclusive discussions where narratives can develop.[29] Through social media the form of public expression changed, through, for instance, liking a post instead of giving a response in writing.[30]

The perception of social media changed over the last years; Zittrain identified waves of thinking about the internet, social media and its governance going from a positive to a more negative notion over time.[31] The first wave was primarily about rights and positive framing, while the second wave was about risks, harms and negative ramifications of the digital landscape and abstract connections.[32] The conversations were primarily about end-user rights following the notion of free speech without any gatekeepers.[33] This period was signified by little to no liability of intermediaries, making them almost immune to any sanctions.[34] The second era, the "public health era"[35] started in the late 2000s. The interlinkages between internet users were emphasized, one's problem can become another's problem. Cybersecurity

[25] For more details about China's filtering system, see Richard Clayton, Steven Murdoch, and Robert Watson, 'Ignoring the Great Firewall of China', in *Privacy Enhancing Technologies* (Berlin: Springer, 2006), 20–35.
[26] Jiayin Lu and Yupei Zhao, 'Implicit and Explicit Control: Modeling the Effect of Internet Censorship on Political Protest in China', *International Journal of Communication* 12 (2018): 3310f.
[27] Stephen A. Meserve and Daniel Pemstein, 'Google Politics: The Political Determinants of Internet Censorship in Democracies', *Political Science Research and Methods* 6, no. 2 (2018): 246.
[28] Meserve and Pemstein, 259.
[29] Ben Wagner, 'Free Expression? Dominant Information Intermediaries as Arbiters of Internet Speech', in *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*, ed. Martin Moore and Damian Tambini (New York: Oxford University Press, 2018), 219.
[30] András Koltay, *New Media and Freedom of Expression: Rethinking the Constitutional Foundations of the Public Sphere*, Hart Studies in Comparative Public Law (2019), 147.
[31] Jonathan Zittrain, 'Three Eras of Digital Governance', *SSRN Electronic Journal*, 2019, 1–8.
[32] Terry Flew and Rosalie Gillett, 'Platform Policy: Evaluating Different Responses to the Challenges of Platform Power', *Journal of Digital Media & Policy* 12, no. 2 (2021): 4.
[33] Zittrain, 'Three Eras of Digital Governance', 2.
[34] Zittrain, 3.
[35] Zittrain, 4.

was a popular topic in this era. After a couple of *public shocks*[36], the wish for more regulation of social media platforms became more dominant. Simultaneously, the public turned more critical of social media and its impacts on society and prevailing political structures. These shocks revealed the truth to the public about how platforms operate and brought unacceptable aspects to light.[37] Zittrain deems the current area as the area of digital governance or legitimacy.[38] As a result, the demand for more internet governance grew.

Internet or more specifically, platform governance is signified by fragmented responsibilities between governments, the intermediaries owning the platforms and users.[39] The task to harmonize their interest is challenging and the concept of power is inherently related to platform governance. Flawed governance can easily burden Freedom of Expression, political participation, social movements, and democratic processes.

All in all, the emergence of the internet and social media platforms created dependences in the distribution of information and thereby, giving platforms immense amounts of power. Through social media, public discourse was substantially expanded and simultaneously, opinions can effortlessly be suppressed.[40] This is due to the dependency on "quasi-monopolistic internet intermediaries"[41] Consequently, this transformation of how we communicate makes expression vulnerable. Interference through regulations is needed but dangerous. The effect of misinformation[42] needs to be mitigated but at the same time, regulation leading to overblocking needs to be avoided. Thus, this paper aims to answer the following research question:

How does platform governance affect users' right to Freedom of Expression?

Accordingly, this paper will first outline the legal framework. Subsequently, the practices of intermediaries and their interests and finally, cases of digital repression will be examined to conclude the current state of the public sphere and Freedom of Expression.

[36] Shocks which highlight platform infrastructural qualities and interrupt the functioning and governance of the platforms. Mike Ananny and Tarleton Gillespie, 'Public Platforms: Beyond the Cycle of Shocks and Exceptions', 2017, 2.
[37] Ananny and Gillespie, 6.
[38] Zittrain, 7f.
[39] Gorwa, 'What Is Platform Governance?', 855.
[40] Koltay, *New Media and Freedom of Expression*, 149.
[41] Wagner, 'Free Expression? Dominant Information Intermediaries as Arbiters of Internet Speech', 220.
[42] Besides the concerns about the effects of mis-/disinformation on the Trump Election in 2016 and the Brexit referendum, there have been reports about e.g., the killing of a law student and his uncle due to the spread of false information about them being them being child abductors. See Patrick McDonnell and Cecilia Sanchez, 'When Fake News Kills: Lynchings in Mexico Are Linked to Viral Child-Kidnap Rumors', Los Angeles Times, September 2018.

This thesis aims to explore the fine line between the legitimate and necessary and the illegitimate and dangerous restriction of content on social media through the lens of Freedom of Expression. For this, the following subquestions will be explored:

I. How is Expression on Social Media protected under law?

To answer this question, it is necessary to take a look at the international human rights framework of the United Nations. Furthermore, this thesis will focus on Europe and EU policies, thus, it is necessary to additionally explore the European Human Rights Framework regarding Freedom of Expression and its restrictions. Both will be delved into in the first chapter. The second subquestion is:

II. How are social media platforms governed?

Platforms are governed by their companies *and* governments. In other words, there is self-governance and external governance.[43] On the one hand, external governance (or content regulation) puts some of the liability back to the companies. NetzDG, an often-cited German law, is an example of this, whereby companies are forced to take down certain content within 24 hours (see also chapter 2, section 1). Self-governance, on the other hand, is enforced by platforms.[44] This question will be discussed in chapter 2 by researching EU legislation and policies and content moderation techniques employed by some of the most popular social media platforms. The third subquestion is:

III. Can platform governance suppress social movements?

To answer this question, this thesis will take on the perspective of digital repression. Digital repression is a concept which explains online censorship which especially targets social movements. Through the exploration of whether such cases can be found in recent years, this thesis aims to answer the question of negative effects on social movements through restricting online speech and Freedom of Expression. This question will be explored in chapter 3.

Subsequently, the last sections of this thesis will discuss the new Digital Services Act by the EU and other potential approaches to protect Freedom of Expression online more comprehensively. Finally, the findings of these sections will be discussed and concluded.

Th following chapter will aim to answer the first subquestion regarding Freedom of Expression in Human Rights Law.

[43] Gorwa, 'What Is Platform Governance?', 862f.
[44] Gorwa, 862.

Chapter I: Freedom of Expression on Social Media

Freedom of Expression is a fundamental right, especially for democratic and pluralist societies. Yet, not everything is allowed to be expressed as some speech can violate protected rights of others. To dive deeper into this, this chapter will outline how Freedom of Expression online is protected under International and European law and under what circumstances it can be restricted.

First, the United Nations framework regarding Freedom of Expression (FoE) with the addition of general comments and reports of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (henceforth: Special Rapporteur or SRFoE) will be researched. It follows a more doctrinal legal approach aiming to answer of when Freedom of Expression is protected and when it can be restricted (subquestion I). Case law selected through a keyword search[45] regarding its relevance to Freedom of Expression on the Internet will support this analysis.

First, before diving deeper into European law, is it important to have a look at International Human Rights law. For this purpose, the following section will examine what the United Nations specify about Freedom of Expression and its legitimate and illegitimate restrictions.

1. The United Nations Framework

"A free, uncensored and unhindered press or other media is essential in any society"[46] describes the notion of the UN toward Freedom of Expression perfectly. The United Nations General Assembly underlined the importance of FoE already in 1946 by identifying it as "fundamental" and "the touchstone of all […] freedoms"[47] while underlining the importance of international cooperation.[48] It was first recognized in Article 19 of the non-binding Universal Declaration of Human Rights (UDHR).[49] Years later, in 1966, an internationally binding instrument was created with the International Covenant on Civil and Political Rights (ICCPR)[50].

[45] The keywords included e.g., "Freedom of Expression" with the addition of "online" or "social media", see appendix 1.
[46] Human Rights Committee, 'General Comment No. 34. Article 19: Freedom of Expression', 2011, para. 13.
[47] United Nations General Assembly, 'Resolution 59(I)'.
[48] Agnès Callamard, 'The Human Rights Obligations of Non-State Actors', in *Human Rights in the Age of Platforms*, ed. Rikke Frank Jørgensen (Cambridge: MIT Press, 2019), 193.
[49] United Nations, 'UDHR' (1948).
[50] The Covenant is currently ratified by 173 states. 'Status of Ratification. International Covenant on Civil and Political Rights', OHCHR, April 2022.

Article 19 ICCPR states:

1. "Everyone shall have the right to hold opinions without interference.
2. Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice.
3. The exercise of the rights provided for in paragraph 2 of this article carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary:
   (a) For respect of the rights or reputations of others;
   (b) For the protection of national security or of public order (ordre public), or of public health or morals."[51]

Article 19 applies "regardless of frontiers" and to "any […] media [...] of choice"[52]. Media includes non-state actors from newspaper owners to internet intermediaries.[53] It refers to not only rights but also duties and responsibilities. The obligation to protect, thus, the horizontal dimension of the right, means that FoE is protected against interference by public authorities and private parties.[54] This also includes the prevention of "monopolistic situations"[55] of privately controlled media. Also, 19(2) includes "electronic and internet-based modes of expression"[56] and protects "even expression that may be regarded as deeply offensive"[57].

States have a positive and negative obligation to refrain from violating FoE and to ensure its protection through, for instance, making sure intermediaries are taking steps to protect the right.[58] They must ensure media diversity, and media independence and that private entities do not interfere with Freedom of Expression and Freedom of Opinion.[59] The Human Rights Committee (HRC), which supervises the ICCPR, emphasised the importance of clear definitions of, for instance, "extremist activity" which is often claimed as a legitimate restriction of FoE online.[60]

Freedom of Expression has a private dimension, Freedom of Opinion (i.e., the right to hold and to form an opinion) which is absolute. Thus, no limitations and restrictions are possible.

---

[51] United Nations, 'International Covenant on Civil and Political Rights' (1966).
[52] United Nations.
[53] Callamard, 'The Human Rights Obligations of Non-State Actors', 198.
[54] Manfred Nowak, *UN Covenant on Civil and Political Rights: CCPR Commentary*, 2nd ed. (Kehl am Rhein: Engel, 2005), 448.
[55] Human Rights Committee, 'United Nations', para. 40.
[56] Human Rights Committee, 'General Comment No. 34. Article 19: Freedom of Expression', 2011, para. 12.
[57] Human Rights Committee, para. 11.
[58] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/HRC/32/38' (Human Rights Council, 2016), para. 8.
[59] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/HRC/38/35' (Human Rights Council, 2018), para. 6.
[60] Human Rights Committee, 'United Nations', para. 46.

Freedom of Opinion has two dimensions; The internal dimension is connected to the Right to Privacy and Freedom of Thought (Article 18) and the external dimension is connected to FoE.[61] 19(1) requires states and private companies no to interfere with Freedom of Opinion. However, the line between what is allowably influencing opinions and what is not is difficult to draw.[62] For instance, state propaganda, advertising and marketing, personal conversations or information dissemination in the media influence people's opinions.[63] The HRC stated that coercive systems which are applied in a discriminatory manner in order to change political opinions of prison inmates constitutes a violation of Freedom of Opinion.[64] Thus, Freedom of Opinion is influenceable. In this regard, mental autonomy is an aspect that needs to be considered when talking about social media.

The public dimension, the right to express an opinion and to receive information, can be restricted which is specified in 19(3). The limitations of FoE need to follow a three-step test; They need to be (1) provided by law, (2) necessary and proportional and (3) legitimate.[65] The restrictions must furthermore be precise and publicly accessible.[66] In other words, any restrictions must be adopted through legal processes under the oversight of independent judicial authorities, they must be demonstrated to be necessary and to impose the least burden on the exercise of said right and the restriction must be made because of the in 19(3) formulated causes.[67] The HRC elaborated on this and warned that 19(3) should not be interpreted as a "license to prohibit unpopular speech"[68]. Further, restrictions must be "appropriate to achieve their protective function" and "must be the least intrusive instrument amongst those which might achieve their protective function".[69] Importantly, restrictions are to be assessed by bodies independent of political, commercial, and other unwarranted influences.[70]

---

[61] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/HRC/47/25' (Human Rights Council, 2021), para. 33.
[62] Nowak, *UN Covenant on Civil and Political Rights: CCPR Commentary*, 442.
[63] Nowak, 442.
[64] Yong Joo-Kang v Republic of Korea, No. CCPR/C/78/D/878/1999 (Human Rights Committee 16 July 2003) para 7.2.
[65] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/HRC/38/35', para. 7.
[66] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/HRC/32/38', para. 7.
[67] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/HRC/38/35', para. 7.
[68] Robert Faurission v France, No. CCPR/C/58/D/550/1993 (HRC 9 November 1996) para 8.
[69] Jacob Mchangama, Natalie Alkiviadou, and Raghav Mendiratta, 'A Framework of First Reference: Decoding a Human Rights Approach to Content Moderation in the Era of "Platformization"', *The Future of Free Speech Project*, 2021, 13; Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/HRC/32/38', 2016, para. 7.
[70] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/67/357' (United Nations General Assembly, 2012), para. 41.

Furthermore, there should be no restrictions on statements that are true and the dissemination of hate speech (unless it has been proven that this happened with the intent to incite discrimination, hostility, or violence) and no one should be restricted by prior censorship.[71] The expression of deep-rooted hatred can undermine the rights of others, thus, under specific circumstances, freedom of expression can be restricted.[72] Article 20(2) states that content that leads to incitement to discrimination, hostility or violence is to be prohibited.[73] Additionally, hate speech targeted toward people of other ethnicities or skin colours is prohibited under Article 4 of the International Convention on the Elimination of All Forms of Racial Discrimination.[74] For instance, in the Jewish Community of Oslo v Norway, the Committee on the Elimination of Racial Discrimination stated that Freedom of Speech is afforded lower protection when dealing with "all ideas based upon racial superiority or hatred".[75] Yet, it is crucial to underline that permitted expression "includes forms of expression that are *offensive, disturbing and shocking*"[76]. The assessment of hate speech must include "the existence of patterns of tensions between […] communities, discrimination against the targeted group, the tone and content of the speech"[77] best done on a case-by-case basis.

One of today's biggest obstacles to information online is the issue of mis- or dis-information (see also chapter 3) which has often been used to incite hatred.[78] Misinformation can be defined as unintentional, and disinformation is the deliberate dissemination of false information.[79] Misinformation can simply be incorrect captions of photos or translations while disinformation is, for instance, intentionally created conspiracy theories. The issue of

---

[71] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, para. 50.

[72] The removal of content online due to hate speech has been increasing in the last five years, e.g., Facebook deleted 2.5 million posts in the first quarter of 2018 and 31.1 million posts in the second quarter of 2021. Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, para. 37; 'Hate Speech. Transparency Center', Meta.

[73] United Nations, International Covenant on Civil and Political Rights.

[74] Article 4 is broader than Article 20(2) as it does not define any threshold (such as "incitement") merely "ideas" are already prohibited to be disseminated. United Nations, 'International Convention on the Elimination of All Forms of Racial Discrimination. Declarations and Reservations', March 1966.

[75] Quoting Article 4(a) of the Convention. The Jewish Community of Oslo v Norway, No. CERD/C/67/D/30/2003 (CERD 15 August 2005).

[76] Emphasis added. Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/67/357', para. 49.

[77] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, para. 46.

[78] Scholars have previously observed the "online disinformation-terrorism-nexus", for more information, see European Foundation for South Asian Studies, 'The Role of Fake News in Fueling Hate Speech and Extremism Online; Promoting Adequate Measures for Tackling the Phenomenon', 2021.

[79] Hossein Derakhshan and Claire Wardle, 'Information Disorder: Definitions', in *Understanding and Addressing the Disinformation Ecosystem*, by Annenberg School for Communication, First Draft, and Knight Foundation, 2017, 9.

disinformation is particularly predominant during the Covid-19 pandemic, calling it an "infodemic"[80]. For instance, the think tank First Draft found that in 2020, posts about vaccinations were prone to include conspiracy theories (with 84% of the conspiracy content being posted on either Facebook or Instagram).[81] Irene Khan, the current Special Rapporteur, criticised state responses to this, calling it "problematic and heavy-handed"[82] and the role of companies "woefully inadequate"[83]. She stated that disinformation can, for instance, be restricted through laws that prohibit false information relating to electoral integrity, but these restrictions need to follow a high threshold.[84]

As social media platforms are operated by private companies, the United Nations Business and Human Rights Framework shows that while the state has the primary responsibility to protect human rights, private businesses have some responsibility as well. According to the Guiding Principles on Business and Human Rights, *every* business has the responsibility to respect human rights.[85] They rest on three pillars: The duty of the state to protect, businesses have a corporate responsibility to protect, and victims need to have access to remedy. Principle 13, for instance, states that businesses should "avoid causing or contributing to adverse human rights impacts" and "seek to prevent or mitigate adverse human rights impacts that are directly linked to their operations"[86]. Businesses need to commit themselves to due diligence in principles 17-21 to assess potential negative human rights impacts. They should be transparent through periodic reports (principle 21) and provide remedies (principle 22). Yet, the principles are non-binding. Thus, enforcement remains a challenge.

Regarding content regulation and moderation, Special Rapporteur Kaye warned that fast and automated content removals written in law result in "new forms of prior restraint"[87]. He emphasized the need for private companies to not make decisions regarding FoE online, as these questions are very complex and primarily a question of law.[88] Furthermore,

---

[80] Rachel Leigh Greenspan and Elizabeth F. Loftus, 'Pandemics and Infodemics: Research on the Effects of Misinformation on Memory', *Human Behavior and Emerging Technologies* 3, no. 1 (January 2021): 8.
[81] Rory Smith, Seb Cubbon, and Claire Wardle, 'Under the Surface: Covid-19 Vaccine Narratives, Misinformation and Data Deficits on Social Media. Executive Summary', *First Draft*, 2020, 10.
[82] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/HRC/47/25', para. 3.
[83] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, para. 3.
[84] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, para. 42.
[85] Endorsed by the Human Rights Council in 2011. High Commissioner of Human Rights and United Nations, 'Guiding Principles on Business and Human Rights', HR/PUB/11/04 § (2011).
[86] High Commissioner of Human Rights and United Nations, principle 13.
[87] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/HRC/38/35', para. 17.
[88] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, para. 17.

intermediaries often lack the means to assess illegal content (see chapter 2 section 2). And although states receive guidelines to prohibit the dissemination of racist ideas and racist expression, there is no technical analysis of themes such as thresholds and delineations between potentially conflicting freedoms such as expression and non-discrimination.[89] With the UN regulations, which are very vague and do not clearly define what is a legitimate restriction and what is not – an issue most laws regarding FoE have – it is possible to protect vulnerable groups, namely children, from seeing harmful content online, or to go against misinformation. Yet, many criticise the possibility of overcensoring under this framework. The next section will outline how European Human Rights law protects FoE.

2. The European Framework

The last section outlined Freedom of Expression in the international human rights framework. This next section will explore the European Convention of Human Rights (ECHR). Article 10 ECHR protects Freedom of Expression and was drafted before the ICCPR. The article further similarly states that Freedom of Expression "carries with it duties and responsibilities"[90] like Article 19 ICCPR. Overall, it is narrower than Article 19 and contains nine legitimate restrictions of FoE stated in 10(2) which also need to follow the three-step test.

Regarding Freedom of Expression online, the European Court of Human Rights (ECtHR or "the Court") held that the internet has an especially important role for FoE[91] whereby audio-visual media have an even more powerful effect than traditional media as it is more immediate.[92] The court further has some case law relevant to social media platforms and their users (see also appendix 1). The ECtHR found in Delfi AS v. Estonia that governments can impose liability on online news outlets if they fail to remove "clearly unlawful" [93] comments, even without notice. This judgment showed tensions between the court and EU legislation (E-Commerce Directive, see chapter 2 section 1) as the Directive prohibits general monitoring and the removal of unlawful comments without notice requires exactly this. The court further stated that the demand for effective measures to take down illegal content is by no means "private censorship"[94]. Moreover, regarding the obligations of individuals, in the recent case

[89] Natalie Alkiviadou, 'The Legal Regulation of Hate Speech - The United Nations Framework as the Common Denominator for Europe and Asia', *European-Asian Journal of Law and Governance*, 2018, 29.
[90] Council of Europe, 'European Convention on Human Rights' (1950), Article 10(2).
[91] E.g., in Ahmet Yildirim v Turkey, No. 3111/10 (ECtHR 18 December 2012) para 49; Times Newspapers Ltd v. the United Kingdom, No. 3002/03, 23676/03 (ECtHR 10 March 2009) para 27; Ooo Informatsionnoye Agentstvo Tambov-Inform v. Russia, No. 43351/12 (ECtHR 18 May 2021) para 84.
[92] Delfi AS v. Estonia, No. 64569/09 (ECtHR 16 June 2015) para 134.
[93] Delfi was seen as the publisher or discloser of the comments and thus, could be held accountable, see Delfi AS v. Estonia, para 159.
[94] Delfi AS v. Estonia para 157.

of Sanchez v. France, the ECtHR ruled that a (Facebook) account holder needs to take down unlawful (in this case, antisemitic) comments.[95] The dissenting judge warned that this judgement created a "very heavy burden on the account holder in terms of monitoring posts since he or she would face criminal prosecution"[96]. The judge feared that this could lead to censorship as the account holder will be more included to protect themselves and err on the side of caution. He further warned that "the chilling effect is self-evident, thus entailing a serious threat to freedom of expression"[97]. Simultaneously, the court states that it is not the role of the judiciary to rewrite history "by ordering the removal […] of all traces of publications"[98] in cases of attack against individual reputations.

Political expression[99] (not only when there are political issues or crimes concerned, but also e.g., sporting issues[100]) or contributions to a debate of public interest[101] are examples of particularly protected speech. In cases of such, the Member States have a narrower margin of appreciation. Protected are also ideas that offend, shock, or disturb. "such are the demands of pluralism, tolerance, and broadmindedness without which there is no democratic society"[102]

Yet, Freedom of Expression is not an absolute right and can be restricted. Interferences with Freedom of Expression are, for instance, the arrest of protestors[103], the prohibition to publish information[104], the seizure of information[105] or the conviction of the owner of a Facebook account for illegal comments posted on the page by third parties[106]. In cases of an interference, the Court follows the three-step test. In the case of FoE on the internet, often two rights conflict, such as with the Respect for Private Life (Article 8). The Court stated that there is no hierarchical relationship between Article 8 and 10[107] and if expression targets, for instance, group identity[108], such as negative stereotyping, it can affect groups' senses of identity, self-worth, and confidence.[109] Under such circumstances, FoE can be restricted.

[95] Sanchez v. France, No. 45581/15 (ECtHR 2 September 2021) para 104.
[96] Sanchez v. France Dissenting Opinion of Judge Mourou-Vikström.
[97] Sanchez v. France Dissenting Opinion of Judge Mourou-Vikström.
[98] Węgrzynowski and Smolczewski v. Poland, No. 33846/07 (ECtHR 16 July 2013) para 65.
[99] Perinçek v. Switzerland, No. 27510/08 (ECtHR 15 October 2015) para 197; Ceylan v. Turkey, No. 23556/94 (ECtHR 8 July 1999) para 34.
[100] Axel Springer Ag v. Germany, No. 39954/08 (ECtHR 7 February 2012) para 90.
[101] Perinçek v. Switzerland at 197; Ceylan v. Turkey para 34; Animal Defenders International v. the United Kingdom, No. 48876/08 (ECtHR 22 April 2013) para 102.
[102] Aksu v. Turkey, No. 4149/04, 41029/04 (ECtHR 15 March 2012) para 64.
[103] Steel and Others v. The United Kingdom, No. 24838/94 (ECtHR 23 September 1998) para 92.
[104] Cumhuriyet Vakfı and Others v. Turkey, No. 28255/07 (ECtHR 8 January 2014) para 46.
[105] Handyside v. the United Kingdom, No. 5493/72 (ECtHR 7 December 1976) paras 44f.
[106] Sanchez v. France para 68.
[107] Timciuc v. Romania, No. 28999/03 (ECtHR 12 October 2010) para 144.
[108] Aksu v. Turkey para 75.
[109] Perinçek v. Switzerland para 200.

Hate Speech is another important restriction of FoE. The ECHR has no equivalent to Article 20(2) ICCPR and thus, there is no explicit obligation for states to prohibit hate speech. The ECtHR usually uses Article 10(2) or 17 (prohibition of abuse of rights) to tackle hate speech. The Council of Europe defined hate speech in more detail as "all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including intolerant expression by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin."[110] This is a relatively broad concept with a low threshold of what constitutes illegal hate speech. Yet, for instance, homophobic or transphobic speech was not mentioned as hate speech, but it was later confirmed that "discrimination based on sexual orientation is as serious as discrimination based on 'race, origin or colour'"[111]. Regarding more specifically which speech is deemed as illegal, the Court stated that "expressions that seek to spread, incite or justify hatred based on intolerance"[112] violate FoE. The assessment of hate speech is very contextual and depends on the circumstances of the case[113][114], the group which the speech addresses[115], the receptors of the speech[116], the manner[117] and the content of the statements (e.g., a direct or indirect call for violence, hatred or intolerance[118]; the advocating for the intensification of armed struggle or glorifying war[119].; or Holocaust denial[120]).

Overall, the European framework relies on much case-law and protects Freedom of Expression extensively. The legitimate restrictions are decided upon context and hardly any generalizations can be made about what content always amounts to hate speech (with Holocaust denial possibly coming closest to a priori hate speech). The next section will briefly discuss the two frameworks.

---

[110] Committee of Ministers, 'Recommendation No. R. (7) 20 of the Committee of Ministers to Member States on "Hate Speech"' (Council of Europe, October 1997), 106.
[111] Vejdeland and Others v Sweden, No. 1813/07 (ECtHR 9 May 2012) para 55.
[112] Gündüz v Turkey, No. 35071/97 (ECtHR 14 June 2004) para 51.
[113] E.g. whether the statements were made during a tense political or social background, see Perinçek v. Switzerland para 205.
[114] E.g., during the clashed between the PKK and the Turkish Security Forces, see Sürek v. Turkey (no. 1), No. 26682/95 (ECtHR 8 July 1999) paras 52,62.
[115] Especially when targeting ethnic, religious or other groups (e.g., leaflets claiming that homosexuality was the reason for the spread of HIV and AIDS where the court stressed that "discrimination based on sexual orientation is as serious as discrimination based on race, origin or colour", see Vejdeland and Others v Sweden para 55.
[116] E.g., younger and more impressionable people, see Vejdeland and Others v Sweden para 56.
[117] E.g., through a poem rather than mass media, see Karatas v. Turkey, No. 23168/94 (ECtHR 8 July 1999).
[118] Perinçek v. Switzerland para 206.
[119] Ozgur Gundem v. Turkey, No. 23144/93 (ECtHR 16 March 2000) para 65.
[120] X v Federal Republic of Germany, No. 9235/81 (ECtHR 16 July 1982) para 5; Peta Deutschland v. Germany, No. 43481/09 (ECtHR 8 November 2012) para 49.

3. Discussion: The Protection of Online Expression

As mentioned above, Freedom of Expression is a fundamental human right. Yet, as the previous sections showed, if expression contains speech against protected groups or violates the rights of others, it can legally be restricted. This balancing is very complex, and many factors need to be considered. As clear from the case-law, states must make sure that effective measures are in place so that the internet is now a law-free space where everything can be published without any oversight. Expression online is protected extensively. In other words, the protection of Freedom of Expression is important but so is its restriction in certain cases. Thus, it is necessary for restrictions of FoE need to follow the three-step test transparently. The restrictions to be provided by law and policies and internal regulations of social media platforms must be formulated in a way that users know exactly what they are allowed to post and what not. Additionally, they must be publicly accessible.

This process is important as, for instance, the targeted harassment of individuals and groups can also have censoring effects.[121] Users have the right to not encounter hate speech and to not have their personality rights violated. Thus, no content moderation can similarly restrict expression. Social media needs to be regulated. History has shown how much harm can be done if there are no regulations, it can seriously undermine democratic processes.

The obligation of states to prevent media monopoles, be it governmental or private, does not reflect reality as well. The accusation of the big companies behind social media (namely, Google and Meta) being monopolistic is not a stretch.[122] Only a few companies have power over the global Freedom of Expression by creating platforms where users can post what they want.

Thus, at first glance, the premises of the UN and EU sound reasonable and realistic. However, within the context of social media, they sound slightly idealistic. The fact that free speech can only be limited through the three-step under the assessment of an independent judge is fundamental but unfortunately, in practice often disregarded. In reality, private monopolistic companies make primarily quick and automated decisions about the blocking of information online (see chapter 2). Thus, the following section will discuss this challenge more concretely by outlining the framework of the EU and Europe and its attempts to implement social media regulations more concretely while respecting Freedom of Expression.

All in all, the legal framework shows that Freedom of Expression is a fundamental human right which can be restricted to protect others' rights among other things. To understand the

[121] Jillian C. York and Ethan Zuckerman, 'Moderating the Public Sphere', in *Human Rights in the Age of Platforms*, ed. Rikke Frank Jørgensen (Massachusetts: MIT, 2019), 154.
[122] E.g., Wagner, 'Free Expression? Dominant Information Intermediaries as Arbiters of Internet Speech', 220.

scope of the law and when it is legitimate to restrict it, a lot of contexts is needed and clear definitions.

This chapter established the provisions regarding Freedom of Expression of the UN and Europe. It highlighted under which conditions Freedom of Expression can be restricted. Accordingly, the next chapter will now outline how the EU more concretely deals with the regulation of social media considering balancing Freedom of Expression and how social media platform moderate user-generated content.

Chapter II: Platform Governance in the European Union

The mid-2010s can be identified as a period in which there was a feeling of urgency for more content regulation and at the same time, populism and misinformation were on the rise. Our current laws developed in this period. The EU's framework regarding content moderation has often been deemed as a "third way" between the digital authoritarianism of China and the relatively extreme unrestricted approach of the US.[123]

This chapter is going to explore this to answer the second subquestion of this thesis (II 'How are social media platforms governed?') by examining EU policies regarding the regulation of social media platforms and subsequently, the platform's self-regulation practices, following a more socio-legal approach.

To clarify, this thesis will refer to content regulation as state initiatives to remove content and content moderation (or self-governance) as the enforcement by the companies. The difference is that content regulation deals with illegal content under either criminal law or legal restrictions of Freedom of Expression while content moderation goes beyond the illegal and restricts legal speech according to the companies' Terms of Service.

1. The European Union's Social Media Regulation

The following section is about the governance *of* platforms. It aims to critically reflect on the current information environment by analysing the system of content regulation and moderation in light of FoE. The EU's main regulation is the E-Commerce Directive of 2000. For this section, the case-law[124] of the Court of Justice of the European Union (ECJ, also "the Court" in this chapter) (see also appendix 1) and the EU's approach to fight Hate Speech will also be drawn upon.

In the European Union, the Charter of Fundamental Rights of the European Union protects Freedom of Expression with Article 11.[125] It only applies against legal acts of the EU, corresponds with the scope of Article 10 ECHR[126] and thus, also applies online[127]. Regarding FoE restrictions, the EU has the non-legally binding 2016 Code of Conduct on Illegal Hate Speech Online. The Code emphasises that offensive, shocking or disturbing speech is

[123] The Freedom House, 'Freedom on the Net 2021' (2021), 15.
[124] For this, case-law primarily about the E-Commerce Directive was researched through keywords such as "Directive 2000/31/EC AND Freedom of Expression".
[125] 'Charter of Fundamental Rights of the European Union', 2000/C 364/01 § (2000).
[126] Article 52(3) states that both rights have the same scope. Charter of Fundamental Rights of the European Union.
[127] Republic of Poland v European Parliament and Council of the European Union, No. C-401/19 (ECJ 26 April 2022).

included in the right to FoE and therefore, protected.[128] The Code defines hate speech as "all conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin"[129]. Certain intermediaries (Facebook, Microsoft, Twitter, YouTube, Instagram, Snapchat, Dailymotion, Jeuxvideo.com, TikTok and most recently, LinkedIn)[130] agreed to take down flagged content in less than 24 hours upon notification. The companies also need to have "effective processes to review notifications regarding illegal hate speech" in place "to disable access to such content"[131]. The wording seems to also steer towards automated responses (besides human intervention) such as filtering which can interfere extensively with FoE (see below). The Code also states that intermediaries need to "develop counter-narratives"[132]. Regarding the latter, this can lead to platforms becoming "carriers of propaganda"[133]. The short time frame in combination with this, can seriously undermine FoE due to Hate Speech prevention.

Regarding regulations relevant to social media platforms, the European Union E-Commerce Directive[134] of 2000 determines currently how social media platforms are supposed to act. Intermediaries are held accountable under Article 12 for "mere conduct", 13 for "caching" and 14 for "hosting". The Directive explicitly states that content takedown needs to be undertaken "in the observance of Freedom of Expression"[135] whereby they refer to Article 10(1) ECHR. It does not refer to hate speech or explicitly outlined what content is illegal and can be taken down. It further prohibited the Member States from imposing monitoring obligations on intermediaries with Article 15.[136]

Article 14(1) of the Directive lifts the liability of intermediaries if they are not having knowledge of illegal activity and act "expeditiously"[137] by removing or disabling the access to illegal content. The ECJ further explained that "knowledge" in this case does *not* mean a

[128] European Commission, 'The EU Code of Conduct on Countering Illegal Hate Speech Online' (2016), 1.
[129] European Commission, 1.
[130] European Commission, 3.
[131] European Commission, 2.
[132] European Commission, 1.
[133] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/HRC/38/35', para. 21.
[134] 'Directive on Certain Legal Aspects of Information Society Services, in Particular Electronic Commerce, in the Internal Market (Directive on Electronic Commerce)', 2000/31/EC Directive § (2000).
[135] Directive on Certain Legal Aspects of Information Society Services, in particular Electronic Commerce, in the Internal Market (Directive on Electronic Commerce), para. 46.
[136] Article 19, 'Germany: Responding to "Hate Speech". 2018 Country Report' (London, 2018); Directive on Certain Legal Aspects of Information Society Services, in particular Electronic Commerce, in the Internal Market (Directive on Electronic Commerce), Article 47.
[137] European Parliament and European Council, 'Directive 2000/31/EC on Certain Legal Aspects of Information Society Services, in Particular Electronic Commerce, in the Internal Market (Directive on Electronic Commerce)', 2000/31/EC § (2000), Article 14(1)(b).

general awareness, it must be based on a notification (i.e., notice and action mechanisms).[138] Regarding the time required to act, the Directive held that intermediaries are not responsible when it acts *quickly* (i.e., remove content upon notification) without indicating a timeframe of said action.

The Directive further allows courts or administrative authorities to require platforms to prevent an infringement (following Member State law) in Article 14(3). Regarding said preventive actions by intermediaries, the ECJ held in Scarlet Extended[139], Belgische Vereeniging[140] and L'Oréal[141] that provisions about preventive measures can lead to automatic filtering which, if that cannot distinguish adequately between legal and illegal speech, it will not sufficiently balance FoE with other fundamental rights. Overall, the Directive creates distance of governments to intermediaries and leads to the primary reliance of self-governance of social media platforms (see section 2 of this chapter).

More recently, the ECJ softened its stance. In a case regarding copyright, Poland argued similarly to the cases mentioned above, that if online content-sharing service providers (i.e., platforms) need to "carry out preventive monitoring"[142], they will use automatic filtering which leads to significant limitations of Article 11 of the Charter. The Court agreed and stated that prior review and prior filtering can restrict Article 11.[143] It held that it is up to the platforms to decide on what measures they want to take in order to respect the rights of the Charter and thus, measures can be necessary when balancing with other rights and proportionate (which is in the duty of the state to ensure).[144] The Advocate General noted that filtering is the logical consequence of provisions asking for preventive measures.[145] The measures must at all times follow the principle of proportionality. He states that the filters are necessary in many cases of copyright infringement.[146] Poland argued that said provision about preventive measures entails the introduction of "general and automated preventive censorship"[147]. The Court furthermore commented that preventive monitoring in the form of

[138] Google LLC, YouTube Inc., YouTube LLC, Google Germany GmbH and Elsevier Inc. v Cyando AG, No. C-682/18 (ECJ 22 June 2021) Paras 35, 118.; Republic of Poland v European Parliament and Council of the European Union, No. C-401/19 (ECJ 26 April 2022), para 50.
[139] In this case, filters could not balance FoE, the right of protection of personal data, the freedom to conduct business adequately with the right to intellectual property. Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM), No. C-70/10 (ECJ 24 November 2011), para 53.
[140] Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV, No. C-360/10 (ECJ 16 February 2012) para 51.
[141] L'Oréal SA and Others v eBay International AG and Others, No. 324/09 (ECJ 12 July 2011) para 139.
[142] Republic of Poland v European Parliament and Council of the European Union para 24.
[143] Republic of Poland v European Parliament and Council of the European Union para 55.
[144] Republic of Poland v European Parliament and Council of the European Union paras 75, 82–84.
[145] Opinion of Advocate General Saugmandsgaard, No. C-401/19 (ECJ 15 July 2021) para 64.
[146] Opinion of Advocate General Saugmandsgaard para 69.
[147] Opinion of Advocate General Saugmandsgaard para 75.

filters would require monitoring and thus, violate Article 15(1) of the Directive.[148] The Advocate General concluded that the provision is presumed to be lawful, but the technological means are the issue.[149] Overall, the ECJ has been saying since 2011 - and continuously repeating - that (1) filtering requires prohibited monitoring[150] and (2) that filtering can violate freedom of expression[151]. Thus, judgements by the ECJ such as Scarlet Extended or Netlog suggest that proactive or preventive measures of intermediaries can conflict with Article 15 of the Directive.[152] In the Scarlet Extended case, the ECJ stated that filtering, as a measure often taken by social media platforms to moderate content is (dangerous), it can violate Freedom of Expression, among other rights, when the filtering information cannot adequately assess whether content is lawful or not.[153]

All in all, the E-Commerce Directive gives a lot of room to intermediaries and little liability over the content published on their platforms. Due to this, the EU introduced more communications that steadily increased intermediary responsibility. Intermediaries were, for instance, "encouraged"[154] to take measures against illegal content. The Code has furthermore been accused of leading to the privatization of FoE enforcement.[155] Moreover, the time frame of 24 hours can be very restricting and combined with unclear definitions can certainly lead to overblocking or even censorship-like practices. The Directive gave Member States a lot of freedom, leading to widely different laws, from more restricted ones such as the German law. Thus, under the EU framework, more restrictive media laws can be adopted. Germany, for instance, has the most far-reaching law. The "Netzwerkdurchsetzungsgesetz" (NetzDG, can be translated to "network enforcement act") poses substantial penalties in case intermediaries fail to respond appropriately to reported content within 24 hours or under certain circumstances, after 7 days or more.[156][157] The companies are also obligated to report

[148] Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV para 38.
[149] Opinion of Advocate General Saugmandsgaard para 208f.
[150] Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM) para 40.
[151] Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM) paras 50, 52 ; Republic of Poland v European Parliament and Council of the European Union para 55.
[152] Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV; Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM).
[153] Thomas Riis and Sebastian Felix Schwemer, 'Leaving the European Safe Harbor, Sailing Towards Algorithmic Content Regulation', *SSRN Electronic Journal* 22, no. 7 (2018): 12.
[154] European Commission, 'Commission Recommendation on Measures to Effectively Tackle Illegal Content Online', 2018/334 § (2018), Chapter II, para. 18.
[155] Eugénie Coche, 'Privatised Enforcement and the Right to Freedom of Expression in a World Confronted with Terrorism Propaganda Online', *Internet Policy Review* 7, no. 4 (November 2018): 11f.
[156] 'Gesetz Zur Verbesserung Der Rechtsdurchsetzung in Sozialen Netzwerken (Netzwerkdurchsetzungsgesetz - NetzDG)' (2017), para. 8.
[157] In the case of YouTube, 87,5% of reported posts were assessed and processed within 24 hours; 43.847 complaints in total, 38.368 posts within 24 hours. Google, 'Entfernungen von Inhalten Nach Dem Netzwerkdurchsetzungsgesetz – Google Transparenzbericht'.

periodically.[158] In general, the reports through NetzDG are progress toward making online content deletion and reporting more transparent. Yet, overall, the process is still very opaque.[159] In the reports, no example cases are shown only broadly the reason for the takedown. They do, however, show that the complaint mechanisms are widely used due to the high numbers. The NGO Article 19 calls the sanctions under NetzDG "severe"[160] and criticises that the response from social media platforms is restrictive and that they have been "over-zealous in removing content"[161]. They further argue that some provisions of the criminal code´, such as Article 185 StGB (insult) and Article 187 StGB (defamation) do not comply with International Human Rights Law.[162]

Yet, the core principles of the Directive are still relevant today and in general, the Directive was received positively. However, it did not give any cooperation mechanism between Member States, leading to legal uncertainties and different degrees of protection for Freedom of Expression and the fear of legal fragmentation.

Thus, the EU relied on a more laissez-faire approach, giving the Member States a wide margin of appreciation for their social media laws. This led to very diverse laws, creating confusion among social media users. Yet, over the last years, the EU increased its regulatory power on the intermediaries, leading to the new Digital Services Act which will amend the Directive and aims to advance EU content regulation (see discussion, section 1).

This section explored the EU approach to regulating social media. For a comprehensive analysis, it is necessary to further highlight how social media platforms implement these regulations and how they overall moderate user-generated content. The following will outline this.

---

[158] See section 2 for extracts from such reports. Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz - NetzDG), Article 1 para. 2.

[159] For instance, Twitter does indicate based on what provision of the German criminal code the measure was taken (e.g., incitement of masses, German: §130 StGB Volksverhetzung, 228.468 complaints with 20.281 measures taken or §185 Insult, German Beleidigung with 171171 complaints and 14.775 measures taken), see Twitter, 'Twitter Netzwerkdurchsetzungsgesetzbericht: Juli - Dezember 2021'.

[160] Article 19, 'Germany: Responding to "Hate Speech". 2018 Country Report', 49.

[161] Article 19, 49.

[162] The Special Rapporteur explicitly stated that speech is allowed to offend. Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/HRC/38/35', para. 26.

2. Content Moderation of Social Media Platforms

The omnipresent worry that state requests and the response of private companies to them can lead to constraints on information access and exchange, limiting the work of journalists and discouraging whistle-blowers and human rights defenders has been voiced frequently.[163] Social media platforms have different interests than states. Overall, they prioritise financial and commercial benefits over human rights (such as selling data and placing advertisements). After previously discussing the stateside, it is important to highlight the possibilities of private companies for implementing regulations and the effects of these measures on Freedom of Expression. Thus, this section is about governance *by* platforms. It tries to uncover the "black box"[164] of content moderation. It will critically analyse the implications of the current framework established in the section before through exploring the more technological side of platform governance.

In the following section, the publicly available Terms of Service (ToS), guidelines and rules of popular social media platforms in Europe will be analysed and compared. Some of the most used social media platforms are Facebook, YouTube, Instagram, TikTok and Twitter.[165] Platforms which do not have a newsfeed, such as direct messengers, were excluded from this analysis due to their content being less curated and less susceptible to content moderation.[166] Thus, Meta (which owns Facebook and Instagram), Google (which owns YouTube), Twitter and TikTok were chosen. Their public community guidelines, hate speech policies and moderation methods will be outlined and discussed in the following.

All chosen intermediaries committed themselves to following the ECHR and the Guiding Principles on Business and Human Rights.[167] In general, intermediaries distinguish between external requests for content removal through the law or the state and internal removal requests through their ToS.[168] ToS are contracts that users need to agree to for participating in the social media platforms. They often contain the companies' policies on what content can be

[163] E.g., Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/HRC/35/22' (Human Rights Council, March 2017).
[164] A criticism of non-transparency, see Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge: Harvard University Press, 2015).
[165] Kemp, 'Digital 2022', slide 123.
[166] E.g., WhatsApp, WeChat, or Snapchat were excluded.
[167] Twitter, 'Twitter's Free Speech and Rights of People'; Miranda Sissons, 'Our Commitment to Human Rights', Meta, March 2021; 'Transparency Center Homepage', TikTok; 'About Human Rights at Google', Google.
[168] Often adopted unilaterally by the companies themselves without any interference from the outside. Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/HRC/38/35', para. 19.

published and what not, often formulated very generally.[169] ToS are more general while community guidelines and policies are more specific and aim to, for instance, target hate speech[170] or misinformation[171]. The agreement of the user has two effects: First, users receive the right to access and publish content, regulate some levels of personalization, identify jurisdictions for dispute resolution and use all other functions of the platform.[172] Second, most of the power is transferred onto the intermediaries regarding user-generated content (UGC). It should be noted that these terms and guidelines are only the tip of the iceberg, content moderation follows many internal (non-public) rules.[173]

Community standards (or guidelines) are part of the ToS and determine what content is allowed on their platform. All chosen intermediaries mention the topic of hate speech or hate full behaviour.[174] They define it themselves and specify groups which are deemed as especially sensitive for the detection of hate speech but newsworthy or content of public interest are sometimes deemed as exceptions (see also appendix 2). Thus, these definitions are usually broader than the ones by governments as they include more groups and cover a wider area. Meta, for instance, defines hate speech as direct attacks against people based on "protected characteristics"[175]. The approach to hate speech is broadly the same between the platforms. Yet, for instance, Twitter provides examples of hate speech and has in general much information about what is allowed and what is not, and the specification of included groups differs slightly.[176]. Meta has a list which ranks content from tier 1 (e.g., generalizations of inferiority) to tier 3 (e.g., calls for action regarding segregation). The detection of violating content is done through a combination of technology or review teams which will be explored in more detail in the following section.[177]

---

[169] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression', May 2016, para. 52; Mchangama, Alkiviadou, and Mendiratta, 'A Framework of First Reference: Decoding a Human Rights Approach to Content Moderation in the Era of "Platformization"', 11.
[170] E.g., 'Hate Speech. Transparency Center'.
[171] E.g., Yoel Roth, 'Introducing Our Crisis Misinformation Policy', 2022.
[172] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/HRC/38/35', para. 24.
[173] For instance, there have been many leaks about guidelines content moderators need to follow, for instance in the case of TikTok, see for instance, Alex Hern, 'Revealed: How TikTok Censors Videos That Do Not Please Beijing', *The Guardian*, September 2019.
[174] 'Hate Speech. Transparency Center'; Twitter, 'Twitter's Policy on Hateful Conduct', Help Center; 'YouTube Hate Speech and Harassment Policy', YouTube; Eric Han, 'Countering Hate on TikTok', TikTok, August 2019.
[175] Which are based on race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender, gender identity, or serious disability or disease. Richard Allan, 'Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community?', *Meta*, June 2017.
[176] YouTube, for instance, defined hate speech as incitement of hatred or violence against 13 characteristics including "victims of a major violent event and their kin" or "veteran status", see 'Hate Speech Policy - YouTube Help', YouTube.
[177] E.g., 'Detecting Violations', Transparency Center; 'How Does YouTube Identify Content That Violates the Community Guidelines?', YouTube Community Guidelines.

Before turning towards different forms of control of social media companies, it is important to explain the most important automated and manual methods used to identify content that could possibly violate ToS.

Algorithms are the automated side of content moderation. They are primarily ex-ante (before publication) "encoded procedures for transforming input data into the desired output, based on specified calculations"[178]. They can also select and exclude (i.e., filter) content according to the users' interests and habits (e.g., TikTok bases recommendations on the likes or shares of users, the accounts they follow, their comments or their posted content)[179]. They can thus regulate what information users receive. They furthermore rank content, with the most relevant posts showing up higher than less relevant ones (e.g., TikTok assess the popularity of content according to captions, sounds or hashtags of videos[180]).[181] One popular content analysing algorithm is 'hashing'. Hashing is used to identify controversial or forbidden images.[182] After manually removing forbidden images, the companies store the digital signature called "hash" in a database – allowing the prevention of the uploading of these images again.[183] A hash is created by changing the pictures to back and white, overlaying them with a grid and then assigning numerical values.[184]

Flagging (also often referred to as reporting) is an ex-post moderation (after publication) method through which users can express their disliking of content through reporting posts. Usually, after content is reported, it will be reviewed by the platforms, primarily through content moderators.[185] This is a widely used practice; According to Google's NetzDG report, 263.663 posts were reported within five months[186] based on reasons such as violations of privacy (11.056), pornography (50.670), hate speech (64.815) or violence (34.123) (see appendix 2 for more information about content take down due to reporting).[187] All platforms in this analysis offer a reporting or flagging mechanism but Google's "YouTube Trusted

---

[178] Leyla Dogruel, Dominique Facciorusso, and Birgit Stark, '"I'm Still the Master of the Machine." Internet Users' Awareness of Algorithmic Decision-Making and Their Perception of Its Effect on Their Autonomy', *Information, Communication & Society*, December 2020, 2.
[179] 'How TikTok Recommends Videos #ForYou', TikTok, August 2019.
[180] 'How TikTok Recommends Videos #ForYou'.
[181] Pascal Jürgens and Birgit Stark, 'The Power of Default on Reddit: A General Model to Measure the Influence of Information Intermediaries: The Influence of Information Intermediaries', *Policy & Internet* 9, no. 4 (2017): 395–419.
[182] An example of such a hashing algorithm is PhotoDNA, see Tracy Ith, 'Microsoft's PhotoDNA: Protecting Children and Businesses in the Cloud', Microsoft, July 2015.
[183] Jørgensen, *Human Rights in the Age of Platforms*, 150.
[184] Kate Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech', *Harvard Law Review* 131, no. 1598 (2018): 1636f.
[185] On how exactly decisions are made during this review process can only be speculated.
[186] between July 2021 and December 2021
[187] 43.847 contents were deleted, most of them due to the accusation of hate speech (14.759). Google, 'Entfernungen von Inhalten Nach Dem Netzwerkdurchsetzungsgesetz – Google Transparenzbericht'.

Flagger Programme"[188] is more advanced. The programme gives NGOs and government agencies more authority through which their reporting of content is prioritized (and thus, receives quicker action) and they receive more insight into decisions about content. In other words, the flaggers are trusted and thus, their reports are taken more seriously which allows for quicker decisions. Flagging is a "complex interplay between users and platforms, humans and algorithms, and the social norms and regulatory structures of social media"[189]. Often, groups systematically report content as a platform violation just to have this content suppressed, even though, in reality, the content just expressed an opinion that was not to their liking.[190] Yet, reporting is a method that gives users the power to determine what content they want to see even though the final decision lies within the intermediaries. Still, reporting is likely the most democratic way of moderating content, even if flawed.

Content moderators are the other widespread ex-post moderation mechanism. Here, people are employed to review online content which has previously been identified as violating the ToS. In the case of Meta, not all moderators review the same content. There are three different classifications; 'Tier 3' is for day-to-day reviewing, 'tier 2' supervises 'tier 3' and reviews prioritized content, and 'tier 1' is made up of lawyers or policymakers.[191][192] Moreover, Meta has the so-called Facebook Oversight Board for Facebook and Instagram which reviews cases regarding content moderation through an appeal mechanism.[193] The board is made up of 40 "experts and civic leaders"[194] from all around the world. Until June 2022, it has made 26 decisions. It is currently thus too early to assess whether this was successful or not, but it seems promising. Further, warning labels are a promising compromise between taking down posts and leaving them online. This is an adequate response to mis- or disinformation, even though some people argue that the trust in social media platforms is so low that users will not trust the warnings. In the case of hate speech, for instance, it is more complex. Instagram, for instance, hides violent content behind such warnings, which at least warns users (and especially children) about the content they will see.

The enforcement is primarily done on the post level and the account level (see appendix 2). The reaction to forbidden or undesired content ranges from labelling (e.g., as misinformation)

---

[188] 'About the YouTube Trusted Flagger Programme', YouTube Help.
[189] Kate Crawford and Tarleton Gillespie, 'What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint', *New Media & Society* 18, no. 3 (2016): 411.
[190] Jørgensen, *Human Rights in the Age of Platforms*, 151.
[191] For more detailed descriptions, see Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech', 1640.
[192] Currently, Facebook employs 15.000 moderators, with critics arguing that they need double the amount. Charlotte Jee, 'Facebook Needs 30,000 of Its Own Content Moderators, Says a New Report', MIT Technology Review, June 2020.
[193] 'Oversight Board'.
[194] 'Meet the Board', Oversight Board.

to the permanent deletion of accounts and content. For instance, Twitter's enforcement is on the tweet-level (i.e., labelling, limiting visibility etc.), message-level (i.e., stopping the sending of messages) or account-level (i.e., temporary restrictions, verification, suspension, etc.).[195] If tweets are not violating rules, they might still be flagged with a notice, age-restricted or geoblocked. Exceptions might be made if tweets are of public interest. Other measures are giving out warnings or strikes, downranking "forbidden" posts or highlighting favoured posts.[196] The severity of the consequences is sometimes dependent on the content (e.g., Meta's tier system) or the number of previous strikes or warnings (e.g., YouTube). Many intermediaries also prohibit false news in their ToS. As a reaction, they add warning labels, suppress the posts, or inform the user about the possibility of deletion.

All in all, these content moderating measures are an essential part of social media platforms. Overall, these methods are used to exercise either *hard* or *soft* control; Hard control is the assessment of content on whether it is acceptable to be published on the platform while soft control is the determination of what content is recommended or restricted.[197] The following will show in more detail what platforms can do with content moderating methods and what effects this can have on users.

2.1. Hard Control of Online Content

As mentioned above, hard control means platforms remove content, block content in specific countries or deactivate or limit accounts and thereby, exercise immense control on online information. In general, the rate of content takedown differs widely between the platforms. According to the NetzDG reports, the takedown rate of content that falls, for instance, under Article 130 StGB (incitement of masses) ranges from under 10 % to over 20 (see appendix 2). Hard control is the obvious (or "overt", see chapter 3) method to moderate content. The four intermediaries of this analysis rely both on technology and user reporting to detect violating content.[198] Most of the time, if the content is found to be violating ToS through hashing, keyword filters, flagging or algorithms, they are withheld or temporarily deleted. Afterwards, they can be reviewed by content moderators or appealed (see appendix 2). TikTok offers a simple appeal process by clicking the button 'submit an appeal' without any further information about the process which follows. Facebook and Instagram users, for instance, have more options than this. They can 'disagree with decision' or request a review. The

[195] 'Our Range of Enforcement Options for Violations', Twitter Help.
[196] Adrian Rauchfleisch and Jonas Kaiser, 'Deplatforming the Far-Right: An Analysis of YouTube and BitChute', *SSRN Electronic Journal*, 2021, 4.
[197] York and Zuckerman, 'Moderating the Public Sphere', 140.
[198] E.g., 'A Safer Twitter'.

former is merely to show Meta that the user disagrees with the deletion of a post but does not lead to human review.[199] The latter usually leads to a review within 24 hours.[200] Afterwards, the Meta user still can send an appeal to the oversight board. Thus, most decisions can be appealed against, but time-sensitive information will be suppressed until further review which can hinder the work of groups such as activists or journalists immensely.

Some automated ex-ante moderation techniques are part of hard control, such as geoblocking or hashing. Meta[201], Twitter[202] and Google[203] use algorithms to detect violating content, and TikTok assumingly as well[204] (see appendix 2). Geoblocking prevents the upload of content according to the location of the end-user, officially used by Google and Twitter.[205] Thus, social media platforms are having the power to block content country by country. YouTube furthermore restrict content regarding the age of the user.[206]

In 2021, internal Meta documents called "Facebook Files" were leaked. They revealed that the algorithmic detection only recognizes 2 to 5 % of hate speech. Francis Haugen, the whistleblower, said that of those only 1 % are cases of violence and incitement in reality.[207] Regarding this, a study in Denmark found that 1.4 % of Facebook comments were determined as "hateful attacks" based on Facebook's ToS, but in reality, only 0.0066 % could be determined as illegal hate speech under Danish law.[208] Thus, the method is not reliable in actually identifying hate speech, thereby illegal speech is left online and legal speech is likely to be blocked. This is especially interesting as Meta has significantly lower hate speech detection rates than for instance, Google and TikTok (see appendix 2). The detection of high amounts of hate speech is worrying as shows that companies are likely to flag content as hate speech even though it might not contain illegal speech or content that violates community guidelines. Much content blocking is due to faulty algorithms (more about this see chapter 2,

---

[199] Recently, the German constitutional court decided that Facebook needs have a provision in its ToS which provides users with the chance to give a statement and receive an explanation regarding the deletion of their accounts. ZR 179/20 (BGH 29 July 2021).
[200] This is, according to Meta, not available for all types of content. 'I Don't Think Facebook Should Have Taken down My Post.', Facebook Help Center.
[201] 97% of detected hate speech was spotted by algorithms Mike Schroepfer, 'Update on Our Progress on AI and Hate Speech Detection', *Meta*, February 2021.
[202] Twitter calls this 'proactive detection'. see 'A Safer Twitter'.
[203] Google uses algorithms to detect content and more specifically, hashes to avoid said content from being re-uploaded. 'YouTube Community Guidelines Enforcement FAQs', Transparency Report Help Center.
[204] TikTok merely indicates that it "uses […] innovative technology […] to identify" violating content. 'Community Guidelines Enforcement Report Jan - Mar 2022', TikTok, June 2022.
[205] See also appendix 2.2., Klonick, 1637; For examples see chapter 3 or Natasha Lomas, 'Twitter Uses Country-Specific Blocking Powers For The First Time To Restrict Neo-Nazi Account In Germany', *TechCrunch*, October 2012.
[206] 'Age-Restricted Content', YouTube Help.
[207] Scott Pelley, 'Whistleblower: Facebook Is Misleading the Public on Progress against Hate Speech, Violence, Misinformation', CBS News, October 2021.
[208] Jacob Mchangama, 'The Real Threat to Social Media Is Europe', *Foreign Policy*.

section 2.2.). Similarly, the effect of mis- or disinformation might be generally overestimated. A 2019 study found that only 1.18% of political tweets of users contained false information and 1% of users (and thereby, voters) accounted for 80% of false information.[209] Currently, there is a high risk of social media companies giving precedence to Artificial Intelligence over post-publication moderation.[210] In general ex-ante measures of hard control are potentially harmful, in particular, if done through erroneous automated means.[211]

Content moderators are the manual side of hard control. The moderators receive content which has been identified through the previously mentioned methods, and they must decide within minutes whether the content violates any rules. The time pressure and the viewing of extreme content can lead to numerous mistakes. The issue with content moderation is its implementation. In the case of TikTok, in 2019, it was expected that moderators check 1000 tickets during an 8-hour shift - giving them around 30 seconds to decide about the content of a 15-second-long video.[212] The content is then divided into four categories: (1) videos that completely violate the platform conditions are deleted, (2) videos are made only visible to the creator, (3) videos are not shown in feeds and (4) where the videos not appearing in feeds or the recommended 'for you' section.[213] Overall, the most extreme consequence is the deplatforming of accounts[214], which is the removal of accounts from social media platforms.

In an interview, anonymous moderators claimed that they are not allowed to talk about their job and the contents they are exposed to every day to anyone.[215] Two interviewees have been cited as saying: "Morally, I find my work to be important. I watch the videos so that other people do not have to watch it."[216] and "the employees are being seen as interchangeable machines that are simply thrown away when they don't function anymore."[217].[218] In May 2020, thousands of moderators joined a class-action lawsuit against

---

[209] Nonetheless, they found that users from the right political spectrum saw and shared significantly more false information. See Nir Grinberg et al., 'Fake News on Twitter during the 2016 U.S. Presidential Election', *Science* 363, no. 6425 (January 2019): 375.
[210] Manuel Ernesto Larrondo and Nicolas Mario Grandi, 'Artificial Intelligence, Algorithms and Freedom of Expression', *Universitas*, no. 34 (February 2021): 173.
[211] As discussed in chapter 2. regarding preventive filtering.
[212] There are three stages of reviewing: The first is done in Barcelona after 50 to 150 video views, the second in Berlin for 8.000 to 15.000 views and the third of videos starting from about 20.000 views. Markus Reuter and Chris Köver, 'TikTok: Cheerfulness and censorship', netzpolitik.org, November 2019.
[213] Reuter and Köver.
[214] A term made popular by right-wing US Americans and has accelerated in popularity in 2018. Rauchfleisch and Kaiser, 'Deplatforming the Far-Right', 2.
[215] Y-Kollektiv, *Content Moderator*innen: Sie Löschen Die Videos Auf Social Media, Die Du Nicht Sehen Sollst*, 2020.
[216] Translated from German, Y-Kollektiv, 6:04-6:09 min.
[217] Translated from German, Y-Kollektiv, 6:15-6:22.
[218] The interviewees have emphasized how traumatizing their job is for them. One of the interviewees stated that she dreams about the contents, especially child pornography and animal abuse has been burdening her. Psychological help was, according to them, available. Y-Kollektiv, 4:40 - 5:20 min.

Facebook in the United States complaining about the psychological toll their work took on them.[219] Facebook ended up settling the case for 52 million US dollars.[220]

The working conditions are not fostering an environment which allows a thorough investigation of the contexts of posts leading to faulty decisions. For instance, content moderators decided to delete the historical picture of Kim Phúc[221] due to Facebook's child nudity ban.[222] This example shows the difficulty when working with oversimplified dichotomies, such as nudity or non-nudity. This decision cannot simply be answered with a yes or no. Context is crucial. Content moderators act on behalf of the companies, and the content of the companies is their capital. The posts are supposed to attract attention. Thus, content moderators often act against their value systems when deciding upon deletion or non-deletion.[223] Still, many argue that human assessment is the better option as automated means. While this could be true and content moderators' work might be slightly less error-prone, their labour is contentious. The number of decisions they have to make is too high for their assessment to be irrefutable and reliable. Furthermore, watching and reading illegal content all day is incredibly burdensome, especially for untrained young workers.

The following section on soft control is more about the interest of the intermediaries to make their platform more attractive and to highlight or suppress certain content.

2.2. Soft Control of Online Content

Soft control is primarily exercised through Artificial Intelligence which is, in this case, human-written and computer code-designed algorithms.[224] This following section focuses more on the invisible side of algorithms, even though they can also be used to filter posts, thereby exercising hard control.

Meta[225], Twitter[226], Google[227] and TikTok[228] use algorithms for the curation of user feeds. The use of algorithms as such has been criticised as hindering individuals' autonomy, their

[219] Manuel Ernesto Larrondo and Nicolas Mario Grandi, 'Artificial Intelligence, Algorithms and Freedom of Expression', *Universitas*, no. 34 (Febuary 2021): 176.
[220] 'Scola v. Facebook', Content Moderators Settlement.
[221] "A Girl in the Picture" by Kim Phúc, an important and iconic photo, see Jillian C. York, 'Facebook's Nudity Ban Affects All Kinds of Users', Electronic Frontier Foundation, September 2016.
[222] Roberts, 'Digital Detritus'.
[223] Roberts.
[224] Larrondo and Grandi, 'Artificial Intelligence, Algorithms and Freedom of Expression', 172.
[225] Facebook uses the algorithm EdgeRank, see 'How Facebook Distributes Content', Meta Business Help Center and Jeff Widman, 'EdgeRank'.
[226] E.g., Trends are determined and the start page is personalized by an algorithm, 'Twitter Trends FAQ', Help Center; 'Personalized Content Based on Your Third-Party Web Activity', Help Center.
[227] 'Trending on YouTube', YouTube Help; 'Manage Your Recommendations and Search Results', YouTube Help.
[228] 'How TikTok Recommends Videos #ForYou'.

functioning and output being too opaque and not giving users control in their interactions.[229] They are designed to promote sensational content, and thus, can, for instance, amplify false information.[230] The main issue is that users are unaware of content that is *not* being displayed. The combination of content omission and promotion creates "filter bubbles" or "echo chambers"[231]. Filter bubbles are "a unique universe of information for each of us"[232]. The term "echo chamber" describes groups' opinion-forming through personalized information environments "which consistently reflect an individual's opinion on themselves, like an echo"[233]. Thus, less diverse information environments. This comes back to a very inherent human desire to avoid cognitive dissonance and thereby to be more attracted to people and opinions like theirs.[234] Joy Buolamwini famously discovered AI biases that discriminate against people of colour and women.[235] There have also been reports of the suppression of LGBTQI+ activism.[236] In general, there is a lot of criticism of algorithms for social media, notably worries about their biases of for instance, certain content[237], locations[238] or groups[239].

As the Facebook Files revealed, Facebook was aware that its algorithms led to users being recommended more extreme, hateful, misogynist, and degrading content.[240] Facebook realized that if the algorithms are changed to be safer, users will spend less time on the platform.[241] As Ms Haugen put it: "Facebook makes more money when you consume more content. People enjoy engaging with things that elicit an emotional reaction. And the more anger they get

---

[229] Dogruel, Facciorusso, and Stark, '"I'm Still the Master of the Machine." Internet Users' Awareness of Algorithmic Decision-Making and Their Perception of Its Effect on Their Autonomy', 3.
[230] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/HRC/47/25', para. 16.
[231] Birgit Stark and Daniel Stegmann, 'Are Algorithms a Threat to Democracy? The Rise of Intermediaries: A Challenge for Public Discourse', Governing Platforms (Algorithm Watch, 2020), 14.
[232] Eli Pariser, *Beware Online 'Filter Bubbles'*, TED Talks, 2011, 9.
[233] Stark and Stegmann, 'Are Algorithms a Threat to Democracy? The Rise of Intermediaries: A Challenge for Public Discourse', 14.
[234] Stark and Stegmann, 3.
[235] Joy Buolamwini, 'The Coded Gaze: Bias in Artificial Intelligence' (2019).
[236] See chapter 3. Ben Bours, 'Facebook's Hate Speech Policies Censor Marginalized Users', Wired, 2017.
[237] Twitter's algorithm has been found to amplify politically right-wing content more than left wing content. For an interesting read, see Ferenc Huszár et al., 'Algorithmic Amplification of Politics on Twitter', *Proceedings of the National Academy of Sciences* 119, no. 1 (2022).
[238] Isaac Johnson et al., 'The Effect of Population and "Structural" Biases on Social Media-Based Algorithms: A Case Study in Geolocation Inference Across the Urban-Rural Spectrum', in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Conference on Human Factors in Computing Systems, Denver Colorado, 2017).
[239] Notably, against African Americans. Maarten Sap et al., 'The Risk of Racial Bias in Hate Speech Detection', *ACL*, 2019, 1–11; Yasin Danyaal, 'Black and Banned: Who Is Free Speech For?', *Index on Censorship*, 2018.
[240] Svea Eckert, Lena Kampf, and Georg Mascolo, 'Neue Enthüllungen setzen Facebook weiter unter Druck', tagesschau.de.
[241] Meta did make Facebook safer in regards to misinformation around the 2020 election but then changed it again. Scott Pelley, 'Whistleblower: Facebook Is Misleading the Public on Progress against Hate Speech, Violence, Misinformation', CBS News, 2021.

exposed to, the more they interact and the more they consume."[242] Special Rapporteur Khan argued that algorithms are driving users towards extremist content and conspiracy theories, and this can, furthermore, through extensive personalisation, undermines users' possibility to choose their "information diet"[243].

The soft control of algorithms is popular because they show users the content they (likely) want to see. This, for one, enhances the satisfaction of the users and for the other, makes it more profitable for the platforms as they can place advertisements tailored to the user's interests – making it more likely for them to buy the advertised products. The following section will discuss the implications of content moderation and regulation on Freedom of Expression.

3. Discussion: The Current Governance of Social Media

Intermediaries primarily moderate user-generated content because they want to protect their brand. They want to boost their advertising revenue, create an enjoyable user experience, and protect themselves from any liability.[244]

Intermediaries in general have a two-fold interest; Individual users are their main target, and their needs are important to the success of the company. Yet, they have a "popularity bias"[245], they care immensely about group engagements and interaction. Companies often merely react rather than prevent. This leads to overly quick and often disproportionate reactions to controversial content. Their provisions in the terms of service and community guidelines tend to be very vague and subjective, giving the users unclear instructions and the platforms a lot of freedom

Hard and soft control affect Freedom of Expression in two distinct ways; Hard control has the potential to violate the publishing of content, thus, the public dimension, while soft control can influence the opinions, the private dimension of FoE.

As previously mentioned, hashing algorithms are the kind of automated preventive means that distinguishes between allowed and not allowed content before publishing (commonly referred to as filtering). Hashing often leads to the automated removal of content and as discussed above, the ECJ has previously warned how this is a foreseeable violation of Freedom of Expressen, even though the court slightly softened its stance recently. Yet, what

---

[242] Pelley.
[243] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/HRC/47/25', para. 66.
[244] Roberts, 'Digital Detritus'.
[245] I.e., they are orientated towards numbers and prefer offers that are preferred by the masses. Stark and Stegmann, 'Are Algorithms a Threat to Democracy? The Rise of Intermediaries: A Challenge for Public Discourse', 10.

is important is that hashing does not classify illegal or legal content, it classifies content which violates community guidelines, not international or European law. As intermediary guidelines tend to be stricter than the law, the deletion of *legal* content can happen when (1) content is wrongly identified as ToS or guidelines-violating or (2) content is correctly classified as breaching guidelines but would not violate human rights law. Algorithms and wrong classification can thus violate the user's rights to receive, investigate and disseminate "which would be limited not by a necessary law with a legitimate and proportional purpose, but by de facto "constellation" made up of inhumane algorithms"[246].

The creation of less diverse information environments through algorithms' soft control makes the platforms more attractive for users and is more likely to present the information they are interested in (including sponsored content). Thus, at first glance, this seems to enhance Freedom of Expression. Yet, at a second glance, the curation and personalization created these bubbles of similar information which are difficult to escape as the feed, trends and search results are personalized. They interfere with the possibility of being confronted with opposing ideas and opinions. This potentially takes away the users' self-determination. This can affect FoE as it limits the possibility of users to evaluate content, idiomatic uses, and cultural aspects of human beings.[247] It fundamentally affects opinion-forming processes. Freedom of Opinion is an absolute right. The question of whether personalized content is a violation of Freedom of Opinion is beyond this thesis and the lack of jurisprudence on the topic does not provide an answer either. What is clear is that it changes the way users educate themselves and thereby form opinions in an unprecedented way.

One solution would be to give the user the option to choose between a curated start page or one with chronological order[248] or the possibility of on and off switch buttons for algorithms. Thereby, the curation of the content is made with consent by the user, and they are free to choose if they want personalized content or not. This would bring more autonomy to the user and make them more aware of how much algorithms affect their information environment.

Flagging furthermore has an interesting effect on FoE; By hitting buttons to report content because, for instance, users are offended, they exercise Freedom of Expression. It is comparable to a 'like button'; It eases the exercise of freedom of expression by changing its nature. Flagging, as mentioned above, gives users enough power to bring attention to possibly illegal content, thus, democratizing content moderation more.

[246] Larrondo and Grandi, 'Artificial Intelligence, Algorithms and Freedom of Expression', 173.
[247] Larrondo and Grandi, 174.
[248] For instance, Twitter and YouTube offer similar options.

It remains that the sheer volume of content posted on these platforms[249] requires automatic regulation. However, even though the work of content moderators is flawed as well, algorithms and automated responses have biases and delete content, thereby restricting Freedom of Expression, often acting in error.

Further, the taking down of forbidden content is very inconsistent. The effects and prevalence of mis-/disinformation and hate speech are also blown out of proportion. They exist, and of course, action needs to be taken against them. Yet, the issue might be less dangerous than many think and the amount of falsely identified information is high. Thereby, particularly language is an issue. Fact-checking is hugely dependent on the state and language and the enforcement differs significantly according to the country and the language being spoken there.[250]

The erroneous nature of content moderation is significant. Content moderation can and does violate Freedom of Expression as it is not able to differentiate between lawful and lawful content. Content moderation privileges or devalues information. Depending on its level of threat – measured against the companies' policies and goals – the content is deleted or not. The detection – human or automated - is unreliable, discriminatory, and biased. Users' right to FoE can be adversely affected by content moderation methods. If they are flawed, it is to be expected that FoE will be violated.

Overall, this analysis was significantly hindered by (1) the information is hidden due to the mere amount and the decentralized way it is published on the internet and (2) the diversity of information published, making the comparison between companies more difficult. The research of the platform community guidelines is challenging as they are distributed in a decentralized way on the internet and the information provided varies drastically between the platforms. Information about, for instance, algorithms is almost drowned out by the amount of information and the distribution over multiple internet sites. Further, there is no single page for algorithms or machine learning. Usually, they are merely mentioned in the context of trends or personalization mechanisms. Yet, overall, Meta and Twitter appear to publish more details about their measures (including statistics, examples of prohibited contents in the case of Twitter and decisions of their algorithms in the case of Meta) while TikTok seemingly only

[249] In February 2020, *every minute* more than 500 hours of videos were posted on YouTube. Further, the website 'internet live stats' shows how many tweets are posted every day in real time, averaging 500 million tweets per day. 'YouTube: Hours of Video Uploaded Every Minute 2020', Statista; 'Twitter Usage Statistics - Internet Live Stats'.

[250] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/HRC/47/25', para. 75.

published the bare minimum. Furthermore, Google's trusted flagger program and Facebook's Social media council are measures that are very progressive and have potential.

Thus, the main issue of content moderation is opacity. Intermediaries periodically publish reports about the content they have taken down where they give information about post removals and (superficially) the reasons. They do not show what the posts exactly say or how much users interact with these posts. Thus, the impact cannot be assessed. Twitter, for example, cut off online services which were tracking deleted tweets of politicians to hinder them from revealing this data.[251] Therefore, the requirement for FoE restrictions to be provided by law is undermined and violations are risked. This opacity is taking away users' agency and hinders investigations and studies about their practices.

The next chapter will outline how and what content can be restricted due to measures by either governments or intermediaries through misuse or errors in the moderation of content. Thus, while this section primarily focused on publicly accessible information, the next section will try to explore the less transparently communicated effects of content moderation methods to see whether content moderation can already be classified as 'digital repression'.

[251] Sarah Perez, 'Twitter Shuts Down Services That Tracked Politicians' Deleted Tweets In 30 Countries', *TechCrunch,* August 2015.

Chapter III: Digital Repression

The last chapter established the regulations and mechanisms of social media for content moderation. It further, reflected on the threats they can post on the exercise of Freedom of Expression online. This chapter will explore how far the weaknesses of said measures can go. It furthermore explores what content tends to be restricted. Thereby it aims to answer the third subquestion of if platform governance can suppress social movements through the restriction of Freedom of Expression (III). FoE is a fundamental right in the sense that other rights such as Freedom of Assembly and Association are inherently connected.[252] Thus, if FoE is violated, it can affect these rights as well. This section builds on this notion and shows how the restriction of Freedom of Expression online can directly affect protest and social movements.

The idea behind restricting or censoring content is that it heightens the costs of accessing and spreading information.[253] Typically, this happens through one of three mechanisms: Fear, friction, or flooding.[254] *Fear* is evoked in social media users of when they access or share certain information there will be (the threat of) repercussions, i.e., some form of punishment.[255] In 2021, in 55 countries social media users were arrested, investigated, or convicted for posting content online.[256] Censorship through *friction* is a mechanism which makes accessing information online more difficult.[257] According to Roberts, the result of friction is that users need to either spend more time or money to spread or have access to information through blocking, reordering search engine results, slowing down (internet) access or shutting the internet down completely.[258] Through friction, information can thus still be accessed but with significantly higher costs. Lastly, *flooding* is the drowning out of information through the coordinated publishing of information online.[259] For instance, protest events in Syria, China, Russia, and Mexico have been overwhelmed by spam tweets.[260]

---

[252] The ECtHR even considers the arrest of protestors to be an interference of FoE, see Steel and Others v. The United Kingdom para 92.
[253] Margaret E. Roberts, 'Resilience to Online Censorship', *Annual Review of Political Science* 23, no. 1 (2020): 403.
[254] Roberts, 403.
[255] For instance, in Iran, the publishing of materials online that insult religion is punished by five years' imprisonment up to the death penalty and creating anxiety and unease in the public's mind can be punished by two years of imprisonment or up to 47 lashes. Article 19, 'Islamic Republic of Iran: Computer Crimes Law' (2012), 19.
[256] The Freedom House, 'Freedom on the Net 2021. The Global Drive to Control Big Tech', 8.
[257] Roberts, 'Resilience to Online Censorship', 403.
[258] Roberts, 403.
[259] Roberts, 403.
[260] John-Paul Verkamp and Minaxi Gupta, 'Five Incidents, One Theme: Twitter Spam as a Weapon to Drown Voices of Protest', 2013, 2.

Yet, while online censorship focuses primarily on governments, censorship today can also be done by private companies, namely internet intermediaries. The concept of digital repression accounts for this and expands online censorship through its effects on contestation and the Right to Assembly and Association. While social media is often cited as having positive effects on social movements, as it did on the Arab Spring protests[261], this theory highlights the other side of the coin.

Thus, digital repression is a phenomenon which happens through censorship practices. It can be defined as "actions directed at a target to raise the target's costs for digital social movement activity and/or the use of digital or social media to raise the costs for social movement activity"[262]. This concept highlights how restrictions on online expressions can affect other rights; In this case, how it can hinder social and protest movements. The theory can be viewed as the consequence of *underregulated* and *overregulated* social media. For instance, underregulated social media can allow authoritarian governments to influence social media platforms while overregulated social media can lead to overblocking.

Earl, Maher, and Pan describe processes that go further than traditional repressive tactics, (such as arrests) which include controlling and challenging information.[263] These tactics pick up Rodger's flooding and friction mechanism and describe them in more detail. According to them, digital censorship has elements of traditional censorship, such as suppressing knowledge through, for instance, banning books or controlling media ownership but it suppresses individual expression as well.[264] The framework is applicable in autocracies and democracies alike.[265] Through this, behaviour and perceptions are affected and movement coordination and mobilization are hindered as they rely hugely on online tools.[266] There is a distinction between overt and covert control (which is comparable to hard and soft control mentioned in chapter 2).

[261] E.g., Christos Frangonikolopoulos, 'Explaining the Role and Impact of Social Media in the "Arab Spring"', *Media Journal* 8, no. 1 (2012): 10–20.
[262] Jennifer Earl, Thomas V. Maher, and Jennifer Pan, 'The Digital Repression of Social Movements, Protest, and Activism: A Synthetic Review', *Science Advances* 8, no. 10 (2022): 1.
[263] Earl, Maher, and Pan, 1, 5.
[264] Earl, Maher, and Pan, 5f.
[265] Earl, Maher, and Pan, 'The Digital Repression of Social Movements, Protest, and Activism'.
[266] Earl, Maher, and Pan, 6.

Table 1. *Information Control*[267]

	Information Coercion		**Information Channelling**	
	overt	covert	overt	covert
State agents tightly coupled with national political officials	-1- Limited national Internet connectivity (e.g., North Korea), temporary Internet blackouts, and state-based content filtering	-2- National content filtering where that filtering is not clear to users (e.g., returning 404 errors for filtered material)	-3- Government accounts posting distracting information and/or flooding online spaces or hashtags with irrelevant material	-4- Government disinformation and/or misrepresentations that influence contention
State agents loosely connected with national political officials	-5- Regional Internet blackouts and/or content filtering	-6- Regional content filtering where that filtering is not clear to users	-7- Local government or police information posting distracting information and/or flooding online spaces or hashtags with irrelevant material	-8- Local government and/or police disinformation and or misrepresentations that influence contention
Private agents	Deplatforming activists or organizations and/or moderating activist or organizational content -9-	Down-ranking, search filtering, shadow banning, throttling the spread of, or otherwise making protest-related material more obscure -10-	Private actors posting distracting information and/or flooding online spaces or hashtags with irrelevant material -11-	Private disinformation and/or misrepresentations that influence contention -12-

Information Coercion (which is like the friction mechanism) includes all restrictions on digital information, namely internet shutdowns (cell 1 and 5).[268] The OHCHR defines internet shutdowns as "all measures that intentionally prevent or disrupt access to, or dissemination of, information online is "shutdowns". Shutdowns come in a wide range of forms, including bandwidth throttling to slow internet access, blocking of specific apps […] and the partial or complete shutdown of access to the internet"[269]. Cell 9 describes the role of private organizations in restricting internet access out in the open. This also includes when companies are forced to remove content based on laws or political pressure. Regarding covered information coercion, cells 2 and 6 include 'secret' censorship, such as distributed denial of service attacks (DDoS)[270], filtering and slowing down the access to information.[271] For instance, Iran developed a 'national internet' which includes website blockings as early as

[267] Earl, Maher, and Pan, 6.
[268] Earl, Maher, and Pan, 'The Digital Repression of Social Movements, Protest, and Activism'.
[269] 'Internet Shutdowns and Human Rights', OHCHR, April 2021.
[270] DDoS attacks are "attacks preventing users of a network or a system from accessing relevant information, services and other resources", see 'Top Cyber Threats in the EU', Council of the EU, April 2022.
[271] Earl, Maher, and Pan, 'The Digital Repression of Social Movements, Protest, and Activism', 7.

2003, banned international media and think tanks and strict monitoring.[272] DDoS attacks are visible in the sense that a website is no longer accessible, yet, the reason why it is not accessible is unclear. Cell 10 includes similar measures by private companies including algorithms which can hinder or block the dissemination of information about protests. Furthermore, user shadow banning can lead to less visibility of content as well.[273]

Information channelling is the redirection of users' attention towards more preferable information through, for instance, changing the topic, including a particular point of view in the discussion, or making it seem as if one side has more support than the other.[274] This leads to people not being able to find the information they would have possibly been interested in (thus, similar to flooding). This can influence users' beliefs, second-order beliefs (beliefs about the beliefs of others), expression and behaviour.[275] Cells 3 and 7 describe situations when higher or lower state actors aim to redirect the attention of the public by introducing information or by concealing protest-relevant information.[276] The involvement of private actors is described in cell 11.[277] Covert information channelling includes the spread of disinformation, flooding of online spaces, or the source of information being misrepresented.[278] The case of state influence, for instance, in other countries domestic politics is described in cell 4.[279] Local governments (cell 8) also use these methods, for instance, the 50-cent army in China, a so-called troll-army[280]. Lastly, private agents also benefit from bots[281], trolls, click farms and influencers that support the agenda of the companies. There is a differentiation to be made between *platform governance* (see chapter 2)

[272] Marcus Michaelsen, 'Transforming Threats to Power: The International Politics of Authoritarian Internet Control in Iran', *International Journal of Communication*, no. 12 (2018): 3861, 3864.
[273] Earl, Maher, and Pan, 'The Digital Repression of Social Movements, Protest, and Activism', 7.
[274] Earl, Maher, and Pan, 7.
[275] Earl, Maher, and Pan, 7.
[276] Earl, Maher, and Pan, 7.
[277] These methods can also be divided into "misdirection" and "smoke screening". Misdirection is a method usually employed by magicians aiming for the audience to look another way while smoke screening hides or obscures content. Norah Abokhodair, Daisy Yoo, and David W. McDonald, 'Dissecting a Social Botnet: Growth, Content and Influence in Twitter', in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (CSCW '15, Vancouver: 2015), 849.
[278] Earl, Maher, and Pan, 'The Digital Repression of Social Movements, Protest, and Activism', 8.
[279] E.g., the South Korean National Intelligence Service published thousands of Twitter messages looking like ordinary citizens to draw attention to positive news, see Earl, Maher, and Pan, 8.
[280] Markus Reuter, 'Fake-News, Bots und Sockenpuppen – eine Begriffsklärung', netzpolitik.org, November 2016.
[281] Bots are "are accounts that can post content or interact with other users in an automates way and without direct human input" tools such as botometer (https://botometer.osome.iu.edu/) analyze whether an account is a bot or not. Stefan Wojcik et al., 'Bots in the Twittersphere', *Pew Research Center: Internet, Science & Tech*, April 2018.

and *digital repression* while digital repression is targeting opportunities for contestation.[282] This chapter will discuss exactly this fine line.

Following this theory, this section explores what cases of information control can be found in recent years. For the sake of this analysis, the differentiation between local and national governments is not relevant to the research question, thus, they will be combined. This section follows an exploratory approach which aims to shed a light and summarize the cases in which social movements were hindered as Freedom of Expression has been restricted online. In the research, a special focus will be put on finding cases in the EU and Europe, but not exclusively. When content about movements from other parts of the world is taken down, this can also affect Europeans. The keyword search was conducted through the LexisNexis database, Google Scholar, Limo, Google, and DuckDuckGo.[283] The search was limited to the last 10 years, thus, from 2012 until April 2022. The next section will outline the results of this small study.

1. Social Media and Digital Repression

Regarding internet shutdowns (cell 1 and 5), in total, 18 internet shutdowns were found in Europe between 2016 and 2021 and 931 globally[284] For instance, in 2019, Russia blocked the internet to hinder anti-government protests[285]. Furthermore, it should be noted that the Russian government requested social media companies to take down posts that would encourage minors to participate in the Navalny[286] protests.[287] Turkey blocked Wikipedia for almost three years from 2017 until 2020[288], and in 2019, it restricted the internet in the South when military operations in Syria were put into motion, likely to protect the troops[289]. India threatened Twitter with prison if they do not delete certain accounts, Twitter deleted hundreds of accounts, but India wanted 500 accounts tweeting about the Farmer's protests more deleted.[290] After more requests, Twitter partly obeyed and banned some hashtags, some are only available outside India, Twitter did not clarify whether the accounts were deleted due to

[282] Jennifer Earl, Thomas V. Maher, and Jennifer Pan, 'The Digital Repression of Social Movements, Protest, and Activism: A Synthetic Review', *Science Advances* 8, no. 10 (2022): 1.
[283] Keyords included, e.g., internet shutdown, remove posts, deplatforming, banning, censor!, etc., combined with social media.
[284] AccessNow, '#KeepItOn STOP Data 2016-2021', Google Docs, 2021; United Nations High Commissioner for Human Rights, 'Internet Shutdowns: Trends, Causes, Legal Implications, and Impacts on a Range of Human Rights' (Geneva: Human Rights Council, May 2022), para. 25.
[285] Kelvin Chan, 'Internet Shutdowns a Growing Tool to Halt Protests', St. Lousi Post-Dispatch, February 2021.
[286] Alexei Navalni was the opposition leader in Russia, currently imprisoned.
[287] Chris, 'Russia Demands Removal of Social Media Posts Encouraging Navalny Protests', Jurist, January 2021.
[288] 'Turkey Restores Access to Wikipedia after 991 Days', *NetBlocks*, January 2020.
[289] 'Twitter, Facebook, WhatsApp and Instagram Restricted in Southern Turkey', *NetBlocks*, October 2019.
[290] Josefine Kulbatzki, 'Meinungsfreiheit in Indien: Twitter und indische Regierung ringen um Kontensperren', netzpolitik.org, Febuary 2021.

them violating ToS or because of the government requests.[291] In 2016, Zuckerberg (the CEO of Facebook) and Israel's Justice Minister and Prime Minister met several times to stop 'incitement' (as defined by Israel) online.[292] As a result, 95 % of content removal requests from the side of the state were reportedly granted by Facebook.[293]

All in all, internet shutdowns are a visible and observable means to halt social movements. Internet shutdowns need to be based on a legal foundation to fulfil the necessity of "provided by law" to explain a restriction of Freedom of Expression. Often, they are followed by ambiguous explanations, if at all. Furthermore, the legitimate aim of internet shutdowns can be dubious and suppressing protest activity is not deemed legitimate.[294] Shutdowns can have impacts on economic activities, education, and health.[295] The ECtHR held that the blocking of websites such as YouTube (which are not easily accessible by other means) is a violation of FoE if it fails to meet the three-step test[296] and the blocking of legal content results out of measures aiming to block illegal content is interfering with FoE of the owners of said legal content or the hosts of such.[297] Thus, measures as such – especially without any legal safeguards – amount to a violation of FoE.

Regarding deplatforming (cell 9), there have been reports about posts or users being deleted due to them being pro-Palestine[298][299], about the Syrian civil war[300] or the situation of

---

[291] Hashtags such as #ModiPlanningFarmerGenocide were likely targeted. Kulbatzki.

[292] 7amleh, 'Facebooks Content Regulation Between "Hate Speech" and Legitimate Political Expression as Freedom of Speech: Bias and Discrimination against Palestinians', Submission to the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 2018, 1.

[293] 7amleh, 1.

[294] Some internet shutdowns can follow arguably more legitimate goals but could still be disproportionate; For instance, the Turkish one to protect troops or when German prosecutors in 1996 demanded Internet Service Providers to block 4 million subscribers from reading sex-related content on parts of the Internet. The shutdown was not possible to be limited to only one region, so individuals worldwide were not able to access the internet. See Philip N Howard, Sheetal D Agarwal, and Muzammil M Hussain, 'The Dictators' Digital Dilemma: When Do States Disconnect Their Digital Networks?', *Issues in Technology Innovation* 13 (2011): 4.

[295] United Nations High Commissioner for Human Rights, 'Internet Shutdowns: Trends, Causes, Legal Implications and Impacts on a Range of Human Rights', paras 33, 36, 38, 39.

[296] Cengiz and Others v. Turkey, No. 48226/10, 14027/11 (ECtHR 1 December 2015) paras 51, 58, 66; See also Kablis v. Russia, No. 48310/16, 59663/17 (ECtHR 30 April 2019) paras 92, 97, 106, 107; Taganrog Lro and Others v. Russia, No. 32401/10, 44285/10 and Others (ECtHR 7 June 2022) para 224.

[297] Vladimir Kharitonov v. Russia, No. 10795/14 (ECtHR 23 June 2020) para 46.

[298] The classification was difficult, as there are reports of the Israeli government requesting the takedowns with 86% of requests being granted by big social media companies, namely Facebook, Twitter, and Google. See Viki Auslender, 'Shomrim - Israel's Censorship Meets Facebook's Compliance', Shomrim. The Center for Media and Democracy, November 2020.

[299] Kelly Kunzl, 'Big Tech Censors Palestinian Advocacy around the World, While Fostering Surge in Jewish Extremism in Israel', Mondoweiss, May 2021, 7; 7amleh - The Arab Center for the Advancement of Social Media, 'Facebooks Content Regulation Between "Hate Speech" and Legitimate Political Expression as Freedom of Speech: Bias and Discrimination against Palestinians'.

[300] On 21st August 2013, an attack with chemical weapons in Damascus killed hundreds. Brown Moses found that 78% of Facebook pages in his sample were connected to YouTube pages that posted information about the August 21 Sarin attacks in Damascus. It should be noted that the sample was small, but his observations are in line with other reports about the deleting of information about the Syrian civil war. See The Associated Press, 'Activists Worry YouTube Erasing Proof of Syria Atrocities', CBS News, September 2017; Michael Pizzi, 'The

the ethnic group Rohingya in Myanmar[301]. Facebook admitted that it failed to "prevent [Facebook] from being used to foment division and incite offline violence"[302] but refused to disclose data regarding the posts[303]. Around 80 accounts with more than 5 million followers advocating for the Occupy Wall Street movement were removed[304] as well as Facebook pages such as the "Free Thought Project" (a free speech page with 3.1 million followers and the "End the Drug War" with around half a million followers.[305]

The deletion of accounts and posts is the most extreme measure platform, as the previous section already established. The Terms of Service dictate what is allowed and what not and they do not specify forbidden topics. Deplatforming has been shown to affect the reach of content, i.e., the posts get fewer interactions.[306] Rauchfleisch and Kaiser deemed Deplatforming as "a highly effective tool in minimizing extreme speech"[307] in their analysis of the removal of far-right YouTubers. As this analysis showed, some topics were systematically targeted. This is an example of social media platforms being reactive to outside pressure and movements that might be controversial in respective regions. The line between overt and covert information coercion is fine as the social media platforms are not very straightforward in whose and what kind of content they delete. Deplatforming involves a notice that the account holder received, thus, this was classified as overt while DDoS attacks and wordfilters are not communicated at all.

Researching cases regarding covered information coercion (cells 2 and 6) was challenging, especially because the origin is never really revealed. Yet, according to Wikipedia Germany, the webpage Wikipedia was down for around 9 hours in 2019 due to a DDoS attack.[308] There have been reports of TikTok censoring content which is against Chinese interests and distortion about other countries' histories.[309] it was revealed that TikTok uses wordfilters leading to words related to LGBTQI+ (e.g., gay, homo, LGBTQ, LGBTQI, queer), sex

---

Syrian Opposition Is Disappearing From Facebook', The Atlantic, February 2014; 'Attacks on Ghouta: Analysis of Alleged Use of Chemical Weapons in Syria' (Human Rights Watch, September 2013); Brown Moses, 'How Facebook Is Destroying History - A Survey Of August 21st', *Brown Moses Blog*, February 2014.

[301] Betsy Swan, 'Exclusive: Facebook Silences Rohingya Reports of Ethnic Cleansing', *The Daily Beast*, September 2017.

[302] The statement was criticised due to the lack of implementation mechanisms recommendations. See Alex Warofka, 'An Independent Assessment of the Human Rights Impact of Facebook in Myanmar', *Meta*, November 2018, ; Alexandra Stevenson, 'Facebook Admits It Was Used to Incite Violence in Myanmar', *The New York Times*, November 2018.

[303] 'US Judge Orders Facebook to Release Anti-Rohingya Account Records', September 2021.

[304] Bours, 'Facebook's Hate Speech Policies Censor Marginalized Users'.

[305] Sanjana Varghese, 'Twitter Has Purged Left-Wing Accounts with No Explanation', *Wired UK*, October 2018.

[306] Rauchfleisch and Kaiser, 'Deplatforming the Far-Right', 22.

[307] Rauchfleisch and Kaiser, 22.

[308] 'Wikipedia Disrupted Globally in Apparent Denial of Service Attack', *NetBlocks*, September 2019.

[309] E.g., the May 1998 riots in Indonesia or the Tiananmen Square incidents. Hern, 'Revealed'.

(prostitution, porn, sex, sex work) and extremist content (e.g., national socialism, terrorist).[310] Furthermore, the word Auschwitz and the name of a Chinese tennis player were blocked.[311] TikTok confirmed that they use an automatic word filter to prevent "potentially harming comments"[312]

DDoS attacks violate Articles 2, 4, 5, and 11 of the Budapest Convention on Cybercrime of 2011 as they involve illegal access to a computer system, system interference, and possible data interference, aiding and abetting crimes.[313] As the access to the internet is hindered, the human rights impacts are similar to the ones of internet shutdowns. Moreover, wordfilters that either suppress or delete content which contains certain words are a drastic method to hinder the circulation of information. In the case of TikTok, the existence of filters was not reported transparently, and the words included cannot be assessed as necessarily leading to hate speech. Thus, this is a gross violation of Freedom of Expression[314], leading to prior censorship and suppressed online discussions.

Regarding cell 10, there have been reports about social media suppressing posts about the shooting of an unarmed black man[315], Uighurs[316] and the Black Lives Matter movement[317]. In 2019, Twitter deleted almost one thousand accounts due to the "deliberately and specifically attempting to sow political discord in Hong Kong" during the protests and about 200.000 accounts due to them violating ToS.[318] Most social media platforms deny shadowbanning or moderating political content, such as Twitter[319] TikTok[320] and Instagram[321]. Yet, there have been reports accusing TikTok, a Chinese-owned platform, of suppressing posts about the Hong Kong protests with hashtags #hongkongprotest, #freehongkong or #joshuawong showing hardly any results. Leaks showed that TikTok suppresses videos by disabled, queer, and overweight creators by instructing its moderators to mark videos of people with

[310] Translated from German. tagesschau.de, 'TikTok nutzt in Deutschland Wortfilter', tagesschau.de.
[311] Peng Shuai accused a high ranking Chinese official of sexual abuse.
[312] tagesschau.de, 'TikTok nutzt in Deutschland Wortfilter'.
[313] The convention currently has 66 state parties. Council of Europe, 'Convention on Cybercrime', 23.XI.2001 § (2001).
[314] TikTok did not explain the filters. Following the lack of explanations, there is no ground to assume that they were provided by law, following a legitimate aim or, of course, being proportionate.
[315] Zeynep Tufekci, 'What Happens to #Ferguson Affects Ferguson': *The Message*, August 2014.
[316] Which has been denied by Tiktok, see Gabriel Nicholas, 'Shadowbanning Is Big Tech's Big Problem', The Atlantic, April 2022. 'Nervous TikTok : Planet Money', NPR.org, 2021.
[317] Posts which contained the #BlackLivesMatter and #George Floyd had significantly less interaction than posts without. Megan McCluskey, 'These Creators Say They're Still Being Suppressed for Posting Black Lives Matter Content on TikTok', Time, 2020.
[318] Twitter Support, 'Information Operations Directed at Hong Kong', Twitter Safety, 2019.
[319] 'What Is a "Shadow Ban," and Is Twitter Doing It to Republican Accounts?', The New York Times, 2020.
[320] Reuter and Köver, 'TikTok'.
[321] Josh Constine, 'How Instagram's Algorithm Works', *TechCrunch*, Nicholas, 'Shadowbanning Is Big Tech's Big Problem'.

disabilities and thereby limiting their reach.[322] Furthermore, TikTok capped the reach of queer and overweight people were listed as special users due to them having a higher risk of bullying.[323] Lastly, contents that might threaten "national security" or show rural poverty, slums, beer bellies, crooked smiles, lack of teeth and houses with cracked walls might further be suppressed.[324] If videos were classified as 'auto r' and they received a certain number of views, they were categorized as not recommended and thereby not being shown on the for you page. This was applied to videos showing people with "facial disfigurements, autism, down syndrome, disabled people, people with some facial problems such as birthmark, slightly squint etc."[325] and "controversial politically motivated demonstrations" including "content involving sensitive leaders such as Putin, Trump, Kim Jong-un with […] topics such as conflicts", the given examples also include protests about the independence of Northern Ireland, Tibet, Taiwan, the holocaust, Tiananmen Square or the "2014 Ukrainian revolution"[326], which is difficult to assess within 15 seconds (see also chapter 2, section 2). Content displaying two men kissing is further deemed as an 'Islam offence' and geoblocked in certain regions. Contents about sexual orientation are categorized as 'Risk 3.4' and blocked in Islamic countries. TikTok responded to this that it is merely abiding by local laws. [327]

Limiting the reach of the specific content is affecting what users see on their feeds. The reception of overly curated content can seriously undermine the autonomy of users. The practices are discriminatory against minorities and, people coming from poorer backgrounds.[328] Thus, from the side of the information recipient, this is worrisome and can influence their opinion-forming. The uploader of content such as the one mentioned above will consequently receive less interaction.

Regarding overt information channelling, flooding today is primarily done through bots. There are reports about bots used in Mexico[329], South Korea[330], Italy[331] and France[332] during

[322] Chris Köver and Markus Reuter, 'Discrimination: TikTok curbed reach for people with disabilities', netzpolitik.org, December 2019.
[323] Sam Biddle, Paulo Victor Ribeiro, and Tatiana Dias, 'TikTok Told Moderators: Suppress Posts by the "Ugly" and Poor', The Intercept, March 2020.
[324] Biddle, Ribeiro, and Dias.
[325] Köver and Reuter, 'Discrimination'.
[326] Reuter and Köver, 'TikTok'.
[327] Reuter and Köver.
[328] Through, for instance, the suppressing of posts with slums or cracked walls in the background.
[329] So called Peñabots, named after the Mexican President Enrique Peña Nieto, have also been used to send out pro-government propaganda Samuel C. Woolley, 'Automating Power: Social Bot Interference in Global Politics', *First Monday* 21, no. 4 (March 2016).
[330] Choe Sang-Hun, 'Prosecutors Detail Attempt to Sway South Korean Election', *The New York Times*, November 2013.
[331] Before the 2017 elections, the party Lega encouraged its followers to turn into selfbots (i.e., bots created by the account holders) amplifying Lega's message. See, Ben Nimmo, '#ElectionWatch: Italy's Self-Made Bots', *DFRLab*, October 2018.

national elections to elevate right-wing voices. In Russia, the share of bots on Twitter rose as high as 85% at times between 2014 and 2015.[333] The bots primarily tweeted news stories, thereby controlling (to a certain degree) the online information environment, as they can promote specific news stories, such as pro-regime news. In general, there is a high presence of Russian propaganda and bots in French[334] and German[335] social media. There have been observations of spam bots which were "bombed" on Twitter to prevent protest activity in Syria, Bahrain, Iran, and Morocco.[336] In the case of Syria during the war, for example, the hashtag #Syria was flooded with messages and pictures about the beauty of Syria.[337] Further, some tweets highlighted natural disasters in other parts of the world.[338] No cases of overt information channelling by private agents could be found.

However, there are also some positive examples of bots, for instance, the WikiEdits bots[339] which disclose when the government edits Wikipedia articles. Bots become more and more human-like and therefore, their research is challenging. Also, some were found to be primarily on the periphery with little interaction.[340] Thus, their impact could be less meaningful than initially thought.[341] Yet, flooding is a method of online censorship which has become more popular in the last decade. It is an innovative way to circumvent the deletion of undesirable information. Subsequently, this could impact the right to opinion as well. The prevalence of topics on social media is often equated with the severity of the situation and it can be very dangerous as it distorts users' perception of reality.

Regarding disinformation (cell 4 and 8) as a method of covered state information channelling, CIA whistleblower Edward Snowden revealed that the GCHQ's[342] secret unit, the Joint Threat Research Intelligence Group tried to manipulate online discourse and activism to "general outcomes it considers desirable" through injecting "all sorts of false material onto the internet to destroy the reputation of its targets" and using "social sciences

---

[332] Ben Nimmo, 'Russian and French Twitter Mobs in Election Push', *DFRLab*, April 2017.
[333] Denis Stukal et al., 'Detecting Bots on Russian Political Twitter', *Big Data* 5, no. 4 (2017): 318.
[334] Ben Nimmo, 'The Kremlin's Audience in France', *DFRLab*, April 2017.
[335] Maks Czuperski, Ben Nimmo, and Barojan Donara, 'Russian Internet: Fake News Haven?', *Medium*, January 2017.
[336] Woolley, 'Automating Power'.
[337] Jillian C. York, 'Syria's Twitter Spambots', *The Guardian*, April 2011.
[338] For instance, some bots posted pictures of the consequences of Hurricane Sandy in the United States. Abokhodair, Yoo, and McDonald, 'Dissecting a Social Botnet', 849.
[339] Heather Ford, Elizabeth Dubois, and Cornelius Puschmann, 'Keeping Ottawa Honest—One Tweet at a Time? Politicians, Journalists, Wikipedians, and Their Twitter Bots', *International Journal of Communication* 10 (2016): 1–24.
[340] In this case the bots were found to be delegitimizing the Black Lives Matter movement. Siobhan Roberts, 'Who's a Bot? Who's Not?', *The New York Times*, June 2020.
[341] A study by Michael Kreil found that most evidence for social bots stems from three sources which he found to have misclassified significant percentages of humans as bots. See Michael Kreil, 'The Army That Never Existed: The Failure of Social Bots Research', 2019.
[342] Government Communications Headquarters, a British intelligence and security organization

and other techniques"[343] including fake victims blog posts, the posting of negative information, discrediting targets by, for instance, emailing their colleagues, discrediting companies (e.g., by stopping deals or ruining business relationships).[344] Furthermore, the official state media outlet of the unrecognized Donetsk People's Republic in Ukraine posted false information about US tanks being sent by the US.[345] In reality, only about one-twentieth of tanks were sent compared to what the article implied. Moreover, some more difficult to categorize examples are the so-called "Putinbots"[346], likely stemming from the Kremlin, posting misinformation targeting the influencing of the 2016 US election.[347] The bots also spread anti-Islam hashtags after the Brussel terror attacks, pro-Leave hashtags on the day of the Brexit referendum, hashtags targeting Macron before the national elections and false news about the shot down of Malaysia Airlines Flight 17.[348][349] Russian-controlled media also posted misinformation about the health of Hilary Clinton, the other presidential candidate in 2016, and favourable reports about Trump.[350]

Finally, private misinformation campaigns were done by, for instance, numerous individuals posting fake news about the upcoming US national election.[351] Most of them run websites publishing (often by stealing) articles that report positively on Trump, often with much misinformation.[352] Furthermore, the hashtag #MacronLeaks reached 47.000 tweets within just three hours two days before the 2017 elections.[353] 10 of the most active accounts tweeting the hashtag, posted over 1.300 tweets in about 3 hours, hinting toward them being bots.

Mis- and disinformation has been widely discussed since the Brexit referendum, the Trump election and the Covid-19 pandemic. Measures to systematically distribute false information

---

[343] Glenn Greenwald, 'How Covert Agents Infiltrate the Internet to Manipulate, Deceive, and Destroy Reputations', The Intercept, February 2014.
[344] Glenn Greenwald, 'Exclusive: Snowden Docs Show British Spies Used Sex and "Dirty Tricks"', NBC News, February 2014.
[345] Ben Nimmo, 'Three Thousand Fake Tanks', *Medium*, January 2017
[346] Andrei Soldatov and Irina Borogan, 'What Spawned Russia's "Troll Army"? Experts on the Red Web Share Their Views', *The Guardian*, September 2015.
[347] Stukal et al., 'Detecting Bots on Russian Political Twitter'.
[348] Stukal et al.
[349] Using hashtags such as #КиевСбилБоинг (Kiev shot the Boing) and #ПровокацияКиева (Kiev's provocation), later, international investigations concluded that this was done by Russia.
[350] Andrew Higgins, Mike McIntire, and Gabriel J. X. Dance, 'Inside a Fake News Sausage Factory: "This Is All About Income"', November 2016.
[351] They publish news that drives up traffic to their website, thus, posts about Trump, and anti-Muslim and anti-Mexican content. Higgins, McIntire, and Dance.
[352] Higgins, McIntire, and Dance.
[353] The hashtag led to supposed leaked emails from Macron. the Digital Forensic Research Lab found the origin of the news to be an alt-right US American whose tweets were reposted by many bots. Ben Nimmo, 'Hashtag Campaign: #MacronLeaks', *DFRLab*, May 2017.

are not compatible with Freedom of Expression and can lead to high levels of distrust and undermine democratic processes.

2. Discussion: The Repression of Social Movements

The sheer number of cases found (see appendix 3 for a full list) highlights the prevalence of digital repression. In other words, through censorship or censorship-alike means, social movements are hindered. Not all measures in this framework are platform governance mechanisms (namely, internet shutdowns) and some even highlight the necessity of regulation (e.g., the high prevalence of bots and misinformation). This does not lessen the fact that keyword filters, algorithms and decisions by content moderators affect the dissemination of information regarding social movements (e.g., regarding the Rohingya minority, feminist or LGBTQI activism). To answer the research question, platform governance definitely gives rise to opportunities to suppress social movements by limiting FoE.

What this exploration also showed is that already marginalized or suppressed groups, such as people with poor backgrounds or the LGBTQI+ community are especially targeted. Through silencing these groups, they receive less attention and will likely end up even more marginalized than before. If certain issues receive less reach, especially in democracies, this will have a significant effect. If an issue receives more attention, people tend to believe that it is serious and important.[354] The interventions are primarily focused on infrastructure; Thus, they are easy to conceal, and most users are not aware of them happening.[355] Users not being aware of them being influenced or manipulated exacerbates these effects.

In theory, content moderation and online censorship are distinct concepts. Yet, this section showed that there are inevitable overlaps in their effects, while their intentions are two ends of a spectrum. When access to the internet is hindered or posts are being removed without any explanation and in some cases, because of outside intervention, Freedom of Expression is significantly impacted. These above-mentioned cases also happened in European or democratic states. Consequently, this phenomenon is not limited to dictatorships. As already discussed, many measures are violating human rights laws. For instance, the right to hold opinions needs to be *without interference*. As our opinions are always influenced, the assessment of whether an interference is lawful or not is challenging. "Any effort to coerce the holding or not holding of any opinion is prohibited"[356] under Article 19 ICCPR. These

[354] Roberts, 'Resilience to Online Censorship', 405.
[355] Earl, Maher, and Pan, 'The Digital Repression of Social Movements, Protest, and Activism', 7.
[356] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/HRC/47/25', para. 35.

methods contravene Freedom of Expression, especially in cases where users are not aware. Thus, there is at least the possibility of seriously interfering with the Freedom of Opinion.

These measures hit society substantially as they can change people's perceptions, knowledge, and values. Therefore, whether China suppresses information about the Uyghurs or Facebook about the Rohingya, it does not matter what the motivation is, their effects are comparable. Subjects as such receive less attention which results in less positive change. This affects especially democratic societies as social media users are voters.

Companies should fight government requests when they find them to be too extreme, in particular coming from governments that have been accused of censorship, propaganda and the suppression of a diverse media landscape. They should rather, follow international (human rights) law.

It should be noted that this analysis was neither exhaustive nor without its limitations. The targeted search for cases of digital repression creates a bias in combination with biases in, for instance, newspapers, as most sources were journalistic articles. This analysis merely showed that there are cases to be found and not any comparison of countries, platforms or social movements is possible due to its limited, explorative, and qualitative nature.

This section highlighted that social media can be an instrument to suppress protest activity and social movements through the restriction of Freedom of Expression online. After establishing the weaknesses and the negative potential that flawed content moderation can have, the next section discusses the proposed Digital Services Act of the EU and other solutions on how these adverse effects could be mitigated.

Solutions and the Digital Services Act

After exploring the flaws and dangers of platform governance in the chapter above, this section will discuss potential solution approaches. First, the new EU Digital Services Act will be outlined and afterwards, the possibility of publicly owned social media and social media councils will be discussed.

Overall, the fact that social media regulation is flawed is widely known. In response, the European Commission proposed the new Digital Services Act (DSA) as part of the Digital Markets Act in late 2020.[357] It will amend the E-Commerce Directive (see chapter 2) whose "initial objectives have not been fully achieved"[358].

Article 3 lifts the intermediaries of liability for information published on their platform if they did not interfere with the information through, for instance, selecting the receivers or modifying the information.[359] Article 3(2) states explicitly that courts can still require online platforms to "terminate or *prevent* an infringement"[360] without mentioning further how. Regarding this, the case-law of the court mentioned in chapter 2 section 1 still applies. While the Directive prohibited monitoring, the DSA only relieves the platforms from "general monitoring or active fact-finding obligations" to protect privacy with Article 7.[361] Article 5(1) of the DSA (similar to Article 14(1) of the Directive) similarly lifts the liability of intermediaries (upon acting "expeditiously" [362] and without any knowledge). The DSA does not specify a time frame further as well. Yet, if companies fail to respond, they lose the immunity provided through the act which will still end up in them taking down content in less clear cases. The DSA further advances the notice and takedown systems of the Directive with Article 14. Paragraph 1 allows users to report what they perceive as illegal content and 14(2) determines what information the platform must disclose. Since the companies must make this information (e.g., reasons why the content is illegal, URL of the content and who reported the content) transparent, they will likely be less inclined to overblock. The ECtHR has previously

[357] 'The Digital Services Act Package'.
[358] European Commission, 'Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services and Amending Directive 2000/31/EC', 7; 'The Digital Services Act Package'.
[359] European Commission, 'Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services and Amending Directive 2000/31/EC', Article 3.
[360] Emphasis added, European Commission, Article 3.
[361] Article 19 published a report supporting this provision, see 'EU: ARTICLE 19's Recommendations for the Digital Services Act Trilogue', Article 19, 2022, 19.
[362] European Parliament and European Council, 'Directive 2000/31/EC on Certain Legal Aspects of Information Society Services, in Particular Electronic Commerce, in the Internal Market (Directive on Electronic Commerce)', 2000/31/EC § (2000), para. 46; European Commission, 'Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services and Amending Directive 2000/31/EC', Article 5.

stated that measures as such can often be an appropriate tool for balancing rights.[363] The NGO Article 19 recommends a 'notice to notice' procedure in which the user gets informed about the complaint regarding their content and thus, gives them the power to assess whether the report was legitimate or not.[364]

Notably, the DSA has the additions to the Directive; It has a separate section for very large platforms, established a Digital Service Coordinator for the Member States and a European Board for Digital Services: First, very Large Online Platforms (VLOPs) - with an average of 45 million or more monthly users - have more responsibilities under the DSA. This implicates most (if not all) known social media platforms from Facebook, YouTube, Instagram, and TikTok to Twitter, Reddit and Quora.[365] These platforms are obliged to perform yearly systemic risk assessments towards the violation of fundamental rights (including hate speech), create mitigation measures, disclose their recommendation and advertisement parameters and publish reports every six months.[366] Systemic risks include illegal speech, such as violations of Article 11 of the Charter and the "intentional manipulation […] with an actual or foreseeable negative effect on the protection of public health, minors, civic discourse or […] electoral processes and public security". [367] Article 19 argued that these provisions do not meet the legality test of international human rights law and that national governments tend to prefer upload filters, general monitoring, and time limits for the takedown of content deemed as illegal in response to vague provisions such as Articles 26 and 27.[368] Article 42(3) about penalties determines that the maximum penalty in case of non-compliance is 6 % of the annual income of the intermediaries and in case of incorrect, incomplete or misleading information the limit is 1%. Article 68 about representation gives users the right to mandate a body to exercise their rights on their behalf.[369]

Secondly, it established an independent Digital Services Coordinator of the respective Member States in Article 38 who will have oversight competencies to ensure compliance and (national and Member State) coordination [370]. Lastly, the DSA sets up a European Board for

[363] Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary, No. 22947/13 (ECtHR 2 February 2016) para 91.
[364] 'Regulation of Notice and Action Procedures in the Digital Services Act', Article 19, 19.
[365] 'Most Popular Social Networks Worldwide of January 2022'.
[366] European Commission, 'Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services and Amending Directive 2000/31/EC' Article 26, 27, 29, 30, and 33.
[367] I.e., the rights to respect for private and family life, freedom of expression and information, non-discrimination, and the rights of the child, see 'Charter of Fundamental Rights of the European Union', 2000/C 364/01 (2000).
[368] 'Due Diligence Obligations in the EU's Digital Services Act', 2.
[369] European Commission, 'Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services and Amending Directive 2000/31/EC', Article 68.
[370] European Commission Article 38.

Digital Services which advises the digital services coordinator, ensures consistency, and supervises very large platforms with the Commission and the digital services coordinator.[371]

The proposal of the DSA was broadly met with approval by, for instance, the Global Network Initiative[372] and Article 19[373]. There has been some criticism about Article 2[374] and Article 21[375]. There has been further concern about the lack of a specific children's clause obligating platforms to assess risks to children's rights and create mitigation measures[376], the lack of measures at the disposal of platforms to deal with systemic risks[377] and the differentiation of smaller and bigger platforms, possibly leading to smaller platforms having more free speech and simultaneously, due to the higher regulations, making it difficult for them to grow[378]. After the (closed door) trilogue meetings in 2022 and the political agreement of the Council and the Parliament, some changes were proposed; "dark patterns"[379], profile-based recommendations of content and targeted advertising for minors and sensitive data will be prohibited.[380] Furthermore, a 'crisis-response mechanism'[381] in response to the war in Ukraine was added which gives more power to the EU in certain situations. Article 19 argued that the decision cannot solely be made through an executive power (i.e., the Commission) and that the definition of crisis needs to be made clear so that the rule of law won't be undermined.[382]

All in all, the DSA requires new transparency and due process measures by intermediaries, notably the out-of-court dispute settlement mechanism can be promising even though difficult to implement. The focus is thus more on transparency and due process, not on viewpoint- or

---

[371] European Commission Article 47.
[372] Chris Sheehy, 'GNI Submission to European Commission Consultation on the Digital Services Act', *Global Network Initiative*, 2021.
[373] 'EU: The Draft Digital Services Act and the Digital Markets Act Must Protect Freedom of Expression - ARTICLE 19', Article 19, December 2020.
[374] There is criticism of the definition of "dissemination to the public" in Article 2(i) as "making information available […] to a potentially unlimited number of third parties" thereby excluding groups with limited participants, such as Telegram groups. See Alexander Peukert et al., 'European Copyright Society – Comment on Copyright and the Digital Services Act Proposal', *IIC - International Review of Intellectual Property and Competition Law* 53, no. 3 (March 2022): 367.
[375] Article 21 asks of intermediaries to even report merely suspicious criminal offences.
[376] 5RightsFoundation, 'Children's Rights in the EU Digital Services Act', Position Paper (SSOAR - GESIS Leibniz Institute for the Social Sciences, 2021).
[377] the fear is that the only possibilities platforms have to deal with systemic issues, such as the disseminating of conspiracy theories, to block users instead of only their content. For more details, see Fredrik Erixon, '"Too Big to Care" or "Too Big to Share":', *European Centre for International Political Economy*, no. 5 (2021): 11.
[378] Erixon, 6.
[379] Dark patterns are "deceptive elements that are intentionally crafted to make the users do actions that they wouldn't do otherwise" such as disguised ads, or fake notifications, see Corina Cara, 'Dark Patterns in the Media: A Systematic Review' VII, no. 14 (2019): 106.
[380] Council of the EU, 'Digital Services Act: Council and European Parliament Provisional Agreement for Making the Internet a Safer Space for European Citizens' (EU Press, April 2022).
[381] Council of the EU, 2.
[382] 'EU: Digital Services Act Crisis Response Mechanism Must Honour Human Rights', ARTICLE 19, April 2022.

content-based regulations which seem to be progress.[383] The conditional immunity for the intermediaries is a way to incentivise them to adhere to international human rights law and national law without making them liable for everything that is posted on their platforms or forcing them to react within 24 hours. Still, once again, the companies are the ones deciding whether the content is illegal or not.[384] The proposal will be amended, so it is yet too early to have a final judgement on the effects this act will have on content moderation practices. The DSA is a promising step towards more regulation through the establishment of different mechanisms, such as the digital services cooperator, which can have an important oversight function to make sure that human rights are protected. In July, the proposal was adopted by the Parliament.[385] The final text was at the writing of this thesis not yet published.

One possible - while not perfect and currently almost utopian - solution is publicly owned social media. Through this, the issue of companies being private is avoided, and governments do not need to rely on private media to disseminate their information. Public media can be held more accountable. Taxpayers can be viewed as stakeholders; Thus, they have a say.

Private platforms have thus an interest in having divisive issues online (if they will not face sanctions for them) as loud and controversial content drives up engagement and interactions. Public social media[386] would have other aims. Just like “normal” public media, no advertisement would be allowed and thus, the issue of "micro-targeting"[387] and concealed influencing would be avoided. This seems unthinkable right now, but not too long ago, public television seemed unthinkable as well. Moreover, public-owned social media would exist parallel to private social media. It would simply be another option for users. Yet, giving governments more power could also become an issue. In the case of, for instance, China, where the Uyghur minority is suppressed, an outside US social media platform would be helpful as the suppression is likely to be less prevalent. Thus, if governments are the owners of social media, they can reinforce their values and worldviews, which could be incompatible with our, European or western, values. This idea, however, needs further research which explores the challenges and opportunities of its implementation.

A compromise would be the setting up of social media platforms that are decoupled from advertising markets, following a model like Wikipedia. Wikipedia is a great example because

---

[383] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, ‘A/HRC/47/25’, para. 59.
[384] ‘Regulation of Notice and Action Procedures in the Digital Services Act’.
[385] ‘Digital Services Package’, European Commission, July 2022.
[386] Mark Coatney, the former director of the social media platform Tumblr, theorized about the idea of non-profit public social media in a very interesting opinion in the New York Times. See, Mark Coatney, ‘We Need a PBS for Social Media’, *The New York Times*, September 2019.
[387] Micro-targeting describes the using of data and demographics to receive curated advertisement, similar to the functioning of algorithms that curate posts shown to users.

it is a non-profit which is dependent on donations, and it has talk or discussion pages where article editors discuss the quality of the content and possible changes.[388] This allows transparency and agency for users in determining what content is published and what is not. Following a similar model of discussion pages could also be possible for the social media platforms mentioned in this thesis.

Ideally, a new independent decision-making body would be created. Another more viable idea is the setting up of social media councils. Ideally, the body should be set up with experts of law while being aware of the sociology and psychological effects of content moderation. The council would be able to create standards for social media platforms, review content moderation practices and decisions, receive reports, be more decentralized (e.g., federal[389]) to make them less prone to outside interference and function as a public forum. Social media councils could (1) provide external oversight based on human rights, (2) receive individual complaints, (3) create more transparency, and (4) represent the diversity of society.[390] Currently, there is a pilot project going on in Ireland with a council set up with representatives from social media companies, media and advertising industries, journalists, academics and civil society organisations.[391] The DSA proposes something similar with the out-of-court dispute settlement set up with Article 18 but it does not elaborate on what requirements this has. Facebook's Oversight Board is also a version of such a council, but its decisions are made according to Facebook's ToS and not human rights standards. Thus, at this moment the council and its effectiveness are too new to judge.

Lastly, some form of incentive could be given to social media platforms to adhere to human rights standards, possibly financially. For instance, states could restrict offshore data transfers to the ICTs (in his case) that comply with international human rights law.[392] The broadness of immunity could also be dependent on, for instance, certificated by NGOs such as the Global Network Initiative or Article 19.[393]

Whether the above-mentioned solutions are feasible is beyond this thesis. Yet, it is clear that the decision-making capacity can not only be trusted with private companies located outside the EU while their practices affect EU citizens as fundamentally as they do now. The following section will discuss the implications of the insights gained from the last sections.

---

[388] 'Help:Talk Pages', *Wikipedia*, April 2022.
[389] Germany has something comparable with the "Rundfunkräte" (i.e., broadcasting council) which monitors the compliance with the public broadcasting mandate.
[390] 'Social Media Councils: One Piece in the Puzzle of Content Moderation', ARTICLE 19, October 2021, 12f.
[391] 'Social Media Councils', 17.
[392] Brian Chang, 'From Internet Referral Units to International Agreements: Censorship of the Internet by the UK and EU', *Columbia Human Rights Law Review*, 2018, 159.
[393] Chang, 159.

Discussion: The Problem of Platform Governance for Freedom of Expression

Overall, this thesis found that platform governance can have adverse effects on Freedom of Expression. This happens through either flawed measures or external regulation including government pressure and companies' economic interests.

As chapter 1 established, Freedom of Expression enjoys broad protection, and any restriction needs to be explained extensively. Case-law showed that there is a demand for effective measures and that the ECtHR expects individuals also to take some responsibility for content posted on their pages. Further, hate speech must be assessed on the basis of the existence of patterns of tension. Yet, no automated means is currently able to achieve that. Content moderators will have a challenging time with this as well due to their time limits and no case-law to draw upon.

Chapter 2 highlighted more technological and implementational challenges of platform governance and how this can affect FoE online. The autonomy of users is quite marginalized, they have little control over what they see when they open social media apps or find out what contents specifically are forbidden. Community-based methods, such as flagging, are a method to give them some authority, but the final decision will always lie with the platforms and users can consider themselves lucky if they get an explanation of why content is deleted or suppressed.

On the surface, regulations by policymakers and platform companies appear similar. In both cases, some expression is to be protected and others to be avoided. Below the surface, however, it is obvious that this protection is differently motivated and affects different expressions. Content on social media is regulated by governments largely to eliminate mis- and disinformation and to protect others' rights. For example, during a period of many terrorist attacks in Europe in 2017, UK's prime minister Theresa May pledged for more social media regulations, putting some blame on intermediaries for the extremist acts.[394][395]

Thus, on the one hand, when states request the taking down of online content, it is usually, due to the protection of the right to privacy, the elimination of hate speech and mis- or disinformation, children's rights, intellectual property and also, and the protection of national security. The current EU approach with the E-Commerce Directive relies heavily on

[394] Jon Stone, 'Theresa May Says the Internet Must Now Be Regulated Following London Bridge Terror Attack', The Independent, June 2017.
[395] Facebook responded with a statement aiming for "Facebook to be a hostile environment for terrorists" and that they "work aggressively to remove terrorist content". Tom DiChristopher, 'Facebook Wants to Be "hostile Environment for Terrorists" as May Calls for Internet Regulations', CNBC, June 2017.

intermediary will. This was criticised in light of the Brexit as the EU prevents states to put more liability on intermediaries.[396] It also received criticism for leading to the over-removal of content resulting in possible (self-)censorship.[397] The main issue regarding content regulation is that the laws that are too broad, too vague, and ill-defined about, for instance, extremism, incitement for hatred, disinformation, or defamation "often serve as pretexts for demanding that companies suppress legitimate discourse"[398]

The issue with platform governance is its decentralized nature. The liability is divided into the users, the company behind the platform and the state. Following the law established in chapter 1, the take-down of content should be assessed by independent judiciary bodies. Unfortunately, this is not possible in practice as the judiciary takes too long to come to a judgement and the online sphere is developing too quickly for them to catch up. Still, illegal speech is sensitive; It should not reach too many people who could get affected negatively by seeing it or false information could be spread and for instance, affect elections. Giving intermediaries a time frame makes sense as the access to illegal speech should not be too long, the longer the posts stay online, the more harm they can do. Yet, this can result in the social media platforms being 'on the safe side' and deleting content "just to make sure". Furthermore, it is unrealistic to expect a judgement from a court within 24 hours or a week.[399]

On the other hand, intermediaries take down content which does not adhere to their ToS. Their definitions are broader than states' definitions but similarly vague. Furthermore, there is no case-law to research to find out what expression is exactly forbidden (except in the few cases of the Meta's Oversight Council). The vagueness of states and their regulations in addition to intermediary liability can also result in excessive filtering. The 'better safe than sorry' approach - where content is taken down to be on the safe side - is a real danger and harms Freedom of Expression worldwide immensely. This contradicts the test of necessity and proportionality. This further does not fulfil the demands of FoE restrictions to be 'provided by law'.

As mentioned in chapter 1, users need to know what they are allowed to publish online and what content might be deleted. Often, content moderation guidelines are only disclosed to the

---

[396] Rajeev Syal, 'Make Facebook Liable for Content, Says Report on UK Election Intimidation', The Guardian, December 2017.
[397] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/HRC/35/22', para. 39; Jørgensen and Zuleta, 'Private Governance of Freedom of Expression on Social Media Platforms', 52.
[398] Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'A/HRC/38/35', para. 13.
[399] In 2021, it took judicial authorities in Council of Europe states on average 778.47 days to take assess the legality of content. Mchangama, Alkiviadou, and Mendiratta, 'A Framework of First Reference: Decoding a Human Rights Approach to Content Moderation in the Era of "Platformization"', 14.

public through leaks which is not adhering to the standards set by the UN and Europe. Furthermore, giving too much liability to intermediaries goes against the legal frameworks as states are the primary duty bearers for the protection of FoE. The issue remains: The publication of hate speech or misinformation is dangerous, yet its quick takedown is also dangerous.

All in all, intermediaries' main objectives are not to protect human rights. They are private companies aiming for growth. The relationship between the user and the company is more determined by the ToS than by human rights treaties or EU policies. There is a clear hierarchy. Questions of restrictions of FoE are questions of the judiciary, not private companies. Any restrictions need to be demonstrated as being proportionate, which is challenging when these restrictions are automated or made by private companies which did not sign any international legal instruments such as the ICCPR.

When creating laws regarding content regulation, states focus on privacy and national security. These laws, however, can pressure intermediaries to (1) acts as quickly as possible and thereby potentially not taking enough context into account, (2) filter pre-publication so they won't have the issue of reviewing massive amounts of data and (3) to penalise very broadly to avoid any repercussions.

Platform governance operates in a *liability vacuum*. Social media platforms must (1) have effective measures in place but not violate Freedom of Expression, (2) not monitor their users but remove illegal content, and (3) follow human rights due diligence. Thereby, one cannot neglect the fact that these private companies have economic interests and operate globally under greatly varying legal systems and civil society expectations.

Chapter 3 showed that the observed suppression of social movements, be it intentional or not, is very worrisome. It showed that unfortunately, some measures are somewhere between methods of digital repression and legitimate platform governance. It puts the notion of the 'democratization of information through social media' into question. If distinct groups' voices are more likely to be blocked or drowned out, the positive effect of digitalization gets contradicted. It could be argued that it only seems as if everyone can post content for everyone to see, while, in reality, there are still more dominant and preferred groups, values, and perspectives. Platform governance is not the issue here per se, its flaws are. Thus, the right answer to subquestion 3 is that *flawed* (or orchestrated) platform governance can suppress social movements.

Finally, the request for effective measures is reasonable as they are necessary to protect other human rights. Regulations are crucial to stop the dissemination of hate speech and

disinformation, otherwise, they can have serious consequences on democratic processes. The ECtHR said that the obligation for effective measures to limit illegal speech is not to be equated to private censorship.[400] Yet, unfortunately, to date, there are no 'effective measures' that can adequately balance Freedom of Expression. As established in chapter 2, technology and the surrounding system (e.g., the conditions under which content moderators operate) do not allow for a reliable assessment of legal or illegal speech (disregarding the fact also that these measures are not aiming to find illegal speech but rather speech that violates their ToS). While the request of the ECtHR does not directly lead to private censorship, it is the logical consequence of this judgement in conjunction with Sanchez v. France[401] and Poland v. European Parliament and Council of the EU[402] of the ECJ is that more measures will be implemented that will endanger Freedom of Expression. In other words, and to answer this paper's research question, platform governance currently is having adverse effects on Freedom of Expression and as a result, social movements.

The regulatory agenda can be viewed as a fight over the monopoly of FoE. While the duty of FoE primarily lies within the state, undoubtedly social media platforms have the executive power. The vagueness of jurisdictions and the provisions on FoE further deepen this power. The EU and its Member States are at the forefront of this fight compared to other states.

The question remains of whether social media empowers users or rather private companies. Arguably, the vagueness of internet governance is as vague as the articles of Freedom of Expression. Freedom of Expression is a very context-dependent concept that seems impossible to regulate without a court. Yet, our courts do not have the capacity and are inherently not designed to judge digital content. If illegal content is online, it cannot simply stay online until the ECtHR decided but simultaneously, possibly illegal content can also not be taken down on the mere suspicion of illegality. Freedom of Expression is a quite sensitive human right, while having absolute elements, restrictions are needed but need to be thought through thoroughly. In an ideal world, judges would function as content moderators and be able to determine the legality of content. Hate speech is a fluent concept which is incredibly context-dependent and individual.

More than other human rights, crucially, Freedom of Expression has a time dimension. Most human rights violations can be trialled in hindsight because the illegal act has already been committed. Regarding FoE, as long as the post (in this case) is online, the 'act' is going on. However, "being on the safe side" because of the time sensitivity of information is also

[400] See chapter 1 section 2., Delfi AS v. Estonia paragraph 157.
[401] Sanchez v. France.
[402] Republic of Poland v European Parliament and Council of the European Union.

not acceptable. Therefore, many laws give time frames to the companies to decide (for instance, around 24 or 48 hours as in the case of NetzDG or the Code of Conduct on Countering Illegal Hate Speech). The time limit is restricting and inevitably ends up detaching context from the analysis. Yet, not giving a time constraint ends up creating a law vacuum in which content (1) may or may not be taken down or (2) stays online and possibly results in the harming of users. As discussed above, opinions enjoy broad protection under human rights law. The forming of opinions in regulated social media is different than in an unregulated one. Regulations lead to the overblocking of content (see chapter 3), filter bubbles or overly curated content (chapter 2, section 2). Thus, there is undoubtedly an effect on the opinion-forming processes of users.

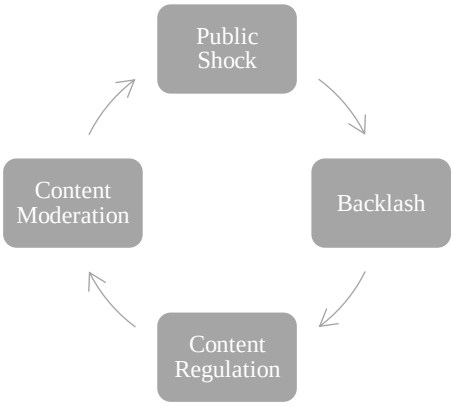
The accusation of the privatization of justice or Freedom of Expression is a dividing question in this field. On the one hand, it can be argued that the standards are set by the courts and the law and on the other hand, it seems reasonable to worry about the excessive penetration of private companies into human rights.[403] Private companies should not decide what groups and movements and speech should be suppressed online. The EU, for instance, stated it is far from excessive to ask intermediaries to balance Freedom of Speech upon notification of illegal content.[404] However, they dismiss the original dilemma which is not about asking private entities to take action on illegal content. The issue is rather that these private companies are the ones who are supposed to *assess* what content is illegal. Chapter 2 showed that the means for said assessment are limited and flawed. Similarly, lawmakers who are tasked with the role of trying to give intermediaries obligations to respect Freedom of Expression without giving them an incentive to overblock through the threat of liability. The horizontal axis and its protection are not clearly defined in the law or case law. The privatization of justice is thus an alarming trend. Social media platforms are not independent and impartial judges, they have economic interests and can be sanctioned if they do not act as precise as a judge.

The cases of digital repression are, especially if they are coming from the private platforms, 'public shocks' which resulted in backlash from the public and forced responses of the companies. The companies, thus, regulated more. Afterwards, these regulations are revealed in more public shocks which results in a cycle that could lead to more and more measures established through trial and error.

---

[403] For instance, 'The Privatisation of Censorship? Self-Regulation and Freedom of Expression', in *Codifying Cyberspace*, by Damian Tambini, Danilo Leonardi, and Chris Marsden, (Routledge, 2007), 432f.
[404] Věra Jourová, 'Code of Conduct – Illegal Online Hate Speech | Questions and Answers', *Directorate-General for Justice and Consumers*, 2016, 2.

Graph 1. The Cycle of Social Media Regulation.



If the rights of users are violated, they want to see change, just like a normal consumer. But the case of most of the covered instances of digital repression, most of them are primarily resulting out of platform governance. The governance was flawed otherwise there would not have been such an outcry. Likely, content moderation is the result of its surrounding power structures, namely the economic interests of the companies. If said interests could be lifted, the companies would be less susceptible to outside interference or internal hunger for growth.

While users generate the content posted on social media, the platforms have the power over said content upon posting (in cases of pre-moderation, even before the publishing). The users agree to the ToS and thereby, give private companies the right to their content. Upon posting, the power shifts from the user to the platform, creating a complete dependency of the user on the platform. The user trusts or hopes that their content will be (1) published, (2) shown to their followers or subscribers on their start page, (3) recommended to new users. Yet, the final decision lies with the companies.

The concept of the regulation of social media is inherently flawed as current regulatory mechanisms cannot assess integral factors of freedom of expression while judges need years to study this. The emergence of social media created this dilemma. It needs to be regulated, there is no doubt, but almost any regulation is flawed.

Conclusion

Coming back to the quote of Mark Zuckerberg from the introduction of this thesis, it seems clear that Zuckerberg was spot on with his statement. Social media platforms have an influence. However, as this thesis showed, not only because they provide a platform to many people but also because they use content moderation techniques. Their exercise affects the right of users to Freedom of Expression. As a result, social movements can be impeded.

It is debatable if comparable protests to the 'Facebook Revolution' part of the Arab Spring could take place today. Social media is becoming less of a free space for dissident but rather an instrument for control. It still has the potential to have this democratization power, but new and more reliable regulations need to be implemented. Thus, for a long time, social media was seen as a way to circumvent censorship, a way of rebellion against bigger structures. This notion transformed into a worry about the excessive power of social media platforms. As this thesis showed, today platform governance and digital repression are overlapping concepts. Platform governance is the umbrella term, while digital repression can be viewed as harsh criticism or a negative consequence of it. If platform governance can be assessed through the lens of digital repression, and examples can be found, we will find violations of Freedom of Expression.

Yet, what should also be remembered when talking about the transformative power of social media is that many of these processes are not completely new. Filtering, for instance, was previously done by journalists who decided what to publish and what not. Furthermore, newspapers and tv channels have also had economic aims in mind, they are not completely detached from the economic dimension. What is different today is that most publishing decisions are made automatically, which is completely legal.

In the end, the regulation of social media is the ultimate challenge of transversality, as it 'penetrates' so many layers within law, society, and sciences. There are many perspectives and most of them are justified, yet they contradict each other. They are not overlapping circles but lines that never cross each other, going in opposite directions. An online activist who owns violent pictures of human rights abuses understandably pledges for an unregulated internet, seeing it as a privilege, while parents of small children urge for more filters and content warnings so their children will not see posts that might traumatize them. The line between content moderation/regulation and censorship is very fine. The methods and the effects are at times identical, only possibly the intention differs. The difference is that content moderation *can* serve the public good, while censorship does not. The issue arises when the

two come too close together; Then, the intention does not matter anymore, and the positive effects of content moderation are marginalized.

Social movements without respected and protected Freedom of Expression are easily suppressed, especially if they threaten the status quo. Platform governance can become a weapon of private companies functioning in a capitalist system and of autocracies spreading their propaganda and hindering the spread of undesired information.

The balance of users' freedom of expression with other rights remains a challenge. Again, unregulated media is dangerous, and this thesis does not aim to pledge for no regulation. It rather discussed the current framework of platform governance and showed how this can violate Freedom of expression and hinder social movements. To emphasise, hate speech and mis- or disinformation are serious threats to our society and human rights. In the end, the question is how we can ideally balance these rights and as this thesis showed, currently this is not done to a satisfactory level, primarily due to the lack of progress in technology and inadequate measures. This thesis showed that current content moderation does not live up to human rights standards and needs to be amended.

In conclusion, what we can observe today is the heightened possibility of restrictions on Freedom of Expression resulting in possible censorship practices and a juxtaposition of private social media platforms and governments trying to exercise power over information which can have serious chilling effects. In other words, the current flawed platform governance does limit Freedom of Expression on social media.

Bibliography

5RightsFoundation. ‘Children’s Rights in the EU Digital Services Act’. Position Paper, 2021. https://www.ssoar.info/ssoar/handle/document/71817.

7amleh. ‘Facebooks Content Regulation Between “Hate Speech” and Legitimate Political Expression as Freedom of Speech: Bias and Discrimination against Palestinians’. Submission to the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 2018.

‘A Safer Twitter’. Accessed 4 July 2022. https://help.twitter.com/en/resources/a-safer-twitter.

Abokhodair, Norah, Daisy Yoo, and David W. McDonald. ‘Dissecting a Social Botnet: Growth, Content and Influence in Twitter’. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver: ACM, 2015), 839–851. https://doi.org/10.1145/2675133.2675208.

‘About Human Rights at Google’. Google. Accessed 10 May 2022. //www.google.com/human-rights/.

‘About the YouTube Trusted Flagger Programme’. YouTube Help. Accessed 5 July 2022. https://support.google.com/youtube/answer/7554338?hl=en-GB.

AccessNow. ‘#KeepItOn STOP Data 2016-2021’. 2021. https://docs.google.com/spreadsheets/d/1DvPAuHNLp5BXGb0nnZDGNoiIwEeu2ogdXEIDvT4Hyfk/edit?usp=sharing&usp=embed_facebook.

‘Account Safety’. TikTok Help Center. Accessed 6 July 2022. https://support.tiktok.com/en/safety-hc/account-and-user-safety/account-safety.

‘Activists Worry YouTube Erasing Proof of Syria Atrocities’. CBS News, September 2017. https://www.cbsnews.com/news/youtube-videos-syria-war-activists-human-rights-violations-war-crimes/.

‘Age-Restricted Content’. YouTube Help. Accessed 5 July 2022. https://support.google.com/youtube/answer/2802167?hl=en&ref_topic=9387060.

Ahmet Yildirim v Turkey, No. 3111/10 (ECtHR 18 December 2012).

Aksu v. Turkey, No. 4149/04, 41029/04 (ECtHR 15 March 2012).

Alkiviadou, Natalie. ‘The Legal Regulation of Hate Speech - The United Nations Framework as the Common Denominator for Europe and Asia’. *European-Asian Journal of Law and Governance*, 2018, 22–41.

Allan, Richard. ‘Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community?’ *Meta* (blog), June 2017. https://about.fb.com/news/2017/06/hard-questions-hate-speech/.

Ananny, Mike, and Tarleton Gillespie. ‘Public Platforms: Beyond the Cycle of Shocks and Exceptions’, 2017, 22.

Animal Defenders International v. the United Kingdom, No. 48876/08 (ECtHR 22 April 2013).

Anonymous [@youranonnews]. ‘Activist Accounts Being Suspended by @twitter b/c of Leaked Info about Russia’. *Twitter*, March 2022. https://twitter.com/youranonnews/status/1507911398835245059.

‘Appeal Community Guidelines Actions’. YouTube Help. Accessed 6 July 2022. https://support.google.com/youtube/answer/185111?hl=en.

Are, Carolina. ‘The Shadowban Cycle: An Autoethnography of Pole Dancing, Nudity and Censorship on Instagram’. *Feminist Media Studies*, May 2021, 1–18. https://doi.org/10.1080/14680777.2021.1928259.

Article 19. ‘Germany: Responding to “Hate Speech”. 2018 Country Report’ (London, United Kingdom, 2018). https://www.article19.org/wp-content/uploads/2018/07/Germany-Responding-to-%E2%80%98hate-speech%E2%80%99-v3-WEB.pdf.

———. ‘Islamic Republic of Iran: Computer Crimes Law’, 2012. https://www.article19.org/data/files/medialibrary/2921/12-01-30-FINAL-iran-WEB[4].pdf.

‘Attacks on Ghouta: Analysis of Alleged Use of Chemical Weapons in Syria’ (Human Rights Watch, September 2013). https://www.hrw.org/report/2013/09/10/attacks-ghouta/analysis-alleged-use-chemical-weapons-syria.

Auslender, Viki. ‘Shomrim - Israel’s Censorship Meets Facebook’s Compliance’. Shomrim. The Center for Media and Democracy, November 2020. https://www.hashomrim.org/eng/353.

Axel Springer Ag v. Germany, No. 39954/08 (ECtHR 7 February 2012).
Belarus Internet Observatory. 'Internet shutdown in Belarus', 2020. https://netobservatory.by/wp-content/themes/netobserver/assets/shutdown/belarus-shutdown-2020-en.pdf.
Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV, No. C-360/10 (ECJ 16 February 2012).
Biddle, Sam, Paulo Victor Ribeiro, and Tatiana Dias. 'TikTok Told Moderators: Suppress Posts by the "Ugly" and Poor'. The Intercept, March 2020. https://theintercept.com/2020/03/16/tiktok-app-moderators-users-discrimination/.
'Block Videos in Specific Territories'. YouTube Help. Accessed 5 July 2022. https://support.google.com/youtube/answer/6303378?hl=en.
Blunt, Danielle, and Zahra Stardust. 'Automating Whorephobia: Sex, Technology and the Violence of Deplatforming: An Interview with Hacking//Hustling'. *Porn Studies* 8, no. 4 (October 2021): 350–366. https://doi.org/10.1080/23268743.2021.1947883.
Bours, Ben. 'Facebook's Hate Speech Policies Censor Marginalized Users'. Wired, August 2017. https://www.wired.com/story/facebooks-hate-speech-policies-censor-marginalized-users/.
Breland, Ali. 'Twitter's New Privacy Policy Makes It Harder to Spread Warnings about Online Fascists'. *Mother Jones*, December 2021. https://www.motherjones.com/politics/2021/12/twitter-privacy-policy/.
Bröckling, Marie. 'Overblocking auf Twitter: Polizisten können zwei Tage lang nicht twittern'. netzpolitik.org, December 2019. https://netzpolitik.org/2019/polizisten-koennen-zwei-tage-lang-nicht-twittern/.
Buolamwini, Joy. 'The Coded Gaze: Bias in Artificial Intelligence' (presented at the Equality Summit, New York, March 2019). https://www.youtube.com/watch?v=eRUEVYndh9c.
Burges, Matt. 'We Finally Know the Full Extent of Russia's Twitter Trolling Campaign'. *Wired UK*, October 2018. https://www.wired.co.uk/article/twitter-troll-data-russia-ira-iran.
Callamard, Agnès. 'The Human Rights Obligations of Non-State Actors'. In *Human Rights in the Age of Platforms*, edited by Rikke Frank Jørgensen (Cambridge, MA: MIT Press, 2019).
Cara, Corina. 'Dark Patterns in the Media: A Systematic Review' VII, no. 14 (2019).
Cengiz and Others v. Turkey, No. 48226/10, 14027/11 (ECtHR 1 December 2015).
Ceylan v. Turkey, No. 23556/94 (ECtHR 8 July 1999).
Chan, Kelvin. 'Internet Shutdowns a Growing Tool to Halt Protests'. St. Louis Post-Dispatch, February 2021.
Chang, Brian. 'From Internet Referral Units to International Agreements: Censorship of the Internet by the UK and EU'. *Columbia Human Rights Law Review*, 2018, 100.
Charter of Fundamental Rights of the European Union, 2000/C 364/01 § (2000).
Chris. 'Russia Demands Removal of Social Media Posts Encouraging Navalny Protests'. Jurist, January 2021. https://www.jurist.org/news/2021/01/russia-demands-removal-of-social-media-posts-encouraging-navalny-protests/.
Clayton, Richard, Steven Murdoch, and Robert Watson. 'Ignoring the Great Firewall of China'. In *Privacy Enhancing Technologies* (Berlin: Springer, 2006), 20–35. https://link-springer-com.kuleuven.e-bronnen.be/content/pdf/10.1007/11957454.pdf.
Coatney, Mark. 'We Need a PBS for Social Media'. *The New York Times*, September 2019. https://www.nytimes.com/2019/09/24/opinion/public-broadcasting-facebook.html.
Coche, Eugénie. 'Privatised Enforcement and the Right to Freedom of Expression in a World Confronted with Terrorism Propaganda Online'. *Internet Policy Review* 7, no. 4 (November 2018): 1–18. https://doi.org/10.14763/2018.4.1382.
Committee of Ministers. 'Recommendation No. R. (7) 20 of the Committee of Ministers to Member States on "Hate Speech"' (Council of Europe, October 1997). https://rm.coe.int/1680505d5b.
'Community Guidelines Enforcement Report Jan - Mar 2022'. TikTok, June 2022. https://www.tiktok.com/transparency/en/community-guidelines-enforcement-2022-1/.
Confessore, Nicholas. 'Cambridge Analytica and Facebook: The Scandal and the Fallout So Far'. *The New York Times*, April 2018. https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html.
Constine, Josh. 'How Instagram's Algorithm Works'. *TechCrunch*. Accessed 27 June 2022. https://tcrn.ch/2LN5kDX.

Cornils, Matthias. ‘Designing Platform Governance’ Governing Platforms (Algorithm Watch, May 2020).
Council of Europe. Convention on Cybercrime, 23.XI.2001 § (2001).
———. European Convention on Human Rights (1950).
Council of the EU. ‘Digital Services Act: Council and European Parliament Provisional Agreement for Making the Internet a Safer Space for European Citizens’ (EU Press, April 2022). https://www.consilium.europa.eu/en/press/press-releases/2022/04/23/digital-services-act-council-and-european-parliament-reach-deal-on-a-safer-online-space/pdf.
Crawford, Kate, and Tarleton Gillespie. ‘What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint’. *New Media & Society* 18, no. 3 (March 2016): 410–428. https://doi.org/10.1177/1461444814543163.
Cumhuriyet Vakfı and Others v. Turkey, No. 28255/07 (ECtHR 8 January 2014).
Czuperski, Maks, Ben Nimmo, and Barojan Donara. ‘Russian Internet: Fake News Haven?’ *Medium* (blog), January 2017. https://medium.com/@DFRLab/russian-internet-fake-news-haven-b5acd9ebd06a.
Danyaal, Yasin. ‘Black and Banned: Who Is Free Speech For?’ *Index on Censorship*, September 2018. https://www.indexoncensorship.org/2018/09/black-and-banned-who-is-free-speech-for/.
Delfi AS v. Estonia, No. 64569/09 (ECtHR 16 June 2015).
Derakhshan, Hossein, and Claire Wardle. ‘Information Disorder: Definitions’. In *Understanding and Addressing the Disinformation Ecosystem*, by Annenberg School for Communication, First Draft, and Knight Foundation, 5–12, 2017.
‘Detecting Violations’. Transparency Center. Accessed 4 July 2022. https://transparency.fb.com/de-de/enforcement/detecting-violations/.
DiChristopher, Tom. ‘Facebook Wants to Be “Hostile Environment for Terrorists” as May Calls for Internet Regulations’. CNBC, June 2017. https://www.cnbc.com/2017/06/04/facebook-wants-to-be-hostile-environment-for-terrorists.html.
Digital, Culture, Media and Sport Committee. ‘Disinformation and “Fake News”: Final Report’ (London, United Kingdom: House of Commons, 2019).
‘Digital Services Package’. European Commission, July 2022. https://ec.europa.eu/commission/presscorner/detail/en/IP_22_4313.
Directive on Certain Legal Aspects of Information Society Services, in particular Electronic Commerce, in the Internal Market (Directive on Electronic Commerce), 2000/31/EC Directive § (2000).
Dogruel, Leyla, Dominique Facciorusso, and Birgit Stark. ‘“I’m Still the Master of the Machine.” Internet Users’ Awareness of Algorithmic Decision-Making and Their Perception of Its Effect on Their Autonomy’. *Information, Communication & Society*, December 2020, 1–22. https://doi.org/10.1080/1369118X.2020.1863999.
‘Due Diligence Obligations in the EU’s Digital Services Act’ (Article 19, May 2021). https://www.article19.org/wp-content/uploads/2021/05/Regulation-of-due-diligence-in-the-EU-DSA-1.pdf.
Earl, Jennifer, Thomas V. Maher, and Jennifer Pan. ‘The Digital Repression of Social Movements, Protest, and Activism: A Synthetic Review’. *Science Advances* 8, no. 10 (2022): 1–15. https://doi.org/10.1126/sciadv.abl8198.
Eckert, Svea, Lena Kampf, and Georg Mascolo. ‘Neue Enthüllungen setzen Facebook weiter unter Druck’. tagesschau.de. Accessed 26 June 2022. https://www.tagesschau.de/investigativ/ndr-wdr/facebook-whistleblower-vorwuerfe-103.html.
‘Egypt Revolution: 18 Days of People Power’. Aljazeera, January 2016. https://www.aljazeera.com/gallery/2016/1/25/egypt-revolution-18-days-of-people-power.
Erixon, Fredrik. ‘“Too Big to Care” or “Too Big to Share”:’ *European Centre for International Political Economy*, no. 5 (2021): 11.
‘EU: Digital Services Act Crisis Response Mechanism Must Honour Human Rights’. Article 19, April 2022. https://www.article19.org/resources/eu-digital-services-act-crisis-response-must-respect-human-rights/.
‘EU: The Draft Digital Services Act and the Digital Markets Act Must Protect Freedom of Expression’. Article 19, December 2020. https://www.article19.org/resources/digital-rights/.

European Commission. Commission Recommendation on measures to effectively tackle illegal content online, 2018/334 § (2018).

———. ‘Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services and Amending Directive 2000/31/EC’ (Brussels, December 2020). https://ec.europa.eu/info/sites/default/files/proposal_for_a_regulation_on_a_single_market_for_digital_services.pdf.

———. The EU Code of conduct on countering illegal hate speech online (2016). https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en.

European Foundation for South Asian Studies. ‘The Role of Fake News in Fueling Hate Speech and Extremism Online; Promoting Adequate Measures for Tackling the Phenomenon’, 2021. https://www.efsas.org/publications/study-papers/the-role-of-fake-news-in-fueling-hate-speech-and-extremism-online/.

European Parliament, and European Council. Directive 2000/31/EC on Certain Legal Aspects of Information Society Services, in particular Electronic Commerce, in the Internal Market (Directive on Electronic Commerce), 2000/31/EC § (2000). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32000L0031&from=EN.

‘Evidence of Internet Disruptions in Russia during Moscow Opposition Protests’. *NetBlocks*, August 2019. https://netblocks.org/reports/evidence-of-internet-disruptions-in-russia-during-moscow-opposition-protests-XADErzBg.

Facebook. ‘NetzDG Transparenzbericht. Januar 2022’, 2022. https://about.fb.com/de/wp-content/uploads/sites/10/2022/01/NetzDG-DE.pdf.

‘Facebook, WhatsApp, Instagram, Messenger down Globally’. *NetBlocks*, April 2019. https://netblocks.org/reports/facebook-whatsapp-instagram-messenger-down-globally-WJBZvWB6.

‘Facebook, WhatsApp, Instagram, Messenger down Globally in Extended Service Outage’. *NetBlocks*, October 2021. https://netblocks.org/reports/facebook-whatsapp-instagram-messenger-down-globally-in-extended-service-outage-JA6pPkyQ.

Flew, Terry, and Rosalie Gillett. ‘Platform Policy: Evaluating Different Responses to the Challenges of Platform Power’. *Journal of Digital Media & Policy* 12, no. 2 (June 2021): 231–246. https://doi.org/10.1386/jdmp_00061_1.

Ford, Heather, Elizabeth Dubois, and Cornelius Puschmann. ‘Keeping Ottawa Honest—One Tweet at a Time? Politicians, Journalists, Wikipedians, and Their Twitter Bots’. *International Journal of Communication* 10 (2016): 1–24.

Frangonikolopoulos, Christos. ‘Explaining the Role and Impact of Social Media in the “Arab Spring”’. *Media Journal* 8, no. 1 (2012): 10–20.

Fung, Brian. ‘Twitter Bans President Trump Permanently’. CNN, January 2021. https://www.cnn.com/2021/01/08/tech/trump-twitter-ban/index.html.

Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz - NetzDG) (2017).

Gillespie, Tarleton. ‘Governance of and by Platforms’. In *Handbook of Social Media*, edited by Jean Burgess, Thomas Poell, and Alice Marwick (SAGE, 2017), 1–30.

Google. ‘Entfernungen von Inhalten Nach Dem Netzwerkdurchsetzungsgesetz – Google Transparenzbericht’. Accessed 16 May 2022. https://transparencyreport.google.com/netzdg/youtube?hl=de.

Google LLC, YouTube Inc., YouTube LLC, Google Germany GmbH and Elsevier Inc. v Cyando AG, No. C-682/18 (ECJ 22 June 2021).

Gorwa, Robert. ‘What Is Platform Governance?’ *Information, Communication & Society* 22, no. 6 (May 2019): 854–871. https://doi.org/10.1080/1369118X.2019.1573914.

Greenspan, Rachel Leigh, and Elizabeth F. Loftus. ‘Pandemics and Infodemics: Research on the Effects of Misinformation on Memory’. *Human Behavior and Emerging Technologies* 3, no. 1 (January 2021): 8–12. https://doi.org/10.1002/hbe2.228.

Greenwald, Glenn. ‘Exclusive: Snowden Docs Show British Spies Used Sex and “Dirty Tricks”’. NBC News, February 2014. https://www.nbcnews.com/feature/edward-snowden-interview/exclusive-snowden-docs-show-british-spies-used-sex-dirty-tricks-n23091.

———. ‘How Covert Agents Infiltrate the Internet to Manipulate, Deceive, and Destroy Reputations’. The Intercept, February 2014. https://theintercept.com/2014/02/24/jtrig-manipulation/.
Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. ‘Fake News on Twitter during the 2016 U.S. Presidential Election’. *Science* 363, no. 6425 (January 2019): 374–378. https://doi.org/10.1126/science.aau2706.
Gündüz v Turkey, No. 35071/97 (ECtHR 14 June 2004).
Han, Eric. ‘Countering Hate on TikTok’. TikTok, August 2019. https://newsroom.tiktok.com/en-us/countering-hate-on-tiktok.
Handyside v. the United Kingdom, No. 5493/72 (ECtHR 7 December 1976).
‘Hate Speech Policy’. YouTube Help. Accessed 26 June 2022. https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=9282436.
‘Hate Speech. Transparency Center’. Meta. Accessed 26 June 2022. https://transparency.fb.com/policies/community-standards/hate-speech/.
‘Help:Talk Pages’. *Wikipedia*, April 2022. https://en.wikipedia.org/w/index.php?title=Help:Talk_pages&oldid=1082183982.
Hern, Alex. ‘Revealed: How TikTok Censors Videos That Do Not Please Beijing’. *The Guardian*, September 2019. https://www.theguardian.com/technology/2019/sep/25/revealed-how-tiktok-censors-videos-that-do-not-please-beijing.
Higgins, Andrew, Mike McIntire, and Gabriel J. X. Dance. ‘Inside a Fake News Sausage Factory: “This Is All About Income”’, November 2016.
High Commissioner of Human Rights, and United Nations. Guiding Principles on Business and Human Rights, HR/PUB/11/04 § (2011).
‘How Does YouTube Identify Content That Violates the Community Guidelines?’ YouTube Community Guidelines. Accessed 5 July 2022. https://www.youtube.com/howyoutubeworks/policies/community-guidelines/.
‘How Facebook Distributes Content’. Meta Business Help Center. Accessed 4 July 2022. https://www.facebook.com/business/help/718033381901819.
‘How TikTok Recommends Videos #ForYou’. TikTok, August 2019. https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you.
Howard, Philip N, Sheetal D Agarwal, and Muzammil M Hussain. ‘The Dictators’ Digital Dilemma: When Do States Disconnect Their Digital Networks?’ *Issues in Technology Innovation* 13 (2011): 11.
Human Rights Committee. ‘General Comment No. 34. Article 19: Freedom of Expression’, 2011. https://www.cambridge.org/core/product/identifier/S0002930000101204/type/journal_article.
Huszár, Ferenc, Sofia Ira Ktena, Conor O’Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. ‘Algorithmic Amplification of Politics on Twitter’. *Proceedings of the National Academy of Sciences* 119, no. 1 (January 2022). https://doi.org/10.1073/pnas.2025334119.
Hyatt, Sophia. ‘Facebook “Blocks Accounts” of Palestinian Journalists’. Aljazeera, September 2016. https://www.aljazeera.com/news/2016/9/25/facebook-blocks-accounts-of-palestinian-journalists.
‘I Don’t Think Facebook Should Have Taken down My Post.’ Facebook Help Center. Accessed 6 July 2022. https://www.facebook.com/help/2090856331203011?helpref=faq_content.
‘Instagram Restricted in Russia as Online Space Continues to Shrink’. *NetBlocks*, March 2022. https://netblocks.org/reports/instagram-restricted-in-russia-as-online-space-continues-to-shrink-JBQXvVAo.
‘Internet Disruptions Registered as Russia Moves in on Ukraine’. *NetBlocks*, February 2022. https://netblocks.org/reports/internet-disruptions-registered-as-russia-moves-in-on-ukraine-W80p4k8K.
‘Investigation of DDoS Attacks against Independent Media Shows Links to Philippine Government and Army – Qurium Media Foundation’. *Qurium*, July 2021. https://www.qurium.org/press-releases/investigation-of-ddos-attacks-against-independent-media-shows-links-to-philippine-government-and-army/.
Ith, Tracy. ‘Microsoft’s PhotoDNA: Protecting Children and Businesses in the Cloud’. Microsoft, July 2015. https://news.microsoft.com/features/microsofts-photodna-protecting-children-and-businesses-in-the-cloud/.
Jee, Charlotte. ‘Facebook Needs 30,000 of Its Own Content Moderators, Says a New Report’. MIT Technology Review, June 2020.

https://www.technologyreview.com/2020/06/08/1002894/facebook-needs-30000-of-its-own-content-moderators-says-a-new-report/.
Johnson, Isaac, Connor McMahon, Johannes Schöning, and Brent Hecht. 'The Effect of Population and "Structural" Biases on Social Media-Based Algorithms: A Case Study in Geolocation Inference Across the Urban-Rural Spectrum'. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver Colorado USA: ACM, 2017), 1167–1178. https://doi.org/10.1145/3025453.3026015.
Jørgensen, Rikke Frank, ed. *Human Rights in the Age of Platforms* Information Policy Series (Cambridge, MA: The MIT Press, 2019).
Jørgensen, Rikke Frank, and Lumi Zuleta. 'Private Governance of Freedom of Expression on Social Media Platforms: EU Content Regulation through the Lens of Human Rights Standards'. *Nordicom Review* 41, no. 1 (2020): 51–67. https://doi.org/10.2478/nor-2020-0003.
Jost, John T., Pablo Barberá, Richard Bonneau, Melanie Langer, Megan Metzger, Jonathan Nagler, Joanna Sterling, and Joshua A. Tucker. 'How Social Media Facilitates Political Protest: Information, Motivation, and Social Networks: Social Media and Political Protest'. *Political Psychology* 39 (February 2018): 85–118. https://doi.org/10.1111/pops.12478.
Jourová, Věra. 'Code of Conduct – Illegal Online Hate Speech | Questions and Answers'. *Directorate-General for Justice and Consumers*, 2016, 4.
Jürgens, Pascal, and Birgit Stark. 'The Power of Default on Reddit: A General Model to Measure the Influence of Information Intermediaries: The Influence of Information Intermediaries'. *Policy & Internet* 9, no. 4 (December 2017): 395–419. https://doi.org/10.1002/poi3.166.
Kablis v. Russia, No. 48310/16, 59663/17 (ECtHR 30 April 2019).
Karatas v. Turkey, No. 23168/94 (ECtHR 8 July 1999).
Kemp, Simon. 'Digital 2022: April Global Statshot Report'. DataReportal, April 2022. https://datareportal.com/reports/digital-2022-april-global-statshot.
Kestenbaum, Sam. '"Antifa's Most Prominent Jew" Booted From Twitter'. The Forward, June 2017. https://forward.com/fast-forward/374276/antifas-most-prominent-jew-booted-from-twitter/.
Kirkpatrick, David. *The Facebook Effect: The Inside Story of the Company That Is Connecting the World*. 1st ed. (New York: Simon & Schuster Paperbacks, 2011).
Klar, Rebecca. 'Tech Companies Seek to Choke out Russian State Media'. Text. *The Hill* (blog), March 2022. https://thehill.com/policy/technology/596813-tech-companies-seek-to-choke-out-russian-state-media/.
Klonick, Kate. 'The New Governors: The People, Rules, and Processes Governing Online Speech'. *Harvard Law Review* 131, no. 1598 (2018): 1596–1670.
Koltay, András. *New Media and Freedom of Expression: Rethinking the Constitutional Foundations of the Public Sphere* Hart Studies in Comparative Public Law (Oxford ; New York: Hart, 2019).
Köver, Chris, and Markus Reuter. 'Discrimination: TikTok curbed reach for people with disabilities'. netzpolitik.org, December 2019. https://netzpolitik.org/2019/discrimination-tiktok-curbed-reach-for-people-with-disabilities/.
Kreil, Michael. 'The Army That Never Existed: The Failure of Social Bots Research', 2019. https://michaelkreil.github.io/openbots/.
Kulbatzki, Josefine. 'Meinungsfreiheit in Indien: Twitter und indische Regierung ringen um Kontensperren'. netzpolitik.org, Febuary 2021. https://netzpolitik.org/2021/meinungsfreiheit-in-indien-twitter-und-indische-regierung-ringen-um-kontensperren/.
Kunzl, Kelly. 'Big Tech Censors Palestinian Advocacy around the World, While Fostering Surge in Jewish Extremism in Israel'. Mondoweiss, May 2021. https://mondoweiss.net/2021/05/big-tech-censors-palestinian-advocacy-around-the-world-while-fostering-surge-in-jewish-extremism-in-israel/.
Larrondo, Manuel Ernesto, and Nicolas Mario Grandi. 'Artificial Intelligence, Algorithms and Freedom of Expression'. *Universitas*, no. 34 (February 2021): 177–194. https://doi.org/10.17163/uni.n34.2021.08.
Lomas, Natasha. 'Twitter Uses Country-Specific Blocking Powers For The First Time To Restrict Neo-Nazi Account In Germany'. *TechCrunch*, October 2012. http://tcrn.ch/OKiIbv.
———. 'YouTube Geoblocks Russia Today, Sputnik Channels in Europe'. Tech Crunch, March 2022. https://tcrn.ch/35kz5uy.
L'Oréal SA and Others v eBay International AG and Others, No. 324/09 (ECJ 12 July 2011).

Lu, Jiayin, and Yupei Zhao. ‘Implicit and Explicit Control: Modeling the Effect of Internet Censorship on Political Protest in China’. *International Journal of Communication* 12 (2018): 3294–2216.
Luhn, Alec. ‘Ukraine Blocks Popular Social Networks as Part of Sanctions on Russia’. *The Guardian*, May 2017. https://www.theguardian.com/world/2017/may/16/ukraine-blocks-popular-russian-websites-kremlin-role-war.
MacFarquhar, Neil. ‘Russian Court Bans Telegram App After 18-Minute Hearing’. *The New York Times*, April 2018, sec. World. https://www.nytimes.com/2018/04/13/world/europe/russia-telegram-encryption.html.
MacKinnon, Rebecca, Elonnai Hickok, Allon Bar, and Hae-in Lim. ‘Fostering Freedom Online. The Role of Internet Intermediaries’ Series on Internet Freedom (UNESCO Publishing, 2014).
Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary, No. 22947/13 (ECtHR 2 February 2016).
‘Manage Your Recommendations and Search Results’. YouTube Help. Accessed 5 July 2022. https://support.google.com/youtube/answer/6342839?hl=en-GB&ref_topic=9257501.
Maza, Cristina. ‘Twitter and Facebook Shut down Anti-Muslim Posts by Far-Right Alternative for Germany Party’. Newsweek, January 2018. https://www.newsweek.com/twitter-facebook-anti-muslim-posts-germany-afd-767869.
McCluskey, Megan. ‘These Creators Say They’re Still Being Suppressed for Posting Black Lives Matter Content on TikTok’. Time, July 2020. https://time.com/5863350/tiktok-black-creators/.
McDonnell, Patrick, and Cecilia Sanchez. ‘When Fake News Kills: Lynchings in Mexico Are Linked to Viral Child-Kidnap Rumors’. Los Angeles Times, September 2018. https://www.latimes.com/world/la-fg-mexico-vigilantes-20180921-story.html.
Mchangama, Jacob. ‘The Real Threat to Social Media Is Europe’. *Foreign Policy*. Accessed 2 July 2022. https://foreignpolicy.com/2022/04/25/the-real-threat-to-social-media-is-europe/.
Mchangama, Jacob, Natalie Alkiviadou, and Raghav Mendiratta. ‘A Framework of First Reference: Decoding a Human Rights Approach to Content Moderation in the Era of “Platformization”’. *The Future of Free Speech Project*, 2021, 1–61.
‘Meet the Board’. Oversight Board. Accessed 28 June 2022. https://www.oversightboard.com/meet-the-board/.
Meserve, Stephen A., and Daniel Pemstein. ‘Google Politics: The Political Determinants of Internet Censorship in Democracies’. *Political Science Research and Methods* 6, no. 2 (April 2018): 245–263. https://doi.org/10.1017/psrm.2017.1.
Michaelsen, Marcus. ‘Transforming Threats to Power: The International Politics of Authoritarian Internet Control in Iran’. *International Journal of Communication*, no. 12 (2018): 21.
‘Mobile Internet Disrupted in Luhansk, Ukraine amid Heightened Tensions with Russia’. *NetBlocks*, February 2022. https://netblocks.org/reports/mobile-internet-disrupted-in-luhansk-ukraine-amid-heightened-tensions-with-russia-l8Wx7LAO.
Moses, Brown. ‘How Facebook Is Destroying History - A Survey Of August 21st’. *Brown Moses Blog*, February 2014. http://brown-moses.blogspot.com/2014/02/how-facebook-is-destroying-history.html.
‘Most Popular Social Networks Worldwide of January 2022’. Statista. Accessed 18 May 2022. http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.
‘Nervous TikTok: Planet Money’. NPR.org, January 2021. https://www.npr.org/2021/01/13/956558906/nervous-tiktok.
‘NetzDG Transparenzbericht’. TikTok, November 2021. https://www.tiktok.com/transparency/de-de/netzdg-2019-2/.
Nicholas, Gabriel. ‘Shadowbanning Is Big Tech’s Big Problem’. The Atlantic, April 2022. https://www.theatlantic.com/technology/archive/2022/04/social-media-shadowbans-tiktok-twitter/629702/.
Nieborg, David B, and Thomas Poell. ‘The Platformization of Cultural Production: Theorizing the Contingent Cultural Commodity’. *New Media & Society* 20, no. 11 (November 2018): 4275–4292. https://doi.org/10.1177/1461444818769694.
Nimmo, Ben. ‘#ElectionWatch: Italy’s Self-Made Bots’. *DFRLab*, October 2018. https://medium.com/dfrlab/electionwatch-italys-self-made-bots-200e2e268d0e.
———. ‘Hashtag Campaign: #MacronLeaks’. *DFRLab*, May 2017. https://medium.com/dfrlab/hashtag-campaign-macronleaks-4a3fb870c4e8.

———. ‘Russian and French Twitter Mobs in Election Push’. *DFRLab*, April 2017. https://medium.com/dfrlab/russian-and-french-twitter-mobs-in-election-push-bca327aa41a5.
———. ‘The Kremlin’s Audience in France’. *DFRLab*, April 2017. https://medium.com/dfrlab/the-kremlins-audience-in-france-884a80515f8b.
———. ‘Three Thousand Fake Tanks’. *Medium*, January 2017. https://medium.com/@DFRLab/three-thousand-fake-tanks-575410c4f64d.
Nowak, Manfred. *UN Covenant on Civil and Political Rights: CCPR Commentary*. 2nd ed. (Kehl am Rhein: Engel, 2005).
OHCHR. ‘Internet Shutdowns and Human Rights’. Office of the High Commissioner, April 2021. https://www.ohchr.org/sites/default/files/Documents/Press/Internet-shutdowns-and-human-rights.pdf.
Ooo Informatsionnoye Agentstvo Tambov-Inform v. Russia, No. 43351/12 (ECtHR 18 May 2021).
Opinion of Advocate General Saugmandsgaard, No. C-401/19 (ECJ 15 July 2021).
‘Our Range of Enforcement Options for Violations’. Twitter Help. Accessed 4 July 2022. https://help.twitter.com/en/rules-and-policies/enforcement-options.
‘Oversight Board’. Accessed 26 June 2022. https://www.oversightboard.com/.
Ozgur Gundem v. Turkey, No. 23144/93 (ECtHR 16 March 2000).
Pariser, Eli. *Beware Online ‘Filter Bubbles’* TED Talks, 2011. https://www.ted.com/talks/eli_pariser_beware_online_filter_bubbles.
Pasquale, Frank. *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge: Harvard University Press, 2015).
Pavia, Will. ‘Twitter Bans 70,000 QAnon Accounts in US Social Media Crackdown’. *The Times*, January 2021. https://www.thetimes.co.uk/article/twitter-bans-70-000-qanon-accounts-in-us-social-media-crackdown-wc9w5026w.
Pelley, Scott. ‘Whistleblower: Facebook Is Misleading the Public on Progress against Hate Speech, Violence, Misinformation’. CBS News, October 2021. https://www.cbsnews.com/news/facebook-whistleblower-frances-haugen-misinformation-public-60-minutes-2021-10-03/.
Perez, Sarah. ‘Twitter Shuts Down Services That Tracked Politicians’ Deleted Tweets In 30 Countries’. *TechCrunch*, August 2015. http://tcrn.ch/1NELAyM.
Perinçek v. Switzerland, No. 27510/08 (ECtHR 15 October 2015).
Perset, Karine. ‘The Economic and Social Role of Internet Intermediaries’. OECD Digital Economy Papers. Vol. 171 (OECD, April 2010). https://doi.org/10.1787/5kmh79zzs8vb-en.
‘Personalized Content Based on Your Third-Party Web Activity’. Twitter Help Center. Accessed 4 July 2022. https://help.twitter.com/en/using-twitter/tailored-suggestions.
Peta Deutschland v. Germany, No. 43481/09 (ECtHR 8 November 2012).
Petey. ‘Twitter Suspended Me for Tweeting Feminist Academic Research. Here’s Why That’s a Problem’. *Center for Civic Media MIT*, 2018. https://civic.mit.edu/index.html%3Fp=2298.html.
Peukert, Alexander, Martin Husovec, Martin Kretschmer, Péter Mezei, and João Pedro Quintais. ‘European Copyright Society – Comment on Copyright and the Digital Services Act Proposal’. *IIC - International Review of Intellectual Property and Competition Law* 53, no. 3 (March 2022): 358–376. https://doi.org/10.1007/s40319-022-01154-1.
Pizzi, Michael. ‘The Syrian Opposition Is Disappearing From Facebook’. The Atlantic, February 2014. https://www.theatlantic.com/international/archive/2014/02/the-syrian-opposition-is-disappearing-from-facebook/283562/.
Ralph. ‘Twitter Admits It Banned Belfast Telegraph Political Editor in Error for the Second Time in One Month’. Independent.ie, May 2021. https://www.independent.ie/irish-news/twitter-admits-it-banned-belfast-telegraph-political-editor-in-error-for-the-second-time-in-one-month-40465569.html.
Rane, Halim, and Sumra Salem. ‘Social Media, Social Movements and the Diffusion of Ideas in the Arab Uprisings’. *Journal of International Communication* 18, no. 1 (April 2012): 97–111. https://doi.org/10.1080/13216597.2012.662168.
Rauchfleisch, Adrian, and Jonas Kaiser. ‘Deplatforming the Far-Right: An Analysis of YouTube and BitChute’. *SSRN Electronic Journal*, 2021. https://doi.org/10.2139/ssrn.3867818.
‘Recommendations for the Digital Services Act Trilogue’. Article 19, February 2022. https://www.article19.org/resources/eu-article-19s-recommendations-for-the-digital-services-act-trilogue/.

‘Regulation of Notice and Action Procedures in the Digital Services Act’. Article 19. Accessed 18 May 2022. https://www.article19.org/resources/eu-regulation-of-notice-and-action-procedures-in-the-digital-services-act/.
‘Removals under the Network Enforcement Law’. Google Transparency Report. Accessed 5 July 2022. https://transparencyreport.google.com/netzdg/youtube?hl=en.
Republic of Poland v European Parliament and Council of the European Union, No. C-401/19 (ECJ 26 April 2022).
Reuter, Markus. ‘Fake-News, Bots und Sockenpuppen – eine Begriffsklärung’. netzpolitik.org, November 2016. https://netzpolitik.org/2016/fakenews-social-bots-sockenpuppen-begriffsklaerung/.
Reuter, Markus, and Chris Köver. ‘TikTok: Cheerfulness and censorship’. netzpolitik.org, November 2019. https://netzpolitik.org/2019/cheerfulness-and-censorship/.
Riis, Thomas, and Sebastian Felix Schwemer. ‘Leaving the European Safe Harbor, Sailing Towards Algorithmic Content Regulation’. *SSRN Electronic Journal* 22, no. 7 (2018): 1–21. https://doi.org/10.2139/ssrn.3300159.
Robert Faurission v France, No. CCPR/C/58/D/550/1993 (HRC 9 November 1996).
Roberts, Margaret E. ‘Resilience to Online Censorship’. *Annual Review of Political Science* 23, no. 1 (2020): 401–419. https://doi.org/10.1146/annurev-polisci-050718-032837.
Roberts, Sarah T. ‘Digital Detritus: “Error” and the Logic of Opacity in Social Media Content Moderation’. *First Monday* 3, no. 23 (March 2018). https://doi.org/10.5210/fm.v23i3.8283.
Roberts, Siobhan. ‘Who’s a Bot? Who’s Not?’ *The New York Times*, June 2020. https://www.nytimes.com/2020/06/16/science/social-media-bots-kazemi.html.
Roth, Yoel. ‘Introducing Our Crisis Misinformation Policy’, May 2022. https://blog.twitter.com/en_us/topics/company/2022/introducing-our-crisis-misinformation-policy.
Sanchez v. France, No. 45581/15 (ECtHR 2 September 2021).
Sang-Hun, Choe. ‘Prosecutors Detail Attempt to Sway South Korean Election’. *The New York Times*, November 2013. https://www.nytimes.com/2013/11/22/world/asia/prosecutors-detail-bid-to-sway-south-korean-election.html.
Sap, Maarten, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. ‘The Risk of Racial Bias in Hate Speech Detection’. *ACL*, 2019, 1–11.
Satariano, Adam, Paul Mozur, and Valerie Hopkins. ‘Shutdowns Of Internet Offer Lesson On Conflicts’. The New York Times, Febuary 2022.
Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM), No. C-70/10 (ECJ 24 November 2011).
Schroepfer, Mike. ‘Update on Our Progress on AI and Hate Speech Detection’. *Meta*, February 2021. https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/.
Schwarz, Karolin, and Lutz Helm. ‘Hoaxmap’. Accessed 29 May 2022. https://hoaxmap.org/index.html.
‘Scola v. Facebook’. Content Moderators Settlement. Accessed 25 June 2022. https://contentmoderatorsettlement.com/.
Shane, Scott. ‘The Fake Americans Russia Created to Influence the Election’, September 2017. https://www.nytimes.com/2017/09/07/us/politics/russia-facebook-twitter-election.html.
Sheehy, Chris. ‘GNI Submission to European Commission Consultation on the Digital Services Act’. *Global Network Initiative*, 2021. https://globalnetworkinitiative.org/dsa-submission-mar-21/.
Shirky, Clay. ‘The Political Power of Social Media: Technology, the Public Sphere, and Political Change’. *Council of Foreign Relations* 90, no. 1 (2011): 28–41.
Sissons, Miranda. ‘Our Commitment to Human Rights’. Meta, March 2021. https://about.fb.com/news/2021/03/our-commitment-to-human-rights/.
Smith, Rory, Seb Cubbon, and Claire Wardle. ‘Under the Surface: Covid-19 Vaccine Narratives, Misinformation and Data Deficits on Social Media. Executive Summary’. *First Draft*, 2020, 24.
‘Social Media Blocked in Turkey as Idlib Military Crisis Escalates’. *NetBlocks*, February 2020. https://netblocks.org/reports/social-media-blocked-in-turkey-as-idlib-military-crisis-escalates-r8VWGXA5.
‘Social Media Councils: One Piece in the Puzzle of Content Moderation’. Article 19, October 2021. https://www.article19.org/resources/social-media-councils-moderation/.

Soldatov, Andrei, and Irina Borogan. ‘What Spawned Russia’s “Troll Army”? Experts on the Red Web Share Their Views’. *The Guardian*, September 2015. http://www.theguardian.com/world/live/2015/sep/08/russia-troll-army-red-web-any-questions.
Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression. ‘A/67/357’ (United Nations General Assembly, September 2012).
———. ‘A/73/348’ (United Nations General Assembly, August 2018).
———. ‘A/HRC/32/38’ (Human Rights Council, May 2016).
———. ‘A/HRC/35/22’ (Human Rights Council, March 2017).
———. ‘A/HRC/38/35’ (Human Rights Council, April 2018).
———. ‘A/HRC/47/25’ Disinformation and Freedom of Opinion and Expression (Human Rights Council, April 2021).
Stack, Liam. ‘What Is a “Shadow Ban,” and Is Twitter Doing It to Republican Accounts?’ The New York Times, July 2018.
Stark, Birgit, and Daniel Stegmann. ‘Are Algorithms a Threat to Democracy? The Rise of Intermediaries: A Challenge for Public Discourse’ Governing Platforms (Algorithm Watch, May 2020).
‘Status of Ratification. International Covenant on Civil and Political Rights’. OHCHR, April 2022. https://indicators.ohchr.org/.
Steel and Others v. The United Kingdom, No. 24838/94 (ECtHR 23 September 1998).
Stevenson, Alexandra. ‘Facebook Admits It Was Used to Incite Violence in Myanmar’. *The New York Times*, November 2018. https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html.
Stone, Jon. ‘Theresa May Says the Internet Must Now Be Regulated Following London Bridge Terror Attack’. The Independent, June 2017. https://www.independent.co.uk/news/uk/politics/theresa-may-internet-regulated-london-bridge-terror-attack-google-facebook-whatsapp-borough-security-latest-a7771896.html.
Stukal, Denis, Sergey Sanovich, Richard Bonneau, and Joshua A. Tucker. ‘Detecting Bots on Russian Political Twitter’. *Big Data* 5, no. 4 (2017): 310–324. https://doi.org/10.1089/big.2017.0038.
Sürek v. Turkey (no. 1), No. 26682/95 (ECtHR 8 July 1999).
Swan, Betsy. ‘Exclusive: Facebook Silences Rohingya Reports of Ethnic Cleansing’. *The Daily Beast*, September 2017. https://www.thedailybeast.com/exclusive-rohingya-activists-say-facebook-silences-them.
Syal, Rajeev. ‘Make Facebook Liable for Content, Says Report on UK Election Intimidation’. The Guardian, December 2017. https://www.theguardian.com/society/2017/dec/13/make-facebook-liable-for-content-says-report-on-uk-election-intimidation.
Taganrog Lro and Others v. Russia, No. 32401/10, 44285/10 and Others (ECtHR 7 June 2022).
tagesschau.de. ‘TikTok nutzt in Deutschland Wortfilter’. tagesschau.de. Accessed 24 March 2022. https://www.tagesschau.de/investigativ/tik-tok-begriffe-blockade-101.html.
‘“The Big Delete:” Inside Facebook’s Crackdown in Germany’, September 2021. https://www.aljazeera.com/news/2021/9/28/the-big-delete-inside-facebooks-crackdown-in-germany.
‘The Digital Services Act Package’. Accessed 19 May 2022. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package.
The Freedom House. ‘Freedom on the Net 2021. The Global Drive to Control Big Tech’ (The Freedom House, 2021).
The Jewish Community of Oslo v Norway, No. CERD/C/67/D/30/2003 (CERD 15 August 2005).
‘The Privatisation of Censorship? Self-Regulation and Freedom of Expression’. In *Codifying Cyberspace* (Routledge, 2007), 279–299. https://doi.org/10.4324/9780203947067-16.
Thompson, Alex. ‘Twitter Appears to Have Fixed “Shadow Ban” of Prominent Republicans like the RNC Chair and Trump Jr.’s Spokesman’. *Vice*, July 2018. https://www.vice.com/en/article/43paqq/twitter-is-shadow-banning-prominent-republicans-like-the-rnc-chair-and-trump-jrs-spokesman.
Timciuc v. Romania, No. 28999/03 (ECtHR 12 October 2010).
Times Newspapers Ltd v. the United Kingdom, No. 3002/03, 23676/03 (ECtHR 10 March 2009).
‘Top Cyber Threats in the EU’. Council of the EU, April 2022. https://www.consilium.europa.eu/en/infographics/cyber-threats-eu/.

‘Transparency Center Homepage’. TikTok. Accessed 10 May 2022. https://www.tiktok.com/transparency/en/.
‘Trending on YouTube’. YouTube Help. Accessed 5 July 2022. https://support.google.com/youtube/answer/7239739?hl=en-GB&ref_topic=9257501.
Tucker, Joshua A., Jonathan Nagler, Megan Metzger, Pablo Barberá, Duncan Penfold-Brown, and Richard Bonneau. ‘Big Data, Social Media, and Protest’. In *Computation Social Science* (Cambridge University Press, 2016).
Tufekci, Zeynep. ‘What Happens to #Ferguson Affects Ferguson’: *The Message* (blog), August 2014. https://medium.com/message/ferguson-is-also-a-net-neutrality-issue-6d2f3db51eb0.
‘Turkey Restores Access to Wikipedia after 991 Days’. *NetBlocks*, January 2020. https://netblocks.org/reports/turkey-restores-wikipedia-access-QyKp568D.
Twitter. ‘Twitter Netzwerkdurchsetzungsgesetzbericht: Juli - Dezember 2021’. Accessed 16 May 2022. https://transparency.twitter.com/content/dam/transparency-twitter/netzdg/NetzDG-Jul-Dec-2021.pdf.
———. ‘Twitter’s Free Speech and Rights of People.’ Accessed 10 May 2022. https://help.twitter.com/en/rules-and-policies/defending-and-respecting-our-users-voice.
———. ‘Twitter’s Policy on Hateful Conduct’. Help Center. Accessed 3 July 2022. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.
‘Twitter and Facebook Restricted in Russia amid Conflict with Ukraine’. *NetBlocks*, February 2022. https://netblocks.org/reports/twitter-and-facebook-restricted-in-russia-amid-conflict-with-ukraine-JBZrogB6.
‘Twitter, Facebook, WhatsApp and Instagram Restricted in Southern Turkey’. *NetBlocks*, October 2019. https://netblocks.org/reports/twitter-facebook-whatsapp-and-instagram-restricted-in-southern-turkey-oy9RzE83.
Twitter Support. ‘Information Operations Directed at Hong Kong’. Twitter Safety, August 2019. https://blog.twitter.com/en_us/topics/company/2019/information_operations_directed_at_Hong_Kong.
‘Twitter Trends FAQ’. Twitter Help Center. Accessed 4 July 2022. https://help.twitter.com/en/using-twitter/twitter-trending-faqs.
‘Twitter Usage Statistics - Internet Live Stats’. Accessed 26 June 2022. https://www.internetlivestats.com/twitter-statistics/.
‘Twitter-Ban Feminist Defends Transgender Views Ahead of Holyrood Meeting’. *BBC News*, May 2019. https://www.bbc.com/news/uk-scotland-48366184.
‘Understanding When Content Is Withheld Based on Country’. Twitter Help Center. Accessed 4 July 2022. https://help.twitter.com/en/rules-and-policies/tweet-withheld-by-country.
United Nations. ‘International Convention on the Elimination of All Forms of Racial Discrimination. Declarations and Reservations’, March 1966. https://treaties.un.org/Pages/ViewDetails.aspx?src=IND&mtdsg_no=IV-2&chapter=4&clang=_en#EndDec.
———. International Covenant on Civil and Political Rights, 2200A (XXI) § (1966).
———. UDHR (1948).
United Nations General Assembly. ‘Resolution 59(I)’. Resolution (United Nations).
United Nations High Commissioner for Human Rights. ‘Internet Shutdowns: Trends, Causes, Legal Implications and Impacts on a Range of Human Rights’ (Geneva: Human Rights Council, May 2022). https://documents-dds-ny.un.org/doc/UNDOC/GEN/G22/341/55/PDF/G2234155.pdf?OpenElement.
‘US Judge Orders Facebook to Release Anti-Rohingya Account Records’, September 2021. https://www.aljazeera.com/news/2021/9/23/us-judge-orders-facebook-to-release-anti-rohingya-account-records.
Varghese, Sanjana. ‘Twitter Has Purged Left-Wing Accounts with No Explanation’. *Wired UK*, October 2018. https://www.wired.co.uk/article/twitter-political-account-ban-us-mid-term-elections.
———. ‘Twitter Has Purged Left-Wing Accounts with No Explanation’. *Wired UK*. Accessed 26 May 2022. https://www.wired.co.uk/article/twitter-political-account-ban-us-mid-term-elections.
Vejdeland and Others v Sweden, No. 1813/07 (ECtHR 9 May 2012).

Verkamp, John-Paul, and Minaxi Gupta. ‘Five Incidents, One Theme: Twitter Spam as a Weapon to Drown Voices of Protest’, 2013. https://www.usenix.org/conference/foci13/workshop-program/presentation/verkamp.

Vladimir Kharitonov v. Russia, No. 10795/14 (ECtHR 23 June 2020).

Wagner, Ben. ‘Free Expression? Dominant Information Intermediaries as Arbiters of Internet Speech’. In *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*, edited by Martin Moore and Damian Tambini (New York, NY: Oxford University Press, 2018), 219–240.

Warofka, Alex. ‘An Independent Assessment of the Human Rights Impact of Facebook in Myanmar’. *Meta*, November 2018. https://about.fb.com/news/2018/11/myanmar-hria/.

‘We Are All Khaled Said’. Accessed 12 July 2022. https://www.facebook.com/elshaheeed.co.uk/.

Węgrzynowski and Smolczewski v. Poland, No. 33846/07 (ECtHR 16 July 2013).

Widman, Jeff. ‘EdgeRank’. Accessed 5 July 2022. http://edgerank.net/.

‘Wikipedia Disrupted Globally in Apparent Denial of Service Attack’. *NetBlocks* (blog), September 2019. https://netblocks.org/reports/wikipedia-disrupted-globally-in-apparent-denial-of-service-attack-RyjoqY8g.

‘Wikipedia Outage with International Impact Detected’. *NetBlocks* (blog), May 2019. https://netblocks.org/reports/wikipedia-outage-with-international-impact-detected-pA2zdJAb.

Wojcik, Stefan, Solomon Messing, Aaron Smith, Lee Rainie, and Paul Hitlin. ‘Bots in the Twittersphere’. *Pew Research Center: Internet, Science & Tech* (blog), April 2018. https://www.pewresearch.org/internet/2018/04/09/bots-in-the-twittersphere/.

Woolley, Samuel C. ‘Automating Power: Social Bot Interference in Global Politics’. *First Monday* 21, no. 4 (March 2016). https://doi.org/10.5210/fm.v21i4.6161.

X v Federal Republic of Germany, No. 9235/81 (ECtHR 16 July 1982).

Y-Kollektiv. *Content Moderator*innen: Sie Löschen Die Videos Auf Social Media, Die Du Nicht Sehen Sollst*, 2020. https://www.youtube.com/watch?v=umafqnmvRsY.

Yong Joo-Kang v Republic of Korea, No. CCPR/C/78/D/878/1999 (Human Rights Committee 16 July 2003).

York, Jillian C. ‘Facebook’s Nudity Ban Affects All Kinds of Users’. Electronic Frontier Foundation, September 2016. https://www.eff.org/deeplinks/2016/09/facebooks-nudity-ban-affects-all-kinds-users.

———. ‘Syria’s Twitter Spambots’. *The Guardian*, April 2011, sec. Opinion. https://www.theguardian.com/commentisfree/2011/apr/21/syria-twitter-spambots-pro-revolution.

York, Jillian C., and Ethan Zuckerman. ‘Moderating the Public Sphere’. In *Human Rights in the Age of Platforms*, edited by Rikke Frank Jørgensen (Massachusetts: Massachusetts Institute of Technology, 2019).

‘YouTube Community Guidelines Enforcement FAQs’. Google Transparency Report Help Center. Accessed 6 July 2022. https://support.google.com/transparencyreport/answer/9209072?hl=en#zippy=%2Chow-does-automated-flagging-work.

‘YouTube, GMail and Google Services down in Multiple Countries’. *NetBlocks*, December 2019. https://netblocks.org/reports/youtube-gmail-and-google-services-down-in-multiple-countries-xyMk4GAZ.

‘YouTube Hate Speech and Harassment Policy’. YouTube. Accessed 26 June 2022. https://www.youtube.com/howyoutubeworks/our-commitments/standing-up-to-hate/.

‘YouTube: Hours of Video Uploaded Every Minute 2020’. Statista. Accessed 19 November 2021. https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/.

Zittrain, Jonathan. ‘Three Eras of Digital Governance’. *SSRN Electronic Journal*, 2019, 1–8. https://doi.org/10.2139/ssrn.3458435.

ZR 179/20 (BGH 29 July 2021).

Новости Роскомнадзора. ‘Социальные Сети Удаляют Призывы к Несовершеннолетним Принять Участие в Незаконных Акциях’. Роскомнадзор, January 2021. https://rkn.gov.ru/news/rsoc/news73344.htm.

Appendix 1: Selection of Relevant Case-law

1.1. The Court of Justice of the EU (ECJ)

Table 2. *Selection of relevant ECJ case-law*

Date	No.	Name	Relevance
26.04.22	401/19	**Poland v. Parliament and Council**	Provisions leading to prior review or filtering are liable to violate FoE (§ 55) but can be necessary when balancing with other rights (§§ 75, 83) and proportionate (§ 82) and thus, does not violate FoE (§ 84). Measures that filter and block lawful content when uploading are incompatible with FoE (§ 86)
22.06.21	682/18	**YouTube v. Cyando**	Hosting platforms are liable if they fail to act regarding illegally uploaded content *if* they received a notification (§§ 35, 118)
06.10.20	511/18, 512/18, 520/18	**La Quadrature du Net and Others v. Premier Ministre and Others**	The Directive is not applicable in cases of the protection of the confidentiality of communications and natural persons (§§ 212, 229)
03.10.19	18/18	**Glawischnig-Piesczek v. Facebook Ireland Limited**	Monitoring in specific cases is not prohibited under Art. 15 of the Directive (§§ 34, 35). Due to the nature of social networks, there is a risk of the reproduction and sharing of illegal information and thus, courts can ask them to block access to the original, identical, or slightly different information containing the same message (§§ 36, 37, 41). Article 15 further does not include a general obligation to seek facts or circumstances indicating illegal activity (§ 47). Further, it does not preclude Member States from ordering the removal or blocking of information worldwide. (§ 53)
15.03.12	292/10	**G v. Cornelius de Visser**	The Member State has jurisdiction in which the service provider is established, if it is unknown, the Directive does not apply. (§§ 71, 72).
16.02.12	360/10	**Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v. Netlog NV**	Provisions leading to filtering systems require general monitoring and thereby, violating Article 15(1) of the Directive. (§ 38). Such an injunction could potentially undermine FoE as it might not distinguish between lawful and unlawful content (§ 50) and thereby, such a provision would not respect the balance with FoE (§ 51).
24.11.11	70/10	**Scarlet Extended SA v. Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM)**	Preventive filtering requires active monitoring and thus, violates Art 15 of the Directive (§ 40) and it would violate FoE if it does not distinguish between lawful and unlawful content (§ 52). Thus, provisions leading to filtering violates FoE (§ 53).

Table 2. *cont.*

Date	No.	Name	Relevance
25.10.11	509/09, 161/10	**eDate Advertising GmbH and Others v. X and Société MGN LIMITED**	The Directive is not intended to achieve harmonisation but rather a coordinate field (§ 57). Art 3 of the Directive must be interpreted that free movement of information society services between Member States is guaranteed (para 65). Thus, Member States must ensure that the provider of an electronic commerce service is not made subject to stricter requirements than those provided in the Member State law following the Directive. (§ 68).
12.07.11	324/09	**L'Oréal and Others v. eBay International AG and Others**	Awareness mentioned in Art. 14 of the Directive means that an intermediary uncovered illegal information through investigation or in specific cases, notification (§ 122). Measures cannot require online service providers to actively monitor all the data of each customer to prevent future intellectual property rights infringements (§ 139).
23.03.10	236/08 to 238/08	**Google France SARL and Google Inc v. Louis Vuitton Malletier SA and Others**	Internet referencing service providers such as Google, fall under Art 14 of the Directive and thus, the exemptions of liability if they do not have played an active role in the data stored (§ 120).

*cases included were regarded provisions in Directive 2000/31/EC (referred to as Directive), namely Article 12-15, and Freedom of Expression (Article 11 CFR).

1.2. European Court of Humana Rights (ECtHR)

Table 3. *Selection of Relevant ECtHR case-law*

Date	No.	Name	Relevance
07.06.22	32401/10, 44285/10, 3488/11, and Others	**Taganrog Lro and Others v. Russia**	The blocking of a Jehovah's Witnesses' website violates Art. 10 as it constitutes an interference by a public authority (§224)
16.11.21	41055/12	**Assotsiatsiya Ggo Golos and Others v. Russia**	The dissemination of materials on websites amounts to the exercise of FoE (§70) and can result in a "public watchdog" role and thereby being comparable to the role of the press (§76).
02.09.21	45581/15	**Sanchez v. France**	Courts can hold Facebook accounts holders liable if they fail to take down unlawful (in this case, antisemitic) third-party comments in their comment section despite their lack of knowledge (§ 99)
18.05.21	43351/12	**Ooo Informatsionnoye Agentstvo Tambov-Inform v. Russia**	The internet's important role for FoE is especially important in election periods (§84).
23.05.20	20159/15	**Bulgakov v. Russia**	The blocking of websites is an extreme measure comparable to banning newspapers or television stations as it disregards the distinction between legal and illegal content (§34)

Table 3. *Cont.*

Date	No.	Name	Relevance
23.06.20	61919/16	**Engels v. Russia**	The blocking of information about technologies (which are content-neutral) interferes with access to all content which can be accessed through them (§30). The legal framework of countries needs to establish safeguards able to protect users from "excessive and arbitrary effects of sweeping blocking measures" (§34).
23.06.20	10795/14	**Vladimir Kharitonov v. Russia**	The blocking of websites is an extreme measure comparable to banning newspapers or television stations as it disregards the distinction between legal and illegal content (§38) Indiscriminate blocking which interferes with lawful content as a collateral effect of measures aimed at illegal content results in the arbitrary interference with the rights of owners of such content or websites (§48).
26.03.20	44229/1	**Pendov v. Bulgaria**	Websites can constitute a means of exercising FoE (§54). The retention of a website by the police followed by limited functionality is interference by a public authority to FoE (§59).
30.04.19	48310/16, 59663/17	**Kablis v. Russia**	The blocking of a social media account without a legal framework with precise rules or judicial review was found to be a violation of Art 10 (§§92, 97, 106, 107)
13.03.18	35285/16	**Nix v. Germany**	The interference with FoE through the conviction of a German blogger for a post with a picture of a Nazi official with a swastika was proportionate, legitimate, and necessary. (§§54-56)
02.02.16	22947/13	**Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary**	Internet Service providers and large news portals enjoy the same principles apply to the press (§61). Internet news portals have duties and responsibilities regarding third-party content (§62). There is a difference if the intermediary has economic interests or not (§64). The court found a violation of Art 10 as the notice-and-take-down system of the intermediary is an effective measure to balance rights (§91).
01.12.15	48226/10, 14027/11	**Cengiz and Others v. Turkey**	YouTube is an important means of exercising FoE which is fostering the emergence of citizen journalism (§52) and blocking it violated Art 10 (§66).
15.10.15	27510/08	**Perinçek v. Switzerland**	Little scope under Article 10(2) to restrict political expression or a debate on questions of public interest (§197), hate speech as a restriction of Art 10 (§205)
16.05.15	64569/09	**Delfi AS v. Estonia**	The audio-visual media often has a more immediate and powerful effect on FoE than the print media (§134). The risk of the internet creating harm through posted content, in particular for the right to respect for private life, is higher than the risk posed by the press (§133). Internet news portals can be held liable for failing to take down unlawful comments (containing hate speech or incitements to violence), in this case, due to insufficient measures (i.e., word-based filters (§ 156) and content moderators (§ 157))
22.04.13	48876/08	**Animal Defenders International v. The United Kingdom**	Debates on the question of public interest are allowed little restrictions under Art 10(2) (§102)
18.12.12	3111/10	**Ahmet Yildirim v. Turkey**	The internet has special importance for the exercise of Art 10 (§49). The dissemination of information and the public receiving them reception also fall under Art 10 (§50) and therefore, the blocking of websites on Google is a violation of Art. 10 (§§ 55, 69)

Table 3. *Cont.*

Date	No.	Name	Relevance/Content
08.11.2012	43481/09	**PETA Deutschland v. Germany**	Balancing with personality rights: Referencing the Holocaust (considering the historical context of Germany) was deemed important (§49) and the sanction of the prevention of publishing was proportionate (§50) to restrict FoE
15.03.12	4149/04 41029/04	**Aksu v. Turkey**	Balancing with Art 8: vulnerable position of Roma/Gypsies (§75)
09.02.12	1813/07	**Vejdeland and Others v. Sweden**	Discrimination based on sexual orientation is as serious as discrimination based on race, origin or colour (§55).
12.10.10	28999/03	**Timciuc v. Romania**	Balancing with Art 8: no hierarchical relationship (§144)
30.06.09	32772/02	**Verein gegen Tierfabriken Schweiz (VgT) v. Switzerland**	Positive obligation by states to organise the execution of the Court's judgements (§97) (the Court's previous judgement ordered the airing of a commercial for animal protection, which Switzerland failed to fulfil)
10.03.2009	3002/03, 23676/03	**Times Newspapers LTD (Nos. 1 and 2) v. The United Kingdom**	Art 10 covers the right to impart information and for the public to receive it, whereby the internet plays an important role in enhancing the public's access (§ 27). Internet archives have a secondary role (to the press) in maintaining and making archives available to the public (§45).
16.03.00	23144/93	**Ozgur Gundem v. Turkey**	Text passages that advocate for intensifying armed struggle, glorify war and "espouse the intention to fight the last drop of blood" and be regarded as encouraging violence and therefore, can be restricted under Art 10(2) (§65)
08.07.99	23556/94	**Ceylan v. Turkey**	Little scope under Art 10(2) for political speech or debates on matters of public interest (§34)
08.07.99	26682/95	**Sürek v. Turkey**	The sensitive situation in South-East Turkey allowed measures to restrict FoE due to the protection of national security, territorial integrity and the prevention of disorder and crime (§52). Stigmatization, context, stirring up hatred, and possible risk of physical violence are relevant reasons to interfere with FoE (§62)

Appendix 2 Content Moderation

Table 4. *ToS, Hate Speech and Enforcement Rates*

	Meta	**Google**	**Twitter**	**TikTok**
	Community Standards	Community Guidelines	Twitter Rules	Community Guidelines
Hate Speech				
Defined Hate Speech Attributes	Race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease	Age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims of a major violent event and their kin, veteran status	Race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease	Race, ethnicity, national origin, religion, caste, sexual orientation, sex, gender, gender identity, serious disease, disability, immigration status
Exceptions	newsworthy, significant, important content for the public interest	Educational and newsworthy content, content of public interest	Content of public interest (especially regarding politics)	Educational, documentary, scientific, artistic, or satirical content, counter-speech, content of social importance
NetzDG Reports				
Rate of § 130 StGB action*	7.76%	22.77%**	8.87%	21.53%
Percentage within 24 hours*	92.50%	88.37%**	97.05%	96.66%
More than 7 days*	1.90%	0.10%**	<0.00	<0.00

*i.e., the rate of removed content classified as hate speech divided through reported hate speech, Calculated with data from the July 2021 – December 2021 NetzDG reports.[405]

**not following §130 but rather classified content as "hate speech or political extremism"

[405] 'Removals under the Network Enforcement Law', Google Transparency Report; Twitter, 'Twitter Netzwerkdurchsetzungsgesetzbericht: Juli - Dezember 2021', 'NetzDG Transparenzbericht. Januar 2022', 2022; 'NetzDG Transparenzbericht', TikTok, November 2021.

Table 5. *Officially reported content moderation techniques*

		Meta	**Google**	**Twitter**	**TikTok**
Hard Control	Geo-blocking	Not Mention	Yes[406] (possible also for the user to determine)	Yes[407]	No Mention
	Keyword Filters	No Mention	No Mention	No mention	No Mention
	Flagging	Yes	Yes, "Trusted Flagger Programme"	Yes	Yes
	Content Moderators	15.000	Intelligence Desk	Yes, no number is publicly available.	Yes, no number is publicly available.
Soft Control	Algorithms	Identification of violating content[408] Personalization [409]	Identification of violating content[410], trends[411], personalization [412]	Identification of violating content[413], trends[414], personalization [415]	Personalization [416]
Measures		Post-Level: Take down, limiting visibility*, labelling**, Account level: restricting accounts, removal	Post-Level: Take-down, limiting visibility*, Age-restricted content Account-Level: Temporary ban, removal	Post-level: Take-down, limiting visibility*, labelling**, demanding removal Account-Level: requiring edits, read-only mode, temporary ban, removal Message-Level: stopping conversations, notice	Post-Level: Take-down, limiting visibility* Account-Level: Temporary bans, removal
Appeal		'Disagree with decision', review request, Oversight Board	Appeal form for strikes, video and account removal[417]	Appeal form for removed tweets or suspended accounts[418]	Appeal button for account and post deletion [419]

*includes making content ineligible for recommendation in feeds, trends, or search results.
** involves usually providing of content, warnings, or fact-checking,

[406] 'Block Videos in Specific Territories', YouTube Help.
[407] 'Understanding When Content Is Withheld Based on Country', Help Center.
[408] Mike Schroepfer, 'Update on Our Progress on AI and Hate Speech Detection', *Meta,* February 2021.
[409] 'How Facebook Distributes Content', Meta Business Help Center.
[410] 'YouTube Community Guidelines Enforcement FAQs', Transparency Report Help Center.
[411] 'Trending on YouTube', YouTube Help.
[412] 'Manage Your Recommendations and Search Results', YouTube Help.
[413] "proactive detection", 'A Safer Twitter'.
[414] 'Twitter Trends FAQ', Help Center.
[415] 'Personalized Content Based on Your Third-Party Web Activity', Help Center.
[416] 'How TikTok Recommends Videos #ForYou', TikTok, August 2019.
[417] 'Appeal Community Guidelines Actions', YouTube Help.
[418] 'Our Range of Enforcement Options for Violations', Twitter Help.
[419] 'Account Safety', TikTok Help Center.

Appendix 3: Cases of Digital Repression

Table 6. *Cases of Digital Repression*

Cell	Country	Year	Details
1 + 5	Belarus	2020	Internet shutdown during a protest in Belarus, a total of 61 hours[420]
		2017	No further information[421]
	Montenegro	2016	No further information[422]
	Turkey	2020	Internet shutdown during an attack on Turkish troops in Syria[423]
		2019	Twitter, Facebook, WhatsApp, and Instagram were restricted in Southern Turkey as military operations were launched in Syria, likely to protect the troops.[424]
		2017-2020	Turkey blocked Wikipedia for almost 3 years until the Constitutional Court rules that the blocking was a violation of FoE[425]
	UK		Internet Shutdown during a protest in London
	Ukraine	2022	in Kharkiv (Ukraine-controlled) and Mariupol and Donetsk on 24.02 and the following months. [426] Before the invasion, there have also been reports of significant internet disruptions.[427] (likely to be foreign influences)
		2017 -	Blocking of popular social networks (VKontakte or Mail.ru) and one email service by decree part of sanctions against Russian companies[428]
	Palestine/ Israel	2018	Israel requested the removal of 14.283 posts, 86% were voluntarily granted by social media companies (namely Facebook, Twitter, and Google)[429] (also cell 9)
		2015	Israel's Cyber Unit worked with social media platforms to block, delete and censor Palestinians' content[430] (also cell 9)
	Russia	2022	Instagram, Twitter, and Facebook restrictions admit the Ukraine war before Russia declared Meta an "extremist organization"[431][432]
		2019	No further information[433]
		2019	targeted internet shutdown in Russia on Saturday 3 August as protestors took to the streets to protest political oppression[434]

[420] Belarus Internet Observatory, 'Internet shutdown in Belarus', 2020; Kelvin Chan, 'Internet Shutdowns a Growing Tool to Halt Protests', St. Lousi Post-Dispatch, February 2021.
[421] AccessNow, '#KeepItOn STOP Data 2016-2021', Google Docs, 2021.
[422] AccessNow.
[423] 'Social Media Blocked in Turkey as Idlib Military Crisis Escalates', *NetBlocks*, February 2020.
[424] 'Twitter, Facebook, WhatsApp and Instagram Restricted in Southern Turkey', *NetBlocks*, October 2019.
[425] 'Turkey Restores Access to Wikipedia after 991 Days', *NetBlocks*, January 2020.
[426] 'Internet Disruptions Registered as Russia Moves in on Ukraine', *NetBlocks*, February 2022.
[427] 'Mobile Internet Disrupted in Luhansk, Ukraine amid Heightened Tensions with Russia', *NetBlocks*, February 2022.
[428] Alec Luhn, 'Ukraine Blocks Popular Social Networks as Part of Sanctions on Russia', *The Guardian*, May 2017.
[429] Viki Auslender, 'Shomrim - Israel's Censorship Meets Facebook's Compliance', Shomrim. The Center for Media and Democracy, November 2020.
[430] 7amleh, 'Facebooks Content Regulation Between "Hate Speech" and Legitimate Political Expression as Freedom of Speech: Bias and Discrimination against Palestinians'.
[431] 'Instagram Restricted in Russia as Online Space Continues to Shrink', *NetBlocks*, March 2022.
[432] 'Twitter and Facebook Restricted in Russia amid Conflict with Ukraine', *NetBlocks*, February 2022; 'Instagram Restricted in Russia as Online Space Continues to Shrink'.
[433] AccessNow, '#KeepItOn STOP Data 2016-2021'.
[434] 'Evidence of Internet Disruptions in Russia during Moscow Opposition Protests', *NetBlocks*, August 2019.

Table 6. *cont.*

Cell	Country	Year	Details
		2018	Ban of the app Telegram in 2018 because the app would not allow the government to read users' text messages[435]
	Worldwide	2021, 2019	Facebook, What-App, Instagram, and Messenger due to internal problems in 2021[436] and 2019[437]
		2019	YouTube, Gmail, and Google services were down in multiple countries due to technical errors[438],
		2019	Wikipedia was down[439]
Cell 9	Germany	2021	Facebook removed 150 accounts and pages linked to anti-lockdown
		2021	demonstrators in Germany due to misinformation short before the national elections[440]
			Around 150 Facebook accounts were deleted by Meta to stop misinformation regarding Covid-19
		2019	11 accounts of the police from north rhine Westphalia were blocked due to "suspicious activities"[441]
		2018	AFD's deputy-later, Germany's right-wing party, Beatrix von Storch's tweet about the police of cologne pandering to "gang-raping hordes" of Muslim men because of them tweeting a post in Arabic.[442] This post violated the new NetzDG law in Germany which requires platforms to take down content online
	Myanmar	2021	Facebook refused to disclose data on posts which showed anti-Rohingya violence in Myanmar[443]
		2018	Facebook admitted that it failed to "prevent [Facebook] from being used to foment division and incite offline violence"[444] the statement was criticised due to the lack of recommendations on implementation mechanisms[445]
		2017	Rohingya activists being deplatformed[446]
	Palestine	2021	Over 500 violations of digital rights of Palestinians primarily on Facebook and Instagram, posts were taken down due to them "violating ToS[447][448]

[435] Neil MacFarquhar, 'Russian Court Bans Telegram App After 18-Minute Hearing', *The New York Times*, April 2018.
[436] 'Facebook, WhatsApp, Instagram, Messenger down Globally in Extended Service Outage', *NetBlocks*, October 2021.
[437] 'Facebook, WhatsApp, Instagram, Messenger down Globally', *NetBlocks*, April 2019.
[438] 'YouTube, GMail and Google Services down in Multiple Countries', *NetBlocks*, December 2019.
[439] 'Wikipedia Outage with International Impact Detected', *NetBlocks*, May 2019.
[440] '"The Big Delete:" Inside Facebook's Crackdown in Germany', September 2021.
[441] Marie Bröckling, 'Overblocking auf Twitter: Polizisten können zwei Tage lang nicht twittern', netzpolitik.org, December 2019.
[442] Cristina Maza, 'Twitter and Facebook Shut down Anti-Muslim Posts by Far-Right Alternative for Germany Party', Newsweek, January 2018.
[443] 'US Judge Orders Facebook to Release Anti-Rohingya Account Records', September 2021.
[444] Alex Warofka, 'An Independent Assessment of the Human Rights Impact of Facebook in Myanmar', *Meta*, November 2018.
[445] Alexandra Stevenson, 'Facebook Admits It Was Used to Incite Violence in Myanmar', *The New York Times*, November 2018.
[446] Betsy Swan, 'Exclusive: Facebook Silences Rohingya Reports of Ethnic Cleansing', *The Daily Beast*, September 2017.
[447] Kelly Kunzl, 'Big Tech Censors Palestinian Advocacy around the World, While Fostering Surge in Jewish Extremism in Israel', Mondoweiss, May 2021; Sophia Hyatt, 'Facebook "Blocks Accounts" of Palestinian Journalists', Aljazeera, September 2016.
[448] 7amleh, 'Facebooks Content Regulation Between "Hate Speech" and Legitimate Political Expression as Freedom of Speech: Bias and Discrimination against Palestinians'.

Table 6. *cont.*

Cell	Country	Year	Details
		2018	Israel requested the removal of 14.283 posts, 86% were voluntarily granted by social media companies (namely Facebook, Twitter and Google)[449] (also cell 1)
		2015	Israel's Cyber Unit worked with social media platforms to block, delete and censor Palestinians' content[450] (also cell 1)
	Syria	2017	Over one hundred YouTube channels reporting on the Syrian civil war were deleted.[451]
		2015	Facebook deleted dozens of accounts of peaceful protestors posting about the civil war in 2014 including pages such as the Syrian Network for Human Rights, a London-based NGO[452] Brown Moses found that 78% of Facebook pages in his sample were connected to YouTube pages that posted information about the August 21 Sarin attacks in Damascus[453][454]
	Ukraine	2022	Anonymous account[455] deleted posting about the Russia-Ukraine war
	United States	2021	US President Trump was permanently banned on Twitter-[456]
		2017	Occupy wall street activists' Twitter accounts were suspended (approx. 80 accounts with an accumulated 5 million followers)[457] due to them having similarities to bot accounts, Facebook deleted pages such as the "Free Thought Project" a free speech page with 3.1 million followers and the "end the drug war" (460.000 followers) due to "inauthentic activity"[458], Twitter suspended the account of the anti-media project and its Facebook page with 2.1 million followers[459], the Twitter account of global revolution live with 55.000 followers was suspended.[460]
	Worldwide	2022	Tech companies[461] and YouTube[462] ban Russian state media (RT, Sputnik etc.) because of disinformation
		2021	Twitter banned 70.000 QAnon Accounts[463]
			Deplatforming of sex workers and activists[464]

[449] Auslender, 'Shomrim - Israel's Censorship Meets Facebook's Compliance'.
[450] 7amleh, 'Facebooks Content Regulation Between "Hate Speech" and Legitimate Political Expression as Freedom of Speech: Bias and Discrimination against Palestinians'.
[451] The Associated Press, 'Activists Worry YouTube Erasing Proof of Syria Atrocities', CBS News, September 2017.
[452] Michael Pizzi, 'The Syrian Opposition Is Disappearing From Facebook', The Atlantic, February 2014.
[453] On 21st August 2013, an attack with chemical weapons in Damascus killed hundreds of Syrians. 'Attacks on Ghouta: Analysis of Alleged Use of Chemical Weapons in Syria' (Human Rights Watch, September 2013).
[454] It should be noted that the sample is very small, but his observations are in line with other reports about the deleting of information about the Syrian civil war. Brown Moses, 'How Facebook Is Destroying History - A Survey Of August 21st', *Brown Moses Blog*, February 2014.
[455] Anonymous [@youranonnews], 'Activist Accounts Being Suspended by @twitter b/c of Leaked Info about Russia', Tweet, *Twitter*, March 2022.
[456] Brian Fung, 'Twitter Bans President Trump Permanently', CNN, January 2021.
[457] Ben Bours, 'Facebook's Hate Speech Policies Censor Marginalized Users', Wired, August 2017.
[458] Sanjana Varghese, 'Twitter Has Purged Left-Wing Accounts with No Explanation', *Wired UK*, accessed 26 May 2022.
[459] Varghese.
[460] Varghese.
[461] Rebecca Klar, 'Tech Companies Seek to Choke out Russian State Media', Text, *The Hill*, March 2022.
[462] Natasha Lomas, 'YouTube Geoblocks Russia Today, Sputnik Channels in Europe', Tech Crunch, March 2022.
[463] Pavia, Will, 'Twitter Bans 70,000 QAnon Accounts in US Social Media Crackdown', *The Times*, January 2021.
[464] Danielle Blunt and Zahra Stardust, 'Automating Whorephobia: Sex, Technology and the Violence of Deplatforming: An Interview with Hacking//Hustling', *Porn Studies* 8, no. 4 (October 2021): 350–366.

Table 6. *cont.*

Cell	Country	Year	Details
			A journalist by the Belfast Telegraph reporting on Dublin crime boss Daniel Kinahan [465]
			Extremist researcher Gwen Snyder was locked out of her account due to her writing tweets about a Proud Boys rally as well as many other extremism researchers mostly due to posts about groups planning assaults, or footage of their rallies etc.; Twitter responded that the deletion of Snyder was a mistake[466]
		2019	Twitter banned a feminist who commented on a new law about transgender rights[467]
		2018	Tweets about feminist research were deleted[468]
		2017	Jewish Activist Daneil Sieradski's account was suspended [469]
Cell 2 + 6	China	2019	TikTok censoring content against Chinese interests and the distortion about other countries' histories (e.g., the May 1998 riots in Indonesia or the Tiananmen Square incidents).[470]
	Philippines	2021	Furthermore, there have been reports about DDoS attacks against Philippine media outlets and human rights groups, which reportedly could have been tied to the government of the Philippines.[471]
	Ukraine	2022	DDoS attacks on two banks, the Ministry of Defence and Ukraine's armed forces which "bore traces of foreign intelligence services"[472]
	Worldwide	2022	TikTok wordfilters (related to LGBTQI+ (e.g., gay, homo, LGBTQ, LGBTQI, queer), sex (prostitution, porn, sex, sex work) and extremist content (e.g., national socialism, terrorist))[473]
		2019	Wikipedia DDoS Attack: Wikipedia was down for about 9 hours reported by Wikipedia Germany, origin of the attack was not established[474]
Cell 10	China	2022	Posts about Uighurs were reportedly suppressed by TikTok[475] which has been refuted by TikTok[476]
	United States	2018	Twitter has been accused of shadowbanning republicans in the US which was fixed overnight[477], however, there has been some criticism of this claim as this technically was not a shadowban[478]
	Worldwide	2021	51.28% of sex workers and AOP reported being shadowbanned on social media [479]

[465] Ralph, 'Twitter Admits It Banned Belfast Telegraph Political Editor in Error for the Second Time in One Month', Independent.ie, May 2021.
[466] Ali Breland, 'Twitter's New Privacy Policy Makes It Harder to Spread Warnings about Online Fascists', *Mother Jones*, December 2021.
[467] 'Twitter-Ban Feminist Defends Transgender Views Ahead of Holyrood Meeting', *BBC News*, May 2019.
[468] petey, 'Twitter Suspended Me for Tweeting Feminist Academic Research. Here's Why That's a Problem. – MIT Center for Civic Media', *Center for Civic Media MIT*, September 2018.
[469] Sam Kestenbaum, '"Antifa's Most Prominent Jew" Booted From Twitter', The Forward, June 2017.
[470] Hern, 'Revealed'.
[471] 'Investigation of DDoS Attacks against Independent Media Shows Links to Philippine Government and Army – Qurium Media Foundation', *Qurium*, July 2021.
[472] Adam Satariano, Paul Mozur, and Valerie Hopkins, 'Shutdowns Of Internet Offer Lesson On Conflicts', The New York Times, February 2022.
[473] Translated from German. tagesschau.de, 'TikTok nutzt in Deutschland Wortfilter', tagesschau.de.
[474] 'Wikipedia Disrupted Globally in Apparent Denial of Service Attack'.
[475] 'Nervous TikTok'.
[476] Nicholas, 'Shadowbanning Is Big Tech's Big Problem'.
[477] Alex Thompson, 'Twitter Appears to Have Fixed "Shadow Ban" of Prominent Republicans like the RNC Chair and Trump Jr.'s Spokesman', *Vice*, July 2018.
[478] Stack, 'What Is a "Shadow Ban," and Is Twitter Doing It to Republican Accounts?'
[479] Blunt and Stardust, 'Automating Whorephobia'.

Table 6. *cont.*

Cell	Country	Year	Details
		2021	An autoethnographic study about an Instagram account about pole dancing being deemed as "vaguely inappropriate content"[480]
		2020	Black lives matter was reportedly suppressed on TikTok[481] with TikTok claiming that this is primarily due to issues with view count displays making it seem as if the videos were suppressed (#BlackLivesMatter and #George Floyd) leaving one creator's video drop views from 75.000 to 1.5000 on the BLM post.[482]
Cell 3 + 7	China	2016	sovereignty movements from Tibet and Taiwan have been drowned out by bots promoting Chinese ideals.[483] The hashtags #tibet and #freetibet have reportedly been flooded by junk tweets from automated accounts.
	Argentina	N/A	Bot usage to demobilize protests[484]
	Bahrain	N/A	Protest Movements were "bombed" on Twitter to prevent protest organization[485]
	China	2012	Spamming of the hashtag #freetibet[486] w2ith 73% Spam tweets
	France	2017	Marine LePen was supported by Russian bots in spreading anti-Macron hashtags.[487]
	Italy	2017	Before the 2017 elections, the party Lega encouraged its followers to turn into 'selfbots' (i.e., bots created by the account holders) amplifying Lega's message.[488]
	Iran	N/A	Protest Movements were "bombed" on Twitter to prevent protest organization[489]
	Mexico	2016	"Peñabots, named after the Mexican President Enrique Peña Nieto, have also been used to send out pro-government propaganda"[490]
		2012	The hashtag #marchaAntiEPN (March against EPN, a presidential candidate) was flooded with around 62% spam tweets[491]
	Morocco	N/A	Protest Movements were "bombed" on Twitter to prevent protest organization[492]
	Russia	N/A	Bot usage to demobilize protests[493]
	South Korea	2012	During the presidential elections, 1.2 million Twitter messages were published from the National Intelligence Service in favour of candidate Park Geun-Hye, who ended up winning the election.[494]

[480] Carolina Are, 'The Shadowban Cycle: An Autoethnography of Pole Dancing, Nudity and Censorship on Instagram', *Feminist Media Studies*, May 2021, 1–18.
[481] McCluskey, 'These Creators Say They're Still Being Suppressed for Posting Black Lives Matter Content on TikTok'.
[482] McCluskey.
[483] Woolley, 'Automating Power'.
[484] Woolley.
[485] Woolley.
[486] Verkamp and Gupta, 'Five Incidents, One Theme', 2.
[487] Nimmo, 'Russian and French Twitter Mobs in Election Push'.
[488] Nimmo, '#ElectionWatch'.
[489] Woolley, 'Automating Power'.
[490] Woolley.
[491] Verkamp and Gupta, 'Five Incidents, One Theme'.
[492] Woolley, 'Automating Power'.
[493] Woolley.
[494] Sang-Hun, 'Prosecutors Detail Attempt to Sway South Korean Election'.

Table 6. *cont.*

Cell	Country	Year	Details
	Syria	2011	Protest Movements were "bombed" on Twitter to prevent protest organizations During the war, for example, the hashtag #Syria was flooded with messages and pictures about the beauty of Syria.[495] Further, some tweets highlighted natural disasters in other parts of the world.[496] (not clear from whom) Some smoke-screening tweets talked about Syria but were not related to the civil war going on- They, for example, posted about Syrian short films.[497]
	Turkey	N/A	Bot usage to demobilize protests[498]
Cell 4 + 9	Great Britain	2014	(Through Snowden leak)[499] The GCHQ's secret unit, the Joint Threat Research Intelligence Group JTRIG with two tactics: "(1) to inject all sorts of false material onto the internet to destroy the reputation of its targets and (2) to use social sciences and other techniques to manipulate online discourse and activism to general outcomes it considers desirable"[500] including fake victims blog posts, the posting of negative information, discrediting targets by, for instance, emailing their colleagues, discrediting companies (by stopping deals or ruining business relationships).
	Germany	2022	Misinformation about refugees in Germany[501]
	Hong Kong	2019	Twitter deleted 936 accounts due to them "deliberately and specifically attempting to sow political discord in Hong Kong" during the protests and about 200.000 accounts due to them violating ToS (Spam, Coordinated Activity, Fake Accounts, Attributed Activity and ban evasion).[502]
	Iran	2018	700 troll accounts could have been liked to Iran [503]
	Turkey	2015	After a Russian plane was shut down by a Turkish fighter jet, bots posted real and fake news about the incident.[504]
	Russia	2022	RT, Sputnik, and other Russian State media dispensing disinformation[505]
		2015 -	"Putinbots"[506], More than 100 Twitter accounts using the #IStandWithPutin violating the "platform manipulation and spam policy" likely being bots
	Ukraine	2017	The official state media outlet of the unrecognized Donetsk People's Republic in Ukraine posted fake news about 3.600 US tanks being sent by the US and therefore NATO.[507] In reality, tanks were sent, but about $\frac{1}{20}^{\text{th}}$ of what the article implied. The story was slowly picked up by other news outlets, finally reaching websites in the US, Canada, and Europe.

[495] York, 'Syria's Twitter Spambots'.
[496] For instance, some bots posted pictures of the consequences of Hurricane Sandy in the United States. Abokhodair, Yoo, and McDonald, 'Dissecting a Social Botnet', 849.
[497] Abokhodair, Yoo, and McDonald, 'Dissecting a Social Botnet'.
[498] Woolley, 'Automating Power'.
[499] Greenwald, 'Exclusive'.
[500] Greenwald, 'How Covert Agents Infiltrate the Internet to Manipulate, Deceive, and Destroy Reputations'.
[501] Karolin Schwarz and Lutz Helm, 'Hoaxmap'.
[502] Twitter Support, 'Information Operations Directed at Hong Kong'.
[503] Matt Burges, 'We Finally Know the Full Extent of Russia's Twitter Trolling Campaign', *Wired UK*, October 2018.
[504] Stukal et al., 'Detecting Bots on Russian Political Twitter', 316.
[505] Klar, 'Tech Companies Seek to Choke out Russian State Media'.
[506] Soldatov and Borogan, 'What Spawned Russia's "Troll Army"?'
[507] Ben Nimmo, 'Three Thousand Fake Tanks', *Medium*, January 2017.

Table 6. *cont.*

Cell	Country	Year	Details
	United States	2016	Russia interference in the Brexit referendum: Researchers found a significant amount of anti-EU biased articles disseminated by primarily RT and Sputnik (kremlin aligned media) with about 134 million potential impressions.[508] Facebook removed accounts which posted frequently content related to "anti-NATO sentiment, protest movements and anti-corruption"[509].[510]
Cell 12	France	2017	The hashtag #MacronLeaks reached 47.000 tweets within just three hours two days before the 2017 elections. The hashtag led to supposed leaked emails from Macron. the Digital Forensic Research Lab found the origin of the news to be an alt-right US American whose tweets were reposted by many bots.[511] 10 of the most active accounts tweeting the hashtag, posted over 1.300 tweets in about 3 hours, hinting toward them being bots.
	United States	2018	Cambridge Analytica build voter profiles from Facebook data to influence political campaigns[512] (with possible inference from Russia, thus, also cell 4+9)
		2016	the New York Times found numerous individuals posting fake news about the upcoming US national election. Most of them run websites publishing (often by stealing) articles that report positively on Trump, often with much misinformation. They publish news that drives up traffic to their website, thus, posts about Trump, and anti-Muslim and anti-Mexican content.[513]

[508] Digital, Culture, Media and Sport Committee, 'Disinformation and "Fake News": Final Report' (London, United Kingdom: House of Commons, 2019), para. 243.
[509] Digital, Culture, Media and Sport Committee, para. 245.
[510] Shane, 'The Fake Americans Russia Created to Influence the Election'.
[511] Nimmo, 'Hashtag Campaign'.
[512] Confessore, 'Cambridge Analytica and Facebook'.
[513] Higgins, McIntire, and Dance, 'Inside a Fake News Sausage Factory: "This Is All About Income"'.