Universiteit Utrecht

Master Thesis

Shifting responsibilities?

Understanding implications of platform regulation by analyzing the discourse in light of the EU *Digital Services Act*

by

Arnika Zinke

Student number: 6842763
MA New Media and Digital Culture

Supervised by Dr. Michiel de Lange
Reader: Dr. Mirko Tobias Schäfer

Citation: Chicago Manual of Style (17th) 2017. “B / Author Date Referencing”
Word count: 11,350

Date: 14 June 2020

Table of Contents

Abstract
This research aims to understand the main tensions around platform regulation in the European Union, in specific regards to hate speech moderation. In deploying critical discourse analysis by Gee, I want to understand how three different stakeholder groups (European Commission, Facebook and third-party-actors OHCHR and EDRi) voice their opinion around this topic. In doing so, I aim to establish a nuanced approach towards discourse on platform regulation that showcases tensions around governance procedures. As the analysis shows, actors in the discourse do not necessarily argue about the concept of regulation but specific governance-based solutions (that could be possibly regulated). Here, the question of responsibility was established as one of the focal-points of the discussion: Who should be given the responsibility to police speech on platforms and to what extent are guidelines influenced by regulatory bodies? In this regard, the Commission as well as Facebook showcased a certain tendency towards techno-determinist approaches, while third-party stakeholders (EDRi and OHCHR) emphasized the importance of human agency. Finally, this research aims to expand on the existing academic debate on platform-governance that takes the current discussion and standpoints of the actors into account.

Keywords: platform regulation, platform governance, hate speech, content moderation, intermediary liability, European Union, E-Commerce Directive, Digital Services Act

Glossary

Due to the partly heavy use of possibly complicated (policy-)language, the following important terms will be introduced briefly. In alphabetical order.

(Commercial) Content moderation: Corporate evaluation of user-generated content posted on platforms. Conducted by humans and artificial intelligence (depending on the content and platform). Please refer to chapter 2.2.2 and 2.3 for detailed description.

Digital Services Act (also DSA): The upcoming regulatory framework for digital services in the European Union. Expected to be presented by the end of 2020, clarifying, among others, platform responsibilities and taking a stand on intermediary liability (exemptions).

E-Commerce Directive: Current (20-year-old) legal framework determining how to regulate platforms (among others) in the European Union.

EDiMA: Brussel-based lobby-group representing the interests of online platforms, such as Facebook, Google, Microsoft, Apple or Amazon. Self-proclaimed "pro-consumer"-oriented organization, promoting "innovation and growth" towards a Digital Single Market for Europe through constant dialogue with EU-institutions.

EDRi: European Digital Rights (EDRi) is an association of civil and human rights organizations from across Europe based in Brussels. Key priorities lie in defending digital rights in areas such as privacy, surveillance, net neutrality and copyright reform.

European Commission: Supranational organ of the European Union, nominated by the European Council, confirmed by the European Parliament. Responsible for proposing legislation, upholding EU treaties. Since 2019, Ursula von der Leyen is the Head of the Commission.

Intermediary liability: (Extent of) liability for intermediaries, also referred to as platforms. In the European Union as well as other global markets, such as the United States, intermediaries enjoy exemptions from intermediary liability. Please refer to chapter 2.2.1 for detailed explanation.

United Nations High Commissioner for Human Rights (also OCHCR): Leading UN entity on human rights, entrusted by the United Nations General Assembly to promote and protect human rights. Documents referred to in this thesis have been issued as reports to the General Assembly by the *Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression.*

1. Introduction

When in October 2019 the European Court of Justice's (ECJ) Advocate General presented his opinion in the case of "Eva Glawischnig-Piesczek v Facebook Ireland Ltd.", it was not just the Court's audience who listened. The case addressed the question to what extent national courts can order platforms like Facebook to take down content which has been deemed illegal. It has been closely monitored by civil rights groups, corporate representatives, lobbyists and academic scholars. Some even called it "one of the most important Internet law cases in recent years" (Keller 2019, 2). The high interest on the case, which was brought to the ECJ by the former Austrian politician Eva Glawischnig-Piesczek, did not stem from its special contents - but its overall implications on future EU platform policy. The case was seen as possible precedent on EU decisions for a new regulatory framework, the Digital Services Act, which is expected to be finished by the end of 2020.

Institutions such as the European Union, as well as academics (Gillespie 2018b; Van Dijck, Poell, and de Waal 2018; Wagner 2019) have been highlighting the growing impact of platforms on society for years. Platforms such as Facebook have become powerful institutions of public discourse. Therefore, some scholars call the discussion on the platforms' influence a debate long overdue – and highlight the need for a critical discussion about the platforms' influence on our society (Van Dijck, Poell, and de Waal 2018, 3). According to a recent study on incivility[1] in online-discourses, conducted by Kümpel & Rieger, this sense for a need of action can also be confirmed empirically.

Currently the new European Commission under Ursula von der Leyen is working on a framework enforcing regulations of online platforms under the "Digital Services Act" (Rudl and Fanta 2019). Whilst the nature of those regulatory policies is still subject to speculation, the European Commission has been planning to tighten the regulation for a few years now (Frosio 2017, 19). This is also due to the fact that the current policy regulating platforms such as Facebook or YouTube, is part of the E-Commerce-Directive established in 2000, making it 20 years old and therefore partly outdated in terms of current technological standards.

[1] Incivility in online discourses has become increasingly visible because of the growing perception for such content (enhanced through technology), as well as individual personal motives (Kümpel and Rieger 2019, 17). Incivility, meaning "in-civil" or "anti-social" behavior, in this context: inappropriate acting on social media, such as through hate speech, with negative consequences for democracy (Kümpel and Rieger 2019, 25; Stark and Stegmann 2020, 39)

It comes as no surprise that the discussion around new laws and regulation has been heated in the last years with arguments driven by cooperate/ public interests and influenced by different cultural values. They center around the *application of regulation* and to what extent regulative efforts might be at odds with human rights, as well as the *modality of content moderation*, sometimes being depicted as a *matter of AI*, other times a deeply *humanized process*. In this sense, especially the question of *governance* over content moderation plays a big role in the discourse. Statements center around the idea of giving platforms *more responsibility* in content policing, which might make them more liable for content, but at the same time shift the evaluation and removal of public speech further into the hands of private actors. The question of *governance-responsibility* stands in the middle of the overarching discussion on institutional *regulation*, and the practice of corporate *content moderation* – which is conducted by platforms and constitutes the process of enforcing governance-guidelines. Figure 1 showcases how these governance structures relate to each other and how regulation and other forces, such as company interests, influence guidelines that ultimately determine content moderation. In the course of this research, I want to understand the discourse and tension over these governance processes, with special regards to hate speech moderation. The discourse on future European policy on platforms is highly influential in our "platform society" (Van Dijck, Poell, and de Waal 2018) as it might have global implications and a huge impact on big American Internet companies (Keller 2019, 2). By conducting a critical discourse analysis, I want to find out how actors argue around the regulation, which ideas are in the center of controversies and potentially determine how such discussions might have an impact on European policy-making in the future. My research will be guided by the question: *What are the tensions in the discourse on EU-hate speech regulation and how do they reflect on future platform governance of the Digital Services Act? (RQ).* In order to receive an answer to this question, the analysis will be guided by a set of sub-question which will help me in answering the main question piece by piece:

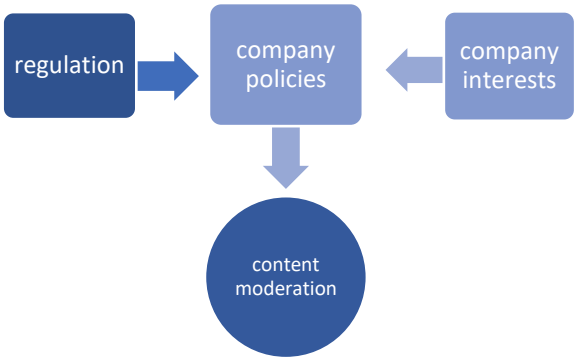

Figure 1 Visual representation of platform governance. Source: By author

SQ1 What are the goals of the stakeholders?

SQ2 To what extent are the topics governance and human agency reflected in the discourse around hate speech regulation?

SQ3 What are the key tensions between stakeholders in the discourse?

SQ4 In what ways are the proposed solutions indicative of a techno-deterministic view on hate speech moderation?

SQ5 What are future possible implications in development of EU platform policy, respectively the Digital Services Act?

During my research I will reflect on different aspects of platform regulation. For the structure of this research this means that I will first explain my theoretical framework, which aims to create an understanding for the academic positioning of platforms, the platform ecosystem, and their societal impact. This includes Van Dijck et al.'s concept of the "platform society" (Van Dijck, Poell, and de Waal 2018). This is followed by an exploration of the concept of governance in relation to platforms and their content moderation practices. Furthermore, the section on "human agency" engages with the recognition of human labor in platform governance practices, also known as "(human) content moderation", and its neglect through the emphasis on artificial intelligence. In the following chapter on methodology, I will explain critical discourse analysis and present my corpus, which consists of documents issued by three stakeholders: the European Union (to be more precise, the European Commission[2]), as regulating force and Facebook, as one of the subjects of regulation. In order to add another set of opinions, this research will be complemented by the statements from third party actors: the digital rights association *EDRi* and the United Nation's *OHCHR* (High Commissioner for Human Rights). In addition, I will provide a detailed description of my analysis, to showcase how I aim to answer my sub-questions and structure the discussion of my results. In order to compare the statements found through the discourse analysis, I have a created a category-set which reflects on both, my theoretical framework and statements examined during the analysis.

[2] For the sake of thoroughness, it is important to clarify that the European Commission is just one body of the European Union and does not stand for a unified European Union opinion on this matter (to the contrary, as the Digital Services Act draft report and its 919 "amendments" submitted by European Parliament members show, see European Parliament 2020a; 2020b). However, as driving force for new regulation, this paper mainly refers to the European Commission's actions and statements. Unless stated otherwise, the Commission's statements might be referred to as "EU" (merely for the purpose of simplification).

Finally, the discussion-chapter will reflect on key subjects (regulation, corporate governance and human agency), and aid me in understanding how stakeholders relate to certain aspects of regulation. Ultimately, this chapter will provide an overview on the key tensions that I have identified and additionally, explore the concept of techno-determinism regarding proposed regulatory solutions. I will use these results in order to explore possible implications on the Digital Services Act, which should serve as a practical addition to the current academic debate on platform governance.

Especially with a Digital Services Act on its way, this thesis could contribute in understanding how discourses shape the development of new digital standards in a platform society. My objective of this research is to create a more nuanced understanding on the stakeholders' goals, but also of the governance procedures that are discussed in this ongoing debate. As the analysis will show, many of the discussed governance practices are hidden behind complicated policy-elements or confusing figures of speech. I believe that a governance-based discussion is only possible when all actors and their objectives are exposed, which is why I want to use this research to untangle the web on platform regulation.

2. Theoretical framework

The following chapters will highlight the theoretical and legal backgrounds of platforms and the implications of technological tools they rely on. It is necessary to comprehend how platforms operate and what role they play in our society. The theories presented will aid me in finding indications of platform regulation, which will later serve as a backbone for the analysis. First, I will look at how platforms have become powerful players of public speech and why regulators see increasing need to govern them. Then I will close in on regulatory efforts so far and the platforms' own governance practices. In doing so, I will specifically highlight the notion of human agency, which will later also help me to analyze my findings from a techno-critical perspective.

2.1 Platform society, ecosystem & corporate ownership

How do media scholars locate platforms within the field, how are platforms structured and what consequences does their corporate ownership and monopole-like power entail? The discourse on hate speech regulation heavily depends on an understanding of these questions, which is why the following section will focus on the theoretical notions on platform society, ecosystem and ownership.

In order to describe the inextricable link between nowadays powerful platforms and modern societies, Van Dijck et al. introduce the concept of "platform society" which refers to the growing dominance of platforms in our everyday lives (Van Dijck, Poell, and de Waal 2018, 2). Similar to Van Dijck et al., Tarleton Gillespie highlights the societal significance of platforms, calling them "architects of public spaces of discourse" (Gillespie 2017, 25). Platforms not only frame the way we talk, but shape the mode of discourse itself: they are powerful enough to swing moods, frame political discourse and shape engagement of users on a global scale (Gillespie 2018a, 23). The influence on the public discourse is precisely why the need for regulation in this regard has risen in recent years. Platforms and their governance methods touch on delicate values such as freedom of expression which is why tools of operation are contested.

Deconstructing the platform society allows us to look closer at the inner workings of platforms[3]. They consist of data, are structured through algorithms and interfaces, are characterized through their ownership and rely on contracts with their users (Van

[3] Van Dijck et al. distinguish two types of platforms: infrastructural and sectoral platforms. Whilst sectoral platforms serve a niche, the real battle for power happens among infrastructural platforms, which contain services from search engines, to data and cloud servers, communication operations like mail and instant messaging, as well social media, analysis software or navigation services, etc. (Van Dijck, Poell, and de Waal 2018, 13).

Dijck, Poell, and Waal 2018, 9). Even though it might seem that there is a huge variety of platforms available, only a few of them share a huge part of the public's attention (Gillespie 2018a, 17). The whole platform "ecosystem" is dominated by "The Big Five"[4], which are characterized by a number of paradoxes that create confusion over their influence: The Internet might seem equal but is actually inscribed in opaque power structures and hierarchy. What should be a place of empowering voices is in fact a marketplace for ad-sellers, what seems free is paid for with the user's own data and finally, the vision of globalized communication is actually inscribed with local, predominantly American, values (Van Dijck, Poell, and de Waal 2018, 12). Understanding these clashes is vital to analyze the discourse on hate-speech regulation. The settings that determine the culture of platforms might fundamentally influence governance procedures. Profit-oriented platforms might inherently think different than institutions, removing content to create a conflict-free environment for their advertising clients (Roberts 2018) – and not a safe, empowering space for its platform users.

The corporate structure of platforms creates a novelty, as technologies of such immense societal impact[5] have never been exclusively in corporate hands before (Van Dijck, Poell, and de Waal 2018, 15). These shifts in ownership might indicate why governments and institutions are struggling to implement rules – the formerly shared power is now in the hand of those whose primarily goal it is to make profit. For the later analysis it is important to understand that such corporate models are deeply inscribed into a platform infrastructure that many people rely on. A few companies dominate a big share of the market and this can lead to concentrated monopolies which governments tend to increasingly worry about.

Adding to that, conflicts between dominant US-platforms and institutions might also be fueled by their differences in culture and ideology (Van Dijck, Poell, and de Waal 2018, 4). According to Van Dijck, Poell and de Waal, many American platforms show an influence of libertarianism, meaning that citizens are left with the major part of responsibility (and thus handing little of it off to the state). Whilst a lot of European countries tend to favor a cooperation-based model that includes individuals, non-governmental-actors and the state - where institutions guard the public from potentially harmful corporate interests (Van Dijck, Poell, and de Waal 2018, 27).

[4] Alphabet-Google, Facebook, Apple, Amazon, and Microsoft
[5] such as trains or highways

Arguments made in the discourse around hate speech regulation, can only be analyzed under consideration of these elements. The motivation of actors is influenced by different societal, economic or cultural factors that determine the way they voice their opinion, which is why the discussion will include these aspects later on.

2.2 Platform governmentality

The structure and influence of platforms, which I have demonstrated in the previous section, leads to certain governance processes, which are an integral part of the ongoing discussion around hate speech regulation. This is why, I want to use this section to show which governance procedures[6] are employed in regard to platforms.

Generally speaking, we can distinguish two types of platform governmentality: governance *of* platforms, meaning rules imposed on platforms by institutions like the European Union, and governance *by* platforms, which includes the platform's own policies (Gillespie 2017, 2). In this sense content moderation tactics, as a practical extension of the company's policies, are of particular interest, as they ultimately execute the deletion of content.

2.2.1 Governance of platforms

Governing platforms on an institutional level entails many challenges often shaped by different discussions around legal platform definition, their responsibility, and institutional restrictions.

In terms of definition, governance of platforms is characterized through struggles with its "middle-ness": Neither are platforms simple ISPs (Internet Service Providers) that just host and spread content, nor are they news publishers in a traditional sense. They are what Gillespie calls "hybrids" (Gillespie 2018b, 210), whose categorization is particularly difficult as they disrupt the media system that has been based on a "century-old distinction" (Gillespie 2018b: 209). This does not only result in confusion in the public debates and but also under the law – and social media

[6] The removal of content through governance processes, be it because of the platform's policies or because of regulation imposed by institutions like the European Union, can be considered as "repressive" strategy to counter hate speech. The alternative to such strategies is what Kümpel and Rieger call "preventive strategies" (2019, 25). Measures in this area include community management by page admins, active involvement into discussions by journalists and use of counter-speech in online debates (Kümpel and Rieger 2019, 25–28). Such measures however place the burden of moderation on the users or require active admins in online-groups. They can be seen as a possible addition to countering hate speech, however, as the analysis will show, are not a substantial part of regulation intentions examined in this research.

companies are making the most out of this (Gillespie 2018b, 207). Van Dijck et al. acknowledge that platforms use these ill-defined, hybrid structures to deliberately bypass legislation that would otherwise limit their operations (Van Dijck, Poell, and de Waal 2018, 21). It is therefore important to clearly define where platforms stand.

For Tarleton Gillespie platforms can no longer be considered as "neutral" - even though they like to portray themselves that way. In the case of Facebook this change in neutrality happened when the company changed its feed from a chronological to an algorithm-based order. The intentional selection of content pushed the platform away from a neutral provider to a content curator (Gillespie 2018b, 211).

Looking at European policy approaches of hate-speech, those operational changes have been reflected in the policy strategy in recent years. Especially in matters of harmful content[7], institutions evidently express their growing concerns towards intermediaries and force them to act (Gillespie 2017, 4). The so-called European Union's Digital Single Market strategy showcased an increasing interest for harsher regulation and tightened the grip on the platform's responsibilities (Frosio 2017, 19). Though not intended, intermediaries have found themselves put into the role of an "internet police". As states do not have the ability or infrastructural means to control free speech, the responsibility falls to the intermediaries who have been subject to growing pressure by states to remove illegal content (Kuczerawy 2018, 32). So far, this pressure was applied through voluntary measures and the notice-and-take-down principle[8] that holds the companies accountable up to a certain point. For now, platforms have implemented self-regulatory measures to avoid regulation by states - concessions made, such as the "Code of Conduct Countering Illegal Hate Speech" (Kuczerawy 2018, 40), might be the result of states putting pressure on private companies threating consequences. (Kuczerawy 2018, 35).

The fact that platforms are not held liable for any uploaded content per se, originates from the current platform[9] policy, which has been introduced almost two decades ago and can be found in a European policy framework on Internet conduct known as the "E-Commerce Directive". These exemptions from so called "intermediary liability" can be traced to similar legislation in the United States (Savin 2018, 1218). Back in the day those exceptions from liability were created to push the emerging

[7] such as hate speech, extremism, misogyny, homophobia, threats of violence or other forms of harassment
[8] This means that unless a platform has been made aware of wrongdoings on their site they do not have to act on it by themselves (but can, if they want to).
[9] The wording platform has only been used since 2015 in EU-documents, before they were commonly referred to as "intermediaries" by EU regulators (Savin 2018, 1217)

Internet market and foster growth (Frosio 2017, 19). Because changes in liability exemptions can lead to more responsibility for platforms, this is also a concern which is frequently voiced in the discourse.

Even though intermediary liability exemptions and their reforms are a delicate topic, scholars like Tarleton Gillespie argue that operational changes of platforms need to be reflected by policy-makers. Gillespie questions the actuality of the current legal framework and highlights that it was created in a time when neither Google, nor Facebook had been created. With the internet growing and changing through time the challenges and responsibilities have altered, yet the legislation on intermediary liability stayed the same (Gillespie 2018b: 206).

2.2.2 Governance by platforms

So far, I have established the current state of platform regulation and its background. But platforms also use their own governance methods for various reasons, as this section will explain.

Gillespie emphasizes that the existing policy frameworks allow platforms to enjoy exemptions from liability but are still allowed to police as they see fit (Gillespie 2018b, 200). Indeed, platforms also regulate on their own behalf - and not just to prevent institutions from imposing regulations. Often corporate policies[10] enforce the company's own goals, i.e. to attract advertising clients (Gillespie 2017, 12). In relation to hate speech, Stark and Stegman critically note a duality in the behavior of platforms: On the one hand companies push emotional postings for engagement purposes, but at the same time they are complicit in removing (previously aggerated) illegal speech (Stark and Stegmann 2020, 41).

To enforce their guidelines, platforms perform "content moderation" (Roberts 2019), which comes with considerable issues. In contrast to the mainstream's believes, most of the moderation work is not conducted by machines or "AI algorithms" but by human "content moderators" who screen huge amounts of content every day (Roberts 2019, 25). The practice of human moderators is opaque and invisible (Roberts 2019, 1-3) yet, considering the platform's impact on public discourse, highly influential.

[10] Even though platforms operate under different goals, their guidelines are similar to EU policy demands. They censor or limit "sexual content or pornography, representations of violence or obscenity, harassment of other users, hate speech, representation or promotion of self-harm (…), drug use" etc. (Gillespie 2017, 14).

That platforms regulate is undeniably necessary[11] (Gillespie 2018b, 202) - but the way they do it, is subject of tension. Moderation happens under questionable circumstances and under an often undisclosed set of rules[12] that neither the public nor governments have access to (Wagner 2019, 113). It is precisely why academics have highlighted that the more Facebook is being turned into a "Internet police" (Kuczerawy 2018; Frosio 2017) by regulators, the more the platform's own dubious, corporate-made measures will dictate public discourse (Frosio 2017, 45). This concern can also be found in documents issued by civil rights groups and human rights bodies (Keller 2019, 5). The clarification of who (content moderators or "algorithms") regulates (voluntarily or not) what, using which standards (transparent or disclosed) is therefore vital for the interpretation and contextualization of the analyzed documents. The role of platforms in content moderation and the shifting responsibility of governance is a frequently discussed issue in the discourse. The arguments for and against it therefore will be represented in the category-set of the analysis (see chapter "Method").

2.3 Addressing human agency in Internet architecture

According to Roberts there is still a knowledge gap and confusion about the human labor behind content moderation (Roberts 2019, 25). False assumptions on how content is removed can result in misleading or inaccurate policy recommendations. Calling for "algorithms" to solve issues of hate-speech emphasizes a techno-deterministic approach and neglects the human role in such practices. Recent publications on algorithms/machine learning tools have questioned the ability of technology to work in that way. Laaksonen et al. address the fact that using technology to regulate hate speech can be difficult, developing functioning "algorithms" not only undermines the complexity of hateful speech but creates a dependency on technology (Laaksonen and et al. 2020, 11). In fact, Internet architecture comes with socio-technological tensions that become evident in discussions on (non-)automated decision making. Addressing human agency in Internet Architecture means questioning "function allocation": what are machines doing and where does (invisible) human labor take place (Wagner 2019, 106)? Only by addressing the human

[11] The sheer amount of content being posted creates an impossible challenge for platforms: they can only react retroactively to illicit content brought to their attention. Yet (as Western societies) we expect the platform to handle illegal content quickly and in a reliable, appropriate manner (Gillespie 2017, 16).
[12] Although some of Facebook moderation guidelines have been leaked in the past, which shows the sometimes strange logics in effect (Fisher 2018; Zinke 2018; Instagram 2019).

component in the technological structure it is possible to question how they work, who they work for and which rules apply. If the human component is neglected, those questions go unanswered and therefore also unregulated.

Facebook's strategy to appear as neutral actor (Gillespie 2018b, 199) did not only serve to avoid tighter legislation but also to hide the fact that there is no high-level technology solution (yet) to solve their issues (Murphy and Murgia 2019). With this in mind, it will be important to analyze how actors in the discourse highlight or neglect the human component in this regard. Scholars have previously noted that there is a discrepancy in addressing human agency as it is either neglected or particularly emphasized (Wagner 2019, 108).

In order to reflect on the theoretical findings, the analysis will assess the stakeholder's opinions in regards to the academic notions presented. In the following chapter I will present how my theory will be incorporated into the analysis of this research.

3. Method

The goal of this research is to understand the actors' key tensions on hate speech regulation in the discourse. To do so, I will use the method of critical discourse analysis (Gee 2014) on set of documents ("corpus"). After explaining how I will conduct my analysis, the last section of this chapter will establish a link between the method's tools and how they can be used to analyze the corpus.

According to Gee critical discourse analysis can be used to filter speech in terms of relevance and contextuality (Gee 2014, 3). In the examination I want to establish the identity of relevant stakeholders in the discourse around hate-speech regulation in a European context and develop an understanding for their opinions. This includes mapping their motives, how their arguments are shaped and how proposed solutions can be considered as techno-deterministic. By preliminary assessing the corpus, I have developed a category-based rubric which strikes a balance between theory and corpus. It will reflect on important theoretical findings, compare and categorize topics frequently voiced in the discourse[13].

I aim to filter fine-grained "nuances" within statements to establish an analytical insight in the positions, actors and actants take in the debate and how they constitute one another. Furthermore, I am interested in how solutions and critique on hate speech regulation are voiced and to what extent they are indicative of platform regulation.

3.1 Corpus

The tension on regulation towards hate-speech evidently focusses on the regulator, in this case the European Union, and the companies that are subject to regulation, here Facebook. In terms of corpus this led to the selection of a limited number of relevant documents, which I will explain in the following section.

The new European Commission named the renewal of the E-Commerce Directive and the establishment of the "Digital Services Act" as one of their key-goals (European Commission 2020, 12). The corpus on the "EU opinion" therefore heavily relies on documents issued by the European Commission on the topic of *hate speech (moderation) / tackling illegal content.* Considering that the last sub-question of this thesis aims to examine future implications on European policy making, I have paid special attention in restricting my research on documents issued in 2019 or 2020.

[13] for further explanation on the category-system, see section "application of method".

Additional to official documents, I have chosen one “leaked” Commission-document which was published on the German digital rights platform “Netzpolitik.org” (Rudl and Fanta 2019), indicating possible future Commission policy-recommendations on hate speech regulation. Since the Commission’s official statements on future regulation were quite vague in terms of substance, the most current framework handling hate speech moderation was added to the corpus of analysis. The “Code of Conduct Tackling Illegal Content Online” was issued in 2016 and defines EU-guidelines for “IT-Companies” to tackle “illegal hate speech online” (EU Commission 2016).

The second part of the discourse concerns Facebook, which is one of the biggest platforms subject to the new European regulation attempts. Regarding hate speech regulation, Mark Zuckerberg, CEO of Facebook, can be identified as one of the leading voices in this discourse. His statements were gathered through the search portal “Zuckerberg Files” and consequently filtered in terms of *hate speech / tackling illegal content / regulation of Tech companies*. Additional to Zuckerberg’s statements, the press release on the “Digital Services Act” issued by the tech-lobby-group “EDiMA”, finalizes this set of the corpus.

In addition to these two actors, I have chosen documents of “third-party stakeholders” to complement the analysis. This will allow me to represent voices in the discourse that are independent from both involved parties. For this corpus, I included documents by the civil rights association *EDRi*, specifically addressing *hate-speech (platform) regulation* and the *Digital Services Act.* Due to its location in Brussels, EDRi also provides unique insights into EU policy-making and, as a civil rights group, adds a bottom-up approach to the analysis emphasizing the enforcement of civil rights. The corpus will be completed with opinions issued by the United Nations High Commissioner for Human Rights (“OHCHR”) on the topic of *(hate speech) content moderation*. Documents from a representative of an international organization are particularly interesting as they clearly do not serve either European or corporate interests, but due to the institutional structure (top-down approach), also defer from NGOs like EDRi.

3.2 Application of method

In this section I will explain the specific use of the previously introduced method and how it can be related to the corpus, theoretical framework and the research questions.

The first sub-question[14] concerns the stakeholders and their "stakes" in the examined discourse, which is why it is necessary to identify the stakeholders and explain their interests as actors in this discourse. This will be a merely descriptive part, however vital in order to understand the discussion of the analyzed documents. Then, the content needs to be examined in terms of the most important subjects in the discourse, which is why I have developed a category scheme based on theoretical findings and the analyzed discourse. Here the theory and the corpus "meet" for the first time. The subject *pro-regulation and regulation-critical* represents opinions voiced for and against the establishment of institutionalized policy for platforms. Here, the section on platform ecosystem and influence is important in order to understand the motives behind calls for more or less regulation. The theoretical section on governance (by and of platforms) is indicative for the categories for *more or less corporate governance*. I have created this category because the call for more responsibility of platforms *or* less power for platforms is a frequently voiced topic in the discourse that poses a clear line of conflict between the different actors. The last set of categories addresses the subject *of human agency and techno-determinism*. As Wagner argues the role of humans in technology is either specifically addressed or completely left out (Wagner 2019, 108). The category of human agency measures which actors mention humans in the work on platforms (creating algorithms, doing content moderation). Whereas the other category, techno-determinism, reflects on the neglect of human agency, as well as the reliance on machine-based (non-human) systems and proposed solutions based on AI technology.

Mapping the topics of the discourse will also help in answering the second sub-question. Here the *Subject Tool* comes into play: Gee explains that this tool can be used to identify why stakeholders relate to specific subjects and how information is organized around those subjects (Gee 2014, 199). In this case this tool can help show how different stakeholders address a subject (such as human agency). It is not unlikely, as the analysis reveals, that two supposedly opposing parties both voice their opinion

[14] RQ: What are the tensions in the discourse on EU-hate speech regulation and how do they reflect on future platform governance of the Digital Services Act?
SQ1 What are the goals of the stakeholders?
SQ2 To what extent are the topics governance and human agency reflected in the discourse around hate speech regulation?
SQ3 What are the key tensions between stakeholders in the discourse?
SQ4 In what ways are the proposed solutions indicative of a techno-deterministic view on hate speech moderation?
SQ5 What are future possible implications in development of EU platform policy, respectively the Digital Services Act?

in the same subject category but ultimately do not share the same goal. It is therefore necessary to understand motivation behind such expressions and analyze them accordingly. This is where the *Doing and Not Just Saying Tool* (Gee 2014, 200) will be helpful. It can be considered as an extension of the subject tool and tries to answer the underlying reasons of the actors' behaviors and expressions. If the CEO of a big American platform calls for more regulation in tech (Zuckerberg 2020a) – can this be interpreted in the same way as when the Commission is calling for more regulation of IT companies (European Commission 2016)? Finally, I will elaborate on the indications of the overall discourse and *key tensions between stakeholders (SQ3)*. This step should allow me to position different opinions within the discourse. My fourth sub-question on *indications for techno-determinism in hate speech moderation* will end this chapter by critically reflecting on solutions.

The goal of the analysis is to establish an understanding of hate-speech regulation that not only deals with pro- and antiregulatory opinions but contextualizes them within the discourse. I want to do this in order to question the dichotomy of "good" and "bad" (pro and anti) and ask how arguments can be mapped within the bigger picture. This also could help addressing the complexity of the topic and reveal a more fine-grained understanding of future governance practices in this field (SQ5).

The outcomes of the analysis are to be treated with usual reservations since the method of discourse analysis comes with several empirical implications. Due to its limited scope, the analysis cannot produce representative, universal outcomes or generalized statements but only small-scale interpretations (Gill 2000, 186). The categories I have chosen in the analysis have been deducted from the questions and theoretical literature research. This means that even though it is my aim to answer the questions under consideration of all aspects, there is still the possibility that not all of them will be covered. However, apart from such considerations, discourse analysis offers a number of advantages compared to other methods: When it comes to discourse, large-scale computerized examination might as well fail in answering nuanced questions or detecting certain discursive codes.

4. Discussing hate speech regulation in light of the DSA

In the following chapter and sections I will introduce my findings of the discourse analysis. The discussion will start with a map of the stakeholders and their goals (SQ1) and will be followed by an examination of the mentioned subjects, which have been categorized and reflected on with the *Subject* and *Doing and not just saying* tool (SQ2). This part will be wrapped up by a reflection on the *key tensions* of the discourse (SQ3). Finally, the results will be examined from a critical view on techno-determinist approaches (SQ4).

4.1 The stakeholders: their background and goals

This chapter aims to identify the stakeholders in the discourse and map their goals relating to hate speech content moderation and platform policy.

4.1.1 The regulator: The European Union

When the new European Commission under Ursula von der Leyen started their work in 2019, the so-called “Digital Services Act” (DSA) was briefly introduced as one of the key goals (Bassot 2020). Since then the Commission’s communication has been closely monitored by different stakeholders.

So far, the Commission has been enforcing the removal of illegal hate speech content through a Code of Conduct[15], which was created in 2016. Yet, according to the new Commission, the efforts made by the companies were not good enough to win the fight against harmful content on social media (European Commission 2020b). In February 2020 the Commission announced its roadmap for their work on ”Europe’s digital future” which presented time-frames for future regulatory documents. The Digital Services Act is to be expected at the end of 2020 (European Commission 2020a). The reason why contents of this new framework are already under discussion, is that the Commission and other stakeholders have been working on the DSA in the background for months, if not years. A leaked internal document on the possible contents of the DSA, which was published in June 2019, confirms such initiatives (EC DG Connect 2019). The leak hints at possible plans of the Commission for platforms and their regulation and suggest a possible regulator for content moderation, a reform of the

[15] The non-legally binding framework was signed by many of the big platforms in the industry, including Facebook, Google, YouTube or Snapchat (EU Commission 2016).

current liability regime and legal incentives for platforms to take illegal content "proactively" down (Rudl and Fanta 2019).

The current intermediary liability regime is one of the most crucial and sensitive topics in the Commission's reform plans. The current regulatory framework[16], the e-Commerce Directive, allows platforms exemptions from liability. This means that Facebook as well as all other smaller or bigger intermediaries are not subject to liability for content as long as they are not notified about any wrongdoings ("notice-and-action"). Reforming the liability regime is a delicate task because more responsibility for platforms might have severe consequences on free speech, as civil rights organizations like EDRi or the OHCHR warn. More responsibility for platforms to take content down might only solve the problem on the surface.

The Advocate General's opinion on the case of "Eva Glawischnig-Piesczek v Facebook Ireland Limited" was considered as another possible milestone for this policy-making progress. The Court's decision[17] was path-leading as it interpreted the existing European legal framework (the e-Commerce Directive) in specific regards to the liability regime in case of hateful speech against a user. In doing so, the Court not only opened the possibility for national European Courts to order Facebook to take down content globally but also specifically mentioned that hosting providers, like Facebook, "may have recourse to automated search tools and technologies" (Court of Justice of the European Union 2019).

4.1.2 The regulated: Facebook

On the side of the regulated parties, there are the platforms, in this specific case Facebook. According to academic scholar Tarleton Gillespie, Facebook has enjoyed certain amounts of freedom in their governance practices (Gillespie 2018b, 210). The liability exemption under the e-Commerce Directive allows the company to use corporate governance (like content moderation) against users who violate their (vaguely) formulated terms of conditions. Those methods are opaque and especially for users not easy to understand. Giving Facebook more responsibility to remove illegal content through changing the liability regime means that there is a risk of endorsing Facebook's methods.

[16] See chapter governance of platforms

[17] The Court ruled that while "general monitoring" and "actively [seeking] facts or circumstances indicating illegal activity" is prohibited by the Directive, national Courts can order Facebook to remove "identical" content which was previously found unlawful as well as content that is "equivalent" under specific circumstances.

There is no understanding of Facebook's governance strategy without looking at the fact that Facebook is a corporate company and therefore by definition has a different goal than the regulator, the European Union. In the case of moderation, this role is always in accordance with profit-making: As Roberts states[18], Facebook's content moderation is fundamentally influenced by advertising-interest (Roberts 2018). So far, Facebook has been very proactive in deleting content[19] that is against ad-sellers interest, yet in other matters, such as hate-speech, it was the European Union who had to apply pressure (such as the Code of Conduct). This profit-orientation is also one of the reasons why Facebook is not keen on making its moderation tactics public. Not everything that Facebook deletes, is deleted for common good but for corporate interests. This fact alone makes it questionable whether it is a good idea to put Facebook in charge of more moderation in the future.

4.1.3 Third party actors: NGO and OHCHR

Even though EDRi and the OHCHR are two very different entities by core, their role as third-party-stakeholders unites them in more than just one regard. Both of them are engaged in the discourse because they see possible harmful consequences on both sides of the stakeholders. They share a very strong emphasis on the protection of fundamental rights. The topic of hate speech regulation is therefore analyzed through the lens of rights-protection. Even though the European Commission claims to hold the same rights in high regard (European Commission 2020b), the solutions between the Commission and the third-party-stakeholders are often far apart from each other.

EDRi and the OHCHR view corporate governance-based solutions skeptically: If the EU asks for more responsibility from Facebook, their opaque governance structures will be further enforced. As a result, public discussions are policed by private actors which might have significant consequences on the freedom of expression. For EDRi one of the main goals is the enforcement of civil rights in the digital sphere. Therefore, civic engagement in content moderation tactics, transparent communication of such practices and civil say in content flagging finds a strong emphasis in all

[18] "Even when platforms do acknowledge their moderation practices and the human workforce that undertakes them, they still are loath to give details about who does the work, where in the world they do it, under what conditions and whom the moderation activity is intended to benefit. To this last point, I argue that content moderation activities are fundamentally and primarily undertaken to protect and enhance platforms' advertising revenue, protect platforms themselves from liability, provide a palatable user experience (highly context-dependent) and, when necessary and in response to specific jurisdictional or regulatory rulings, to comply with legal mandates. In short, the role of content moderation is fundamentally a matter of brand protection for the firm" (Roberts 2018)

[19] For examples of such cases please see Byström, Soda, and Kraus 2016; Faust 2017; Olszanowski 2014.

documents analyzed. For the OHCHR it is the application of human rights which is held is highest regard, arguments made are generally put in context of rights protection and fulfillment of the Human Rights Conventions.

4.2 Identifying subjects in the discourse

The following sections will introduce different subject categories in the discourse on hate speech regulation. As previously explained (see method section), the subject categories are deducted from theory and extended through a corpus analysis. The creation of such categories facilitates comparing statements from different texts and actors.

4.2.1 Regulation as field of tension?

This section will analyze how the stakeholders voice their opinion on the topic of platform regulation, especially in regard to hate speech. In order to understand possible tensions regarding this subject, two categories “pro-regulation“ and “regulation-sceptic” have been established for the analysis. This subject addresses how the stakeholders address the general notion of applying regulation[20] in the area of hate speech moderation, not specific areas regulation touches upon. These, such as governance-authority and content moderation, will be reflected on in more detail later.

It comes as no surprise that documents issued by the European Commission showed a high occurrence in the category “pro regulation”. The Commission seems to agree with the theoretical arguments on the need for an update of the current frameworks and promises to improve them by “increasing and harmonizing the responsibilities of online platforms and information service providers and reinforce the oversight over platforms’ content policies” (European Commission 2020a). The documents addressing regulation read as a letter of complaint, stating a number of issues, such as lack of authority, control or effectiveness (EC DG Connect 2019). It seems out of question that the Commission wants to change the framework under which platforms currently operate, however when it comes to concrete regulatory proposals, the statements made explicitly towards regulation, remain vague. As a regulator, the Commission might not be very vocal during this time, considering that the Digital Services Act is still in progress. However, the Commission shows a possible

[20] In the theory see “governance *of* platforms”.

tendency by considering that “online intermediaries can put in place proactive measures without losing the liability exemption the e-Commerce Directive” (European Commission 2020b).

In contrast to the regulator side, the other stakeholders in the discourse seem to be more articulate on their views on new regulation. As the analysis shows, Facebook does not seem to shy away from regulatory options. In fact, Mark Zuckerberg has been lobbying publicly for better regulation since the beginning of 2020, even writing an article in the Financial Times stating that “Big Tech needs more regulation” (Zuckerberg 2020a). The platform publicly participates in the discourse on regulation rather than dismissing it overall. Facebook’s *public* concessions to regulators are extensive, reaching from willingness for external audits in content moderation to calls for transparency, even hoping for more “guidance and regulation from states” (Zuckerberg 2020b). Such statements are not just surprising coming from a company, that has been reluctant to show transparency in content moderation aspects (Roberts 2019) but also because they completely disagree with policy recommendations issued by its own tech-lobbyists at EDiMA.

EDiMA, contrary to Zuckerberg’s statements, has been very clear about wanting much flexibility for service providers to conduct business. Platforms should decide what policies suit them best and, according to the lobbyists, “the scope of the new framework should be broadly defined, technology-neutral, and principles-based, applying proportionately to a variety of different online services rather than a specific list” (EDiMA 2020, 3)[21]. Such statements correspond with current academic views on the governance of platforms, where it is said that platforms use confusion around their definition to circumvent policies that could restrict them (Van Dijck, Poell, and de Waal 2018, 21).

Neither Zuckerberg’s editorial in the Financial Times or his coordinated (and almost word-alike) statements calling for more regulation (Zuckerberg 2020b; 2020a; 2020c), nor EDiMA’s detailed policy recommendations, can be considered as “accidents” in the discourse. This leads to the conclusion that apparently Facebook’s CEO seems to *say* one thing, his lobbyists *do* something else. This makes understanding the discourse on hate speech regulation much more difficult.

[21] Furthermore EDiMA does not wish for any oversight that has power over provider’s decisions, which again, can only be interpreted as the opposite of Zuckerberg’s call for “more oversight and accountability” where “global technology platforms answer to someone, so regulation should hold companies accountable when they make mistakes” (Zuckerberg 2020a, 3).

Stakeholders that seemingly belong to the same group, share different views on the same topic.

The OHCHR does not necessarily disagree with regulation as such, however remains sceptic on certain ideas. States should not push companies further into responsibility, leading to the use of automated tools threatening freedom of expression and ending in "prepublication censorship" (United Nations General Assembly 2019). EDRi voices similar concerns, stating that while the Commission does address that platforms should not be pushed into deleting content, frameworks like the Code of Conduct shift the burden to the platforms and ultimately pressure them to act through deleting content (EDRi 2019). Such concerns are consistent with academic standpoints on regulatory governance approaches (see Kuczerawy 2018; Frosio 2017). From a civil-rights perspective, the Commission fails to serve its citizens by not giving them more agency[22]. Instead of more corporate governance through regulation, states should aim not move closer towards what the OHCHR calls "smart regulation" which is based on strengthening the user's agency and enhancing transparency in content moderation (United Nations General Assembly 2018b, 19). Ultimately, the third parties' reservations against regulation stand and fall with the amount of responsibility regulators like the European Union hand over to the companies. It is here where publicly oriented institutions could lose their grip on public discourse while trying to protect its citizens from infringement against such.

To summarize, none of the actors are completely against regulation, however do share skepticism (for corporate or human rights reasons). These findings make obvious, that the discussion is not necessarily about implementing regulation but the content and enforcement of such. Investigating the nuances further is therefore necessary for an understanding of the goals and arguments in the discourse.

4.2.2 Shifting responsibilities in areas of governance

"Holding companies responsible" is a buzzword that the Commission uses on many occasions but this responsibility can mean different things. In the following section I will reflect on my findings in the categories on "more corporate governance" and "less/anti corporate governance".

[22] "It [the Code of Conduct] does not (…) improve legal certainty for users, nor does it provide for proper review and counter-notice mechanisms, or allow for investigations into whether or not the removed material was even illegal" (EDRi 2019).

In order to tackle hate-speech on platforms there are different approaches mentioned by the various stakeholders. The main tension between those approaches focusses on the question of governance and responsibility of the platforms for illegal content. The category called “more corporate governance” indicates statements that call for more governance-responsibility of companies in hate speech moderation and subsequently support company policies to remove content. The category “less/anti governance” highlights statements that oppose such ideas and articulate possible risks of handing over decisions of speech into the hands of private actors.

Gillespie’s notions on platforms aiming to strengthen their own power on enforcing their policies seem to be confirmed by the statements found through the analysis (Gillespie 2017, 12-14). The category of “pro corporate governance” shows that Facebook is pushing for corporate governance-authority on many occasions and seems to be keen to enhance its position to execute its own policies. The tech-lobby EDiMA makes it very clear, that if platforms are asked for more responsibility in taking down content, such measures must come with limited liability to not “perversely” police tech-companies for taking content down. In doing so, the lobby defines its own meaning for responsibility and argues: “Service providers would define the kind of measures which best suit their unique situation, and which are the least intrusive for users” (EDiMA 2020, 2). This practice is particularly delicate because platforms demand full autonomy in this regard, meaning that regulatory bodies or users have no say in the creation or enforcement of moderation rules. Revisiting Gillespie’s governance theory, such patterns in the discourse of platforms are not new and can be explained by the company’s interest to moderate as they see fit (Gillespie 2018b, 200) with the purpose of pleasing advertising clients (Roberts 2018). Giving Facebook more responsibility without monitoring it (in case they do not act according to a certain set of rules), would entail that Facebook can continue operating under non-disclosed rules and remove contain without having to justify why and under what circumstances.

Therefore, the Commission’s calls for “more responsibility” for platforms can only change the current circumstances if they are bound to transparent rules. Yet, so far the Commission’s statements remained very vague, such as in terms of concrete content moderation principles or ideas for regulatory provisions, such as a Public Ombudsman (as suggested by Gillespie 2018b, 214). They simply mention that the ultimate goal is to ensure responsibility of platforms and create “a common approach

to quickly and proactively detect, remove and prevent the reappearance of content" (European Commission 2020b).

This is why third-party-stakeholders in this discourse challenge regulatory approaches that might result in a shift of governance towards private actors. Giving corporate companies more responsibilities to take down content, comes with a lot of concerns, as the analysis revealed. According to the OHCHR, this leads to "pressure on companies such that they may remove lawful content in a broad effort to avoid liability" (United Nations General Assembly 2018b). This is particularly worrisome as it is unclear what companies consider as "hateful"[23], which results in very little accountability for those possibly undermining freedom of expression (United Nations General Assembly 2018b). For EDRi the calls for more responsible behavior of platforms are a seemingly easy way out of a problem that needs to be treated from its societal core (Berthélémy 2019). The Commission's approach to push companies into further removals through their terms of service is "a convenient way of removing legal content as they are vague and redress mechanisms are often ineffective" (EDRi 2019).

Both the Commission and Facebook seem to be aware of these potential risks, with the Commission considering the idea of public regulatory oversight and aiming for protection of fundamental rights (EC DG Connect 2019) and Zuckerberg publicly admitting that people "don't want private companies making so many decisions about how to balance social equities without a more democratic process" (Zuckerberg 2020b, 21:27).

Further pushing the governance onto private actors has practical implications on the actual process of content moderation. The following section will therefore highlight concrete measures to do so and examine to what extent the stakeholders refer to the specificities in content moderation during the discourse.

4.2.3 Human agency in hate speech moderation: A crucial point

Why is the acknowledgement of humans so integral for the discussion around hate speech moderation and regulation? This is what I would like to explain in the course of this discussion point.

[23] For the sake of thoroughness, I want to mention that there is also a discussion on the policing of hate speech content, as the term "hateful" (often used in policy-papers and guidelines) can be interpreted in many ways and sometimes includes content that is not illegal per se (which makes its removal from public speech even more delicate). However, due to the limited scope of this paper (which cannot include a legal-analysis), the discussions around the removal of borderline illegal hate-speech content, will not be handled here.

The subject category of “human agency” examines if and to what extent the analyzed documents mention the role of humans in the process of content moderation. The execution of content moderation, be it human or automated, starts with the guidelines they are based on. Moderation can only happen on the basis of such guidelines, which is why they are crucial for this process. The interplay of humans and technology can happen at different stages. The clearer human work is addressed, the more precise its assessment can be (Wagner 2019, 106). This means in reverse: when actors fail to disclose human practices, they fail to acknowledge and answer to any consequences that come with it.

Even though the phrasing of guidelines is therefore a particularly delicate one, and possible point of intervention for regulators, only the OHCHR mentions it in detail (United Nations General Assembly 2018b). The documents issued by the OHCHR refer to all possible steps of moderation processes in detail and in specific reference to human agency, which results in a very holistic view on the topic of moderation.

The next step of moderation, the actual assessment of content, is the main part of this process. As explained in the theory section, those conditions have been also academically examined in the past (Roberts 2019). The analysis reveals that again all reports published by the OHCHR specifically mention the human component in this step. This not only includes a detailed description of content moderation labor (flagging, selecting, reviewing), but also mentions the increasing issues around this work due to the growing scope and pressure by governments (United Nations General Assembly 2018b). On the regulator side (EU Commission), the existence of flaggers and human moderators is acknowledged but not necessarily mentioned as integral part of the process and sense-making. However, even though the EU Commission’s documents lack specificity on the work conducted by human moderators, their labor conditions are specifically mentioned[24].

The topic of human agency finds repeated mentions in its relation to the use of AI for the removal of content. Here, the European regulators seem particularly keen to maintain “humans-in-the-loop” in automated content moderation processes (European Commission 2019, 7). The specific mention might stem from a general distrust in the use of technology for the purpose of removing content. The OHCHR reports address issues on the use of AI for content moderation works extensively. This shows that the

[24] In the Commission’s statement on “Europe’s digital future” the roadmap mentions a plan for the improvement of labor conditions by 2021 (European Commission 2020a).

reports place a high emphasis on the human influence in technology. The reason for the heavy emphasis is that, as the reports argue, automated tools are not able to assess questionable content properly. For one, there are potential issues of discrimination with a potential risk of "overmoderation" (United Nations General Assembly 2018a, 8–9). Furthermore, the reports refer to studies that proof a lack of AI-understanding when it comes to the examination of context or specificity such as irony. From a human rights perspective, which the OHCHR represents, calls for automated decision-making in delicate areas such as freedom of expression, are therefore to be treated with extreme caution. In this regard, the OHCHR specifically targets states and supranational institutions for their pressure-making in the implementation of automated tools for the purposes of hate speech moderation.

None of the above-mentioned issues in content moderation and fears around the use of AI can be found in Facebook's statements on hate speech and its moderation. In fact, the human component in moderation labor was not mentioned in a single text analyzed for this research. This corresponds with Gillespie's notion of Facebook depicting itself as neutral actor (Gillespie 2018b, 199) and Roberts assessment of the platforms' inability to address human moderation labor (Roberts 2019, 37).

4.3 Central controversies and techno-determinism in the discourse

Looking at the *key tensions between the stakeholders* (SQ3), this research revealed that the question of *regulation* is not the focal point of the discussion. Rather, main arguments center around *who should be given the burden of responsibility to interfere with speech online*. It is exactly where the controversies start to become nuanced and the understanding of arguments becomes difficult. Mark Zuckerberg calling for regulation even though it "may hurt Facebook's business" (Zuckerberg 2020a, 3), might appeal to a broad audience. But regulation itself is not the controversial topic (anymore?) – the controversy on how the regulation should look like is. It is necessary to keep in mind that even if Facebook's Zuckerberg and the European Commission are both calling for regulation, it does not mean that they have the same intentions in mind. Which is why there is a need to fragment the concept of platform regulation and question what this broad term entails. And indeed, as the analysis shows, being "pro regulation" and "pro regulation" mean different things, as soon as the arguments, the subjects, are called into question. Ultimately the heart of the controversy is around the

governance-authority, the “responsibility”. This includes the definition of guidelines, the enforcement of such guidelines and respectively the removal of content (with all the consequences this entails). And while Facebook states to consider “oversight” as a possible solution for the future (Zuckerberg 2020a, 3), unwillingness to open up about their own governance processes and inability to address human labor in content moderation says something different.

The following Figure 2 bears witness to the key tensions established through this analysis and can be seen as an extension to the first visual (presented in the Introduction-chapter).

The shaded arrows depict possible points of regulatory intervention, touching on corporate policies and content moderation itself. This depiction shows a more nuanced and detailed visualization of how regulation could address different areas of governance.

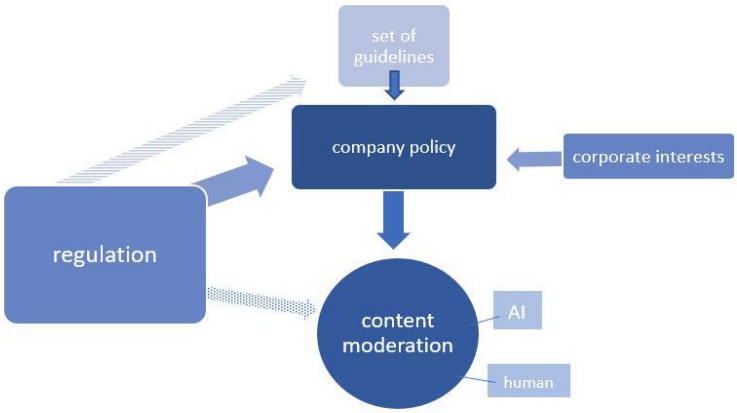

Figure 2: Governance visual extended. Source: The author

4.4 Techno-determinism in regulation and hate speech moderation

In the last section of this analysis, I want to address possible regulatory solutions in the process of governance from a techno-critical perspective (SQ4). Statements which have been attributed to this category (“techno-determinism”) show willingness or emphasis for technological solutions in hate speech regulation[25]. The corpus material

[25] It is important to note that despite some of the solutions discussed by actors in the discourse, artificial intelligence is – to this date – not able to understand context in the analysis of hate-speech content (Laaksonen and et al. 2020; Murphy and Murgia 2019). So far, programmers have been successful in creating artificial intelligence (AI) solutions that are able to detect hate-speech content, however struggle when it comes to putting them into context, as Jerome Pesenti, Facebook’s Head of Artificial Intelligence, recently noted: “Deep learning and current AI, if you are really honest, has a lot of limitations. We are very very far from human intelligence, and there are some criticisms that are valid: It can propagate human biases, it’s not easy to explain, it doesn't have common sense, it’s more on the level of pattern matching than robust semantic understanding” (Knight 2019). Even though big tech companies have invested huge amounts of money and work-load into the development of hate-speech detection-technology, it is evident that “algorithms” still cannot live up to the expectations put into them. Yet, they are still met with a high level of trust and hope in order to solve hate-speech issues (Laaksonen and et al. 2020, 4)

revealed a considerable use of metaphors to describe the use of technology, which is why I will pay specific attention to them.

As introduced in the chapter on human agency, Facebook does not disclose human labor in content moderation. However, this does not mean that Facebook does not address its content moderation work publicly – they simply fail to mention one of its most integral parts: humans. Instead, Facebook, in many cases Zuckerberg himself, defers to the (successful) use of artificial intelligence in content moderation. In doing so, the emphasis is on the improvement of AI in detecting content, rather than its failure. Often Zuckerberg uses the success of AI technology in one area of content moderation (i.e. terrorist content) to relate it to the assessment of hate speech moderation. Instead of admitting that AI is not able to precisely detect hate speech, Zuckerberg uses statements like “this isn't a problem that you ever fully solve” or “but I do feel like we are improving” (Zuckerberg 2020b, 08:52-10:09) to stir away from the problem. In other statements he even diminishes the sophistication of hate speech, stating that “people who (…) say hateful things aren't necessarily getting smarter at saying hateful things” (Zuckerberg 2020b, 09:24). Such statements lay the discursive groundwork for calls on the use of automated tools to conduct content moderation.

In the discourse there is a particularly high use of metaphors when stakeholders address the use of technology in content moderation. EDiMA, for instance, wishes to “protect consumers” by implementing “additional measures” (EDiMA 2020). “Additional measures” hint to use of AI-technology to facilitate removal of content. Similar wording can be found in the Commission’s documents, often using phrases like “proactive tools”, or “proactive measures” as an indication for the use of algorithms/filters/artificial technology. The European Commission’s opinions are particularly sensitive as future policy-making could have a heavy impact on the development of platforms and their impact on freedom of expression. While the wording is a little more cautious, the call for “proactive” measures is consistent in many of the documents[26] (EC DG Connect 2019, 2).

Circumventing the specific wording might distract some readers from the fact that technological solutions are discussed. But this is exactly where discourse meets regulation and where statements made are crucial. Not “punishing” platforms for “additional measures” (as demanded by EDiMA), such as AI or automated filters, could

[26] In the Commission’s leaked document from July 2019, the authors go as far as calling for more clarity in order not to “disincentivize” platforms to take proactive measures. Interestingly, a very similar phrasing can be found in the EDiMA-document addressed to the EU policy-makers just a few months later (EDiMA 2020).

potentially allow tech-companies to employ technology that is unable to understand hate-speech. Through the implementation of their own rules, they could determine not only which expressions are to be taken down but also use methods to do so that might lead to automated filtering of speech. This would further strengthen their influence in an already heavily platform-dominated system, in other words, our platform society. Considering the huge impact of platforms on public discourse (Gillespie 2017, 25; Van Dijck, Poell, and de Waal 2018), this could lead to a further shift from public into private discursive oversight. This is why, there are very different, opposing opinions among the stakeholders in this category. The OHCHR, with its consistent mentions of human agency in content moderation, views the use of automated filters very skeptically, whereas Facebook's statements are dominated by positive examples of artificial intelligence.

So, are the proposed solutions indicative of a techno-deterministic view on hate speech moderation? The Commission and especially Facebook show techno-determinist tendencies in their solution-proposals. While the Commission focusses on the possibility of "proactive measures" as an aid for content moderation, Facebook seems to be very articulate in painting a techno-utopian scenario where – eventually - technology will reach the superiority to solve such issues. While this might be a favorable scenario for Facebook as a tech-company, it is also one that can most likely never happen (keeping the previously articulated technological limitations in mind).

Overall, this section highlighted the need for a thorough understanding of platform governance methods that not only address issues in this context but also the practice of solving them. This is a notion that could be observed in many parts of this discussion and showcases some of the nuances in the discourse on platform regulation. My final chapter will therefore elaborate on the most interesting findings, reflect on the method and consider possibilities for future policy and research.

5. Conclusion: Implications for the Digital Services Act

The overall outcome of the analysis shows that the tensions on EU hate speech regulation do not necessarily center around the necessity of regulation but specific governance methods. The notion of responsibility has been established as a key-phrase of this discussion: Who is to be made responsible for illicit content and what does this responsibility entail for governance procedures? As regulator, the European Union will need to answer these questions carefully and with the necessary foresight. Calling on platforms to take more responsibility will not be enough to counter the issue of hate speech on platforms. The European Union might end up giving Facebook and other platforms even more power over speech, which is particularly worrisome, considering that institutions seem to be already struggling with the growing influence of platforms (Van Dijck, Poell, and de Waal 2018, 27).

Goals of the actors are driven by certain values or economic interests. This corresponds with the theoretical framework of the research. The discourse analysis revealed that questions over governance, which are at the heart of this discussion, are extensively addressed by actors, specifically in the relation to content moderation. Hate speech is often specifically mentioned as an example. However, when it comes to human agency, the analysis shows big differences among actors. While third-party-actors emphasize human agency, the European Commission seems to lack concrete references in many aspects, often just mentioning humans in their relation to artificial intelligence. One of the most crucial results is that Facebook's statements entirely lack human components in hate speech moderation (or, as a matter of fact, in any part of governance procedures). This aspect is particularly interesting from a techno-critical view, as it hints to a techno-determinist argumentation. The last chapter confirms such tendencies, especially in Facebook's discourse. Here, the platform seems to have developed a particular rhetoric to circumvent the notion of human agency, while heavily emphasizing technological solutions (that are, from today's standpoint, especially in hate speech moderation, still unrealistic). Only when putting the arguments from Facebook (be it EDiMA or Zuckerberg) into context, the true intentions of the platform become obvious: Facebook seems to have two faces while ultimately arguing for one goal: maintaining governance authority with little regulative intervention, especially when it comes to (human) content moderation practices. Here, it is particularly interesting to see which statements are voiced and how they can be interpreted, once

they are set into context or related to other statements made in the discourse. The method of critical discourse analysis proved to be particularly useful as it helped uncover how stakeholders refer to common subjects but also how statements might indicate one thing and mean something different.

For the development of the Digital Services Act this means that a detailed assessment of corporate governance and an exact understanding of content moderation practices will be crucial for regulatory consequences on platforms. In order to make this more tangible, let me use my findings to paint the following scenario: If the European Union simply remains on the standpoint of ordering platforms to take down certain content, Facebook and other platforms will be strengthened in their policing activities. A change in intermediary liability exemptions could furthermore lead to platforms having a legal obligation to remove content. This, paired with the European Commission's multiple mentions and the European Court of Justice's recent ruling on the use of "pro-active measures", could finally lead to the establishment of online filters. Why? When platforms are responsible to take down content (and will be punished if they do not) and are allowed to employ "proactive measures", then filters might be the consequence. Vice versa, the clear establishment of common and transparent rules for platforms, the reinforcement of user agency in content moderation, a public oversight body and constant measures for accountability in the policing of speech, could take power over discourse away from platforms and strengthen the individual user.

Looking back at the earlier presented "Eva Glawischnig-Piesczek v Facebook Ireland Ltd." case, Keller was definitely not wrong to state that the case might have big implications for the digital infrastructure of the future (Keller 2019, 2). By tackling hate speech through regulation, the European Commission is targeting a big issue of public discourse online, which has the power to ultimately frame the future of freedom of speech.

Once the Digital Services Act takes form, this research could lay the groundwork for future policy analysis of the Commission's decisions in this regard. Still, I could only touch on one aspect, the discourse, in this policy-work-in-progress and there are many more to uncover. As mentioned before, the legal basis of the framework, such as the definition of hate-speech and its issues around (il-)legality are just one of the potential weaknesses that need to be revisited from an academic standpoint. Furthermore,

insights into the policy-making process, including the impact of lobby-groups during such, could contribute to an academic evaluation of governance processes.

Finally, as a critical new media scholar, I hope to have contributed to a better understanding of regulation in our field. While social media and its impact on society are frequent topics of the academic debate, regulation is often left aside. Considering the above-mentioned impact of regulation on the digital sphere, I believe that European digital regulation should earn a more prominent position in our field. And I hope that this research was able to showcase why there is a need to do so.

Bibliography

Bassot, Etienne. 2020. 'The von Der Leyen Commission's Priorities for 2019-2024 - Think Tank'. European Parliament Think Tank. 28 January 2020. https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI(2020)646148.

Berthélémy, Chloé. 2019. 'Interoperability: A Way to Escape Toxic Online Environments'. EDRi.org. 4 December 2019. https://edri.org/interoperability-way-to-escape-toxic-online-environments/.

Byström, Arvida, Molly Soda, and Chris Kraus. 2016. *Pics or It Didn't Happen. Images Banned from Instagram*. Munich: Prestel.

Court of Justice of the European Union. 2019. 'EU Law Does Not Preclude a Host Provider Such as Facebook from Being Ordered to Remove Identical and, in Certain Circumstances, Equivalent Comments Previously Declared to Be Illegal'. *Judgment in Case C-18/18 Eva Glawischnig-Piesczek v Facebook Ireland Limited*, no. PRESS RELEASE No 128/19 (October): 2.

EC DG Connect. 2019. 'Leaked Document on Digital Services Act'. European Commission / Directorate-General Connect.

EDiMA. 2020. 'EDiMA Calls for a New Online Responsibility Framework'. 7 January 2020. https://edima-eu.org/news/edima-calls-for-a-new-online-responsibility-framework/.

EDRi. 2019. 'More Responsibility to Online Platforms- but at What Cost?' EDRi.org. 19 June 2019. https://edri.org/more-responsibility-to-online-platforms-but-at-what-cost/.

EU Commission. 2016. 'The EU Code of Conduct on Countering Illegal Hate Speech Online'. Text. European Commission - European Commission. 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 Commission. 2016. 'Illegal Content on Online Platforms'. Text. Shaping Europe's Digital Future - European Commission. 29 September 2016. https://ec.europa.eu/digital-single-market/en/illegal-content-online-platforms.

———. 2019. 'Assessment on the Code of Conduct on Hate Speech on Line State of Play'.

———. 2020a. 'Communication on "Shaping Europe's Digital Future"'. 19 February 2020. https://www.europeansources.info/record/communication-on-shaping-europes-digital-future/.

———. 2020b. 'Illegal Content on Online Platforms: Shaping Europe's Digital Future'. March 2020. https://ec.europa.eu/digital-single-market/en/illegal-content-online-platforms.

European Parliament. 2020a. 'DRAFT REPORT with Recommendations to the Commission on Digital Services Act: Improving the Functioning of the Single Market'. Rapporteur: Alex Agius Saliba. Committee on the Internal Market and Consumer Protection. https://www.europarl.europa.eu/doceo/document/IMCO-PR-648474_EN.pdf.

———. 2020b. 'AMENDMENTS 599 - 919 Draft Report Alex Agius Saliba (PE648.474v02-00)'. Committee on the Internal Market and Consumer Protection. https://www.europarl.europa.eu/doceo/document/IMCO-AM-652305_EN.pdf.

Faust, Gretchen. 2017. *Hair, Blood and the Nipple Instagram Censorship and the Female Body*. Digital Environments. Ethnographic Perspectives across Global Online and Offline Spaces. Bielefeld: transcript.

Fisher, Max. 2018. 'Inside Facebook's Secret Rulebook for Global Political Speech'. *The New York Times*, 27 December 2018, sec. World. https://www.nytimes.com/2018/12/27/world/facebook-moderators.html.

Frosio, Giancarlo F. 2017. 'Reforming Intermediary Liability in the Platform Economy: A European Digital Single Market Strategy'. *Northwestern University Law Review Online* 112: 18–46.

Gee, James Paul. 2014. *How to Do Discourse Analysis: A Toolkit*. London, UNITED KINGDOM: Routledge. http://ebookcentral.proquest.com/lib/uunl/detail.action?docID=1600495.

Gill, Rosalind. 2000. *'Discourse Analysis.' Qualitative Researching with Text, Image and Sound*. Vol. 1.
Gillespie, Tarleton. 2017. 'Governance of and by Platforms'. Edited by Jean Burges, Thomas Poell, and Alice Marwick. *SAGE Handbook of Social Media*, 30.
———. 2018a. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven, UNITED STATES: Yale University Press. http://ebookcentral.proquest.com/lib/uunl/detail.action?docID=5431574.
———. 2018b. 'Platforms Are Not Intermediaries'. *Georgetown Law Technology Review* 198.
Instagram. 2019. 'Instagram Policy Changes on Self-Harm Related Content - Protecting Vulnerable Users'. Instagram Blog. 7 February 2019. https://about.instagram.com/blog/announcements/supporting-and-protecting-vulnerable-people-on-instagram.
Keller, Daphne. 2019. 'Filtering Facebook: Introducing Dolphins in the Net, a New Stanford CIS White Paper - OR - Why Internet Users and EU Policymakers Should Worry about the Advocate General's Opinion in Glawischnig-Piesczek'. 5 September 2019. /blog/2019/09/filtering-facebook-introducing-dolphins-net-new-stanford-cis-white-paper-or-why.
Knight, Will. 2019. 'Facebook's Head of AI Says the Field Will Soon "Hit the Wall"'. *Wired*, 4 December 2019. https://www.wired.com/story/facebooks-ai-says-field-hit-wall/.
Kuczerawy, Aleksandra. 2018. *Intermediary Liability and Freedom of Expression in the EU: From Concepts to Safeguards*. KU Leuven Centre for IT & IP Law Series. Cambridge, United Kingdom: Intersentia.
Kümpel, Anna Sophie, and Diana Rieger. 2019. *Wandel der Sprach- und Debattenkultur in sozialen Online-Medien: Ein Literaturüberblick zu Ursachen und Wirkungen von inziviler Kommunikation*. Ludwig-Maximilians-Universität München. https://doi.org/10.5282/ubm/epub.68880.
Laaksonen, Salla-Maaria, and et al. 2020. 'The Datafication of Hate: Expectations and Challenges in Automated Hate Speech Monitoring | Big Data'. 5 February 2020. https://www.frontiersin.org/articles/10.3389/fdata.2020.00003/full.
Murphy, Hannah, and Madhumita Murgia. 2019. 'Can Facebook Really Rely on Artificial Intelligence to Spot Abuse?' 8 November 2019. https://www.ft.com/content/69869f3a-018a-11ea-b7bc-f3fa4e77dd47.
Olszanowski, Magdalena. 2014. 'Feminist Self-Imaging and Instagram: Tactics of Circumventing Sensorship'. *Visual Communication Quarterly* 21 (2).
Roberts, Sarah T. 2018. 'Digital Detritus: "Error" and the Logic of Opacity in Social Media Content Moderation'. *First Monday*, March. https://doi.org/10.5210/fm.v23i3.8283.
———. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven, UNITED STATES: Yale University Press. http://ebookcentral.proquest.com/lib/uunl/detail.action?docID=5783696.
Rudl, Tomas, and Alexander Fanta. 2019. 'Geleaktes Arbeitspapier - EU-Kommission erwägt neues Gesetz für Plattformen'. *netzpolitik.org* (blog). 15 July 2019. https://netzpolitik.org/2019/geleaktes-arbeitspapier-eu-kommission-erwaegt-neues-gesetz-fuer-plattformen/.
Savin, Andrej. 2018. 'Regulating Internet Platforms in the EU - The Emergence of the "Level Playing Field"'. *Computer Law & Security Review* 34 (6): 1215–31. https://doi.org/10.1016/j.clsr.2018.08.008.
Stark, Birgit, and Daniel Stegmann. 2020. 'Are Algorithms a Threat to Democracy? The Rise of Intermediaries: A Challenge for Public Discourse'. Algorithm Watch. https://algorithmwatch.org/wp-content/uploads/2020/05/Governing-Platforms-communications-study-Stark-May-2020-AlgorithmWatch.pdf.
United Nations General Assembly. 2018a. 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, A/73/348'.
———. 2018b. 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, A/HRC/38/35'.

———. 2019. ‘Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression A/74/486’.
Van Dijck, José, Thomas Poell, and Martijn de Waal. 2018. *The Platform Society: Public Values in a Connective World*. Oxford, New York: Oxford University Press.
Wagner, Ben. 2019. ‘Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems’. *Policy & Internet* 11 (1): 104–22. https://doi.org/10.1002/poi3.198.
Zinke, Arnika. 2018. ‘“The Cleaners” - Die schockierende Realität hinter deinem Facebook-Feed’. Wienerin. 12 September 2018. https://wienerin.at/node/26784.
Zuckerberg, Mark. 2020a. ‘Financial Times Editorial: Big Tech Needs More Regulation’. Zuckerberg Transcripts. 1097. https://epublications.marquette.edu/zuckerberg_files_transcripts/1097/.
———. 2020b. ‘Mark Zuckerberg at Munich Security Conference’. Zuckerberg Transcripts.
———. 2020c. ‘MZ Post about Quarterly Community Update’. Zuckerberg Transcripts.

Appendix

Overview

Corpus Category Overview		Date	Pro-EU-Regulation	Pro-Corporate Governance	Less Corporate Governance	Regulation sceptic	Human Agency	Techno-deterministic
OHCHR. 2018. Report on Content Moderation. A/HRC/38/35	3rd party	Apr-18	x		x	x	x	
OHCHR. 2018. Report on AI. A/73/348	3rd party	Aug-18	x		x	x	x	
OHCHR. 2019. Report on hate speech and regulatory solutions. A/74/486	3rd party	Oct-19			x	x	x	
EDRi. 2019. "Interoperability: A way to escape toxic online environments"	3rd party	Dec-19			x			
EDRi. 2019. "A privately managed public space?	3rd party	Nov-19			x			
EDRi. 2019 Leaked Commission document. "More responsibility to online platforms– but at what cost?"	3rd party	Jul-19			x	x		
EDRi. 2019. "E-Commerce review: Opening Pandora's box?	3rd party	Jun-19				x		
EDiMA. 2020. "EDiMA calls for a new Online Responsibility Framework"	FB	Jan-20		x		x		x
Zuckerberg, Mark. 2020, "MZ post about Quarterly Community Update	FB	Jan-20	x	x				
Zuckerberg, Mark. 2020. "Financial Times editorial: Big Tech Needs More Regulation"	FB	Feb-20	x	x		x		
Zuckerberg, Mark. 2019. "Interview with Washington Post on Free Speech"	FB	Oct-19				x		
Zuckerberg, Mark. 2019. "Standing For Voice and Free Expression"	FB	Oct-19		x	x	x		x
Zuckerberg, Mark. 2020. Munich Security Conference	FB	Feb-20	x	x	x			x
European Commission / Directorate-General Connect. 2019. "Leaked Document on Digital Services Act"	EU	Jun-19	x	x	x			x
European Commission. 2020. "Shaping Europe's digital future"	EU	Feb-20	x				x	
European Commission. 2019. "Assessment of the Code of Conduct on Hate Speech on line"	EU	Sep-19					x	x
European Commission. 2020 "Illegal content on online platforms". Communication on Website	EU	Mar-20	x	x			x	x
European Commission. 2016. "The EU Code of conduct on countering illegal hate speech online"	EU	May-16		x			x	x

Discourse Analysis Coding

Color Coding: EU, Facebook, 3rd party

Parts that relate to the category have been marked **bold.**

Pro regulation and less regulation

Pro regulation	**OHCHR. 2018. Report on Content Moderation. A/HRC/38/35**[27]
	Smart regulation, not heavy-handed viewpoint-based regulation, should be the norm, focused on ensuring company transparency and remediation to enable the public to make choices about how and whether to engage in online forums (p 19)
	OHCHR. 2018. Report on AI. A/73/348 [28]
	State approaches may involve enhanced transparency and disclosure obligations on companies and robust data protection **legislation** that addresses AI-related concerns (p16)
	States should ensure that human rights are central to private sector design, deployment and implementation of AI systems. This includes **updating and applying existing regulation**, particularly data protection regulation, to the AI domain, pursuing **regulatory or co-regulatory** schemes designed to require businesses to undertake impact assessments and audits of AI technologies and ensuring effective external accountability mechanisms. Where applicable, **sectoral regulation of particular AI applications** may be necessary and effective for the protection of human rights. To the extent that such restrictions introduce or facilitate interferences with freedom of expression, States should ensure that they are necessary and proportionate to accomplish a legitimate objective in accordance with article 19 (3) of the Covenant. AI-related **regulation** should also be developed through extensive public consultation involving engagement with civil society, human rights groups and representatives of marginalized or underrepresented end users. (p 20)
	Zuckerberg, Mark. 2020, "MZ post about Quarterly Community Update"[29]
	When it comes to these important social issues, I don't think private companies should be making so many important decisions by themselves. I don't think each service should have to individually decide what content or advertising is allowed during elections, or what content is harmful overall. There should be a **more democratic process for determining these rules and regulations**. For these issues, it's not enough for us to just make principled decisions -- the decisions also need to be seen as legitimate and reflecting what the community wants. That's why I've called for **clearer regulation for our industry** (p 2)
	Zuckerberg, Mark. 2020. "Financial Times editorial: Big Tech Needs More Regulation"[30]
	I don't think **private companies should make so many decisions alone** when they touch on fundamental democratic values. That is why last year I **called for regulation** in four areas: elections, harmful content, privacy and data portability (p 2)

[27] United Nations General Assembly / Human Rights Council. 2018. "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression". A/HRC/38/35. 6 April 2018.
[28] United Nations General Assembly. 2018. Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. A/73/348. 29 August 2018
[29] Zuckerberg, Mark. 2020. "MZ post about Quarterly Community Update". Zuckerberg Transcripts . 1081. https://dc.uwm.edu/zuckerberg_files_transcripts/1081
[30] Zuckerberg, Mark. 2020. "Financial Times editorial: Big Tech Needs More Regulation". Zuckerberg Transcripts. 1097 https://dc.uwm.edu/zuckerberg_files_transcripts

	One is transparency. Governments often tell us it's hard to **design content regulation** because they don't have insight into how our systems work. Facebook already publishes more detailed reports about harmful content than any other major internet service, and we've shown regulators how our systems operate. We're also looking at opening up our content moderation systems for external audit (p 2)
	Lastly, we **need more oversight and accountability**. People need to feel that global technology platforms answer to someone, so **regulation should hold companies accountable when they make mistakes** (p 3).
	I believe **good regulation** may hurt Facebook's business in the near term but it will be better for everyone, including us, over the long term. These are problems that need to be fixed and that affect our industry as a whole. If we don't create standards that people feel are legitimate, they won't trust institutions or technology (p 3).
	Zuckerberg, Mark. 2020. Munich Security Conference [31]
	Um, but at some level, I- I do think that we don't want private companies making so many decisions about how to balance social equities without a more democratic process. So I- I think that where- where the lines, i- in my opinion should be drawn, is there should be more, uh, **guidance and regulation from the States** on, um, on- on basically on- on what kind of, you know, take political advertising as an example, um, you know, what discourse should be allowed, um, or on the balance of, um, of free expression and- and some things that people call harmful expression, where do you draw the line? What kinds of systems should companies have to- have to develop? (21:27; p 9)
	European Commission / Directorate-General Connect. 2019. "Leaked Document on Digital Services Act"[32]
	Digital collaborative economy services increasingly face **uncoordinated national or even regional regulation** of their services and no standards exists for information exchange with local or national authorities (e.g. on tax matters). As a result of this legal fragmentation lack of enforcement (e.g. of the E-Commerce Directive), and the **lack of information for regulators**, home-grown collaborative economy start-ups such as Taxify cannot scale-up across the EU and grow to compete with US rivals such as Uber (p 1)
	The extremely fast evolution of digital services and the high complexity of issues resulting from the wide take-up of digital services raises **structural problems in the ability of regulators to implement, enforce and adapt rules** dynamically and in a timely and effective manner. Although digital services regulators exist for Data Protection, Audio-visual media, Competition, Electronic Communication Services, and Consumer Protection etc, there is currently **no dedicated "platform regulator" in the EU, which could exercise effective oversight and enforcement**, e.g. in areas such as content moderation or advertising transparency. Many of the existing regulators also lack the digital capacities needed to interface with online platforms today. At the same time, **no regulatory authority** is presently available to provide quick and reliable EU-wide guidance on emerging, unforeseen issues, such as the recent organised abuse of multiple platforms seen in the Christchurch attack, or such as the ever-changing issues around online harms for minors (p 2)
	The perceived **lack of control over the activities of globally operating service providers** is also one of the drivers for increasing national regulatory activity in this area (p 3)
	Regulating content moderation. Uniform rules for the removal of illegal content such as illegal hate speedi would be made binding across the EU, building on the Recommendation on illegal content and relevant case-law, and include a robust set of fundamental rights safeguards. Such notice-and action rules could be tailored to the types of services, e.g. whether the service is a social network, a mere conduit, or a collaborative economy service, and where necessary to the types of content in question, while maintaining the maximum simplicity of rules. The feasibility of introducing thresholds could be examined in this context, taking due account of the size and nature of the service provider and of the nature of

[31] Zuckerberg, Mark. 2020. "Mark Zuckerberg at Munich Security Conference". Zuckerberg Transcripts. 1091 https://dc.uwm.edu/zuckerberg_files_transcripts
[32] European Commission / Directorate-General Connect. 2019. "Leaked Document on Digital Services Act". Retrieved via "Netzpolitik.org", 16.7.2019. https://netzpolitik.org/2019/leaked-document-eu-commission-mulls-new-law-to-regulate-online-platforms/

	the potential Obligations to be imposed on them. Building on the Recommendation on illegal Content, binding transparency Obligations would also be at the heart of a **more effective accountability framework for content moderation at scale,** and would complement recently adopted rules on AVMS or Copyright. Options for transparency for algorithmic recommendation systems of public relevance such as newsfeeds should also be examined. At the same time, these rules would avoid that Member States impose parallel transparency Obligations at national level, thus providing for a simple set of rules in the Single Market (p 5)
	Finally, a binding "Good Samaritan provision" would **encourage and incentivise proactive measures**, by clarifying the lack of liability as a result of Such measures, on the basis of the notions already included in the Illegal Content Communication. (p 5)
	However, a clear **distinction will be made between illegal and harmful content when it comes to exploring policy options**. For instance, the ever changing nature of harmful content seems to make it unsuitable for strict notice and action type Obligations; in case of harmful content, codes of conduct and user empowerment in choosing sources could be given higher prominence; **the role of the regulator could be strengthened** (e.g. via approval of such codes of conduct) (p 5-6).
	Regulatory oversight. A **dedicated regulatory structure should ensure oversight and enforcement of the rules**, in particular for cross-border situations, but also partnerships and guidance for emerging issues, and with appropriate digital capacities and competences, inter alia to help translate rules into technical solutions. **The nature of the regulatory structure will depend on the specific mission**, and could involve a **central regulator, a decentralised system, or an extension of powers of existing regulatory authorities**. Possible roles and powers of such **regulatory structures** will be explored, including reporting requirements, powers to require additional information, complaint handling, the power to impose fines or other corrective action, or the approval of codes of conduct. This analysis will draw on external advice (e-g. through the Observatory of the Online Economy) and any insights to be gained from **existing or planned regulatory structures**, both at EU and national level (p 6).
	European Commission. 2020. "Shaping Europe's digital future"[33]
	Key actions - New and revised rules to deepen the Internal Market for Digital Services, by **increasing and harmonising the responsibilities of online platforms and information service providers and reinforce the oversight over platforms' content policies** in the EU. (Q4 2020, as part of the Digital Services Act package).
	European Commission. 2020 "Illegal content on online platforms". Communication on Website[34]
	The Commission considers that online intermediaries can put in place proactive measures **without losing the liability exemption the the e-Commerce Directive.**

Regulation-sceptic	**OHCHR. 2018. Report on AI. A/73/348 [35]**
	States, meanwhile, are pressing for efficient, speedy automated moderation across a range of separate challenges, from child sexual abuse and terrorist content, where AI is already extensively deployed, to copyright and the removal of '"extremist" and "hateful" content. The

[33] European Commission. 2020. "Shaping Europe's digital future". COM(2020) 67 final. 19.2.2020.
[34] European Commission. 2020. "Illegal content on online platforms. Policy". ec.europa.eu, accessed on March, 23 2020. https://ec.europa.eu/digital-single-market/en/illegal-content-online-platforms
[35] United Nations General Assembly. 2018. Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. A/73/348. 29 August 2018

	European Commission Recommendation on measures to further improve the effectiveness of the fight against illegal content online of March 2018 calls upon Internet platforms to use automatic filters to detect and remove terrorist content, with human review in some cases suggested as a **necessary counterweight to the inevitable errors caused by the automated systems** (p 8)
	OHCHR. 2018. Report on Content Moderation. A/HRC/38/35[36]
	In the light of legitimate State concerns such as privacy and national security, the appeal of regulation is understandable. However, such rules **involve risks to freedom of expression, putting significant pressure on companies such that they may remove lawful content in a broad effort to avoid liability.** They also involve the delegation of regulatory functions to private actors that lack basic tools of accountability. Demands for quick, automatic removals risk new forms of prior restraint that already threaten creative endeavours in the context of copyright. Complex questions of fact and law should generally be **adjudicated by public institutions, not private actors** whose current processes may be inconsistent with due process standards and whose motives are principally economic (p 7)
	The 2016 European Union Code of Conduct on countering illegal hate speech online involves agreement between the European Union and four major companies to remove content, committing them to collaborate with "trusted flaggers" and promote "independent counter-narratives". While the promotion of counter-narratives may be attractive in the face of "extremist" or "terrorist" content, **pressure for such approaches runs the risk of transforming platforms into carriers of propaganda well beyond established areas of legitimate concern** (p 8)
	Government demands not based on national law. Companies distinguish between requests for the removal of allegedly illegal content submitted through regular legal channels and requests for removal based on the companies' terms of service. (Legal removals generally apply only in the requesting jurisdiction; terms of service removals generally apply globally.) **State authorities increasingly seek content removals outside of legal process or even through terms of service requests.** Several have established specialized government units to refer content to companies for removal. The European Union Internet Referral Unit, for instance, "flag[s] terrorist and violent extremist content online and cooperat[es] with online service providers with the aim of removing this content" (p 8)
	OHCHR. 2019. Report on hate speech and regulatory solutions. A/74/486[37]
	Legislative efforts to incentivize the removal of online hate speech and impose liability on Internet companies for the failure to do so must meet the necessity and proportionality standards identified above. In recent years, States have pushed companies towards a nearly immediate takedown of content, demanding that they develop filters that would disable the upload of content deemed harmful. **The pressure is for automated tools that would serve as a form of pre-publication censorship**. Problematically, an upload filter requirement "would enable the blocking of content without any form of due process even before it is published, reversing the wellestablished presumption that States, not individuals, bear the burden of justifying restrictions on freedom of expression". Because such filters are notoriously unable to address the kind of natural language that typically constitutes hateful content, they can cause significant disproportionate outcomes. Furthermore, there is research suggesting that such filters disproportionately harm historically underrepresented communities (p 14)
	It is useful to contemplate a hypothetical State that is considering legislation that would hold online intermediaries liable for the failure to take specified action against hate speech. Such an "intermediary liability" law is typically aimed at restricting expression, whether of the users of a particular platform or of the platform itself, sometimes with a view to fulfilling the obligation under article 20 (2) of the Covenant. Any legal evaluation of such a proposal must address the cumulative conditions established under article 19 (3) to ensure consistency with international standards on free expression (p 13)

[36] United Nations General Assembly / Human Rights Council. 2018. "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression". A/HRC/38/35. 6 April 2018.
[37] United Nations General Assembly. 2019. "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression". A/74/486. 9 October 2019.

	EDRi. 2019 Leaked Commission document. "More responsibility to online platforms– but at what cost?"[38]
	The Commission acknowledges that when platform companies are pushed to take measures against potentially illegal and harmful content, their balancing of interests pushes them to over-block legal speech and monitor people's communications to prevent legal liability for user content. At the same time, the note proposes that harmful content should best be dealt with through voluntary codes of conduct, which **shifts the censorship burden to the platform companies**. However, companies' terms of service are often a convenient way of removing legal content as they are vague and redress mechanisms are often ineffective. Drawing from the experience of the EU's Code of Conduct on Hate Speech and the Code of Practice on Disinformation (https://edri.org/civil-society-calls-for-evidence-basedsolutions-to-disinformation/), **this approach pushes platform companies to measure their success only based on the number of deleted accounts or removed pieces of content as well as on how speedy those deletions have been carried out. It does not, however, improve legal certainty for users, nor does it provide for proper review and counter-notice mechanisms, or allow for investigations into whether or not the removed material was even illegal**
	The leaked Commission note claims that recent sector-specific content regulation laws such as the disastrous Copyright Directive (https://edri.org/censorship-machine-takesover-eu-internet/) or the proposed Terrorist Content Regulation (https://edri.org/openletter-regulation-on-terrorist-content-online-endangers-freedom-of-expression/) had left "most of" the current E-Commerce Directive unaffected. This is euphemistic at the very least. According to these pieces of legislation, **all online platforms are required to pro-actively monitor and search for certain types of content to prevent their upload, which makes them "active" under current case law and should flush their liability exemption down the toilet.** This is not changed by the Copyright Directive's claim on paper that it shall not affect the E-Commerce's liability rules.
	EDRi. 2019. "E-Commerce review: Opening Pandora's box?"[39]
	(…) we witness more government pressure on companies to implement voluntary mechanisms against alleged illegal or "harmful" content. These two parallel developments **resulted in an increasing number of wrongful removals and blocking of legitimate speech**. In the past months, the Directorate-General for Communications Networks, Content and Technology (DG Connect) of the EU Commission already started the process of exploring policy options for content moderation that will be presented to the incoming College of Commissioners. A reform of the ECD to attempt the harmonisation of liability exemptions and content moderation rules seems to have become unavoidable.
	The big question is: will the review of the ECommerce Directive (ECD) (https://ec.europa.eu/digital-singlemarket/en/e-commerce-directive) open Pandora's box and become one of this decade's **biggest threat to citizens' rights and freedoms online – or will it be a chance to clarify and improve the current situation?**
	Zuckerberg, Mark. 2020. "Financial Times editorial: Big Tech Needs More Regulation"[40]
	Of course, we won't agree with every proposal. Regulation can have unintended consequences, especially for small businesses that can't do sophisticated data analysis and marketing on their own. Millions of small businesses rely on companies like ours to do this for them. If regulation makes it harder for them to share data and use these tools, that could disproportionately hurt them and inadvertently advantage larger companies that can (p 3)
	EDiMA. 2020. "EDiMA calls for a new Online Responsibility Framework"[41]

[38] EDRi. 2019. "More responsibility to online platforms- but at what cost?" EDRi.org. June 19th 2019. https://edri.org/more-responsibility-to-online-platforms-but-at-what-cost/
[39] Fiedler, Kristen. 2019. "E-Commerce review: Opening Pandora's box?". EDRi.org. https://edri.org/e-commerce-review-1-pandoras-box/
[40] Zuckerberg, Mark. 2020. "Financial Times editorial: Big Tech Needs More Regulation". Zuckerberg Transcripts. 1097 https://dc.uwm.edu/zuckerberg_files_transcripts
[41] EDiMA. 2020. "EDiMA calls for a new Online Responsibility Framework". January 20th, 2020. https://edima-eu.org/news/edima-calls-for-a-new-online-responsibility-framework/

	The law should continue to **assign primary liability to those users that act illegally** or harm others and limit the liability of online service providers whose services are abused by others. The notice and action regime which accompanies the **limited liability regime** should remain the key set of rules governing specific illegalities – and in fact further clarity on notice and action rules would be welcome (p 2)
	Built-in safeguards would be required to ensure that measures taken under this framework of responsibility would **not compromise service providers' limited liability**. This would reconcile responsibility with online service providers' freedom to conduct a business, the need for legal certainty for both private sectors and competent authorities, and **ensure that service providers are not perversely incentivised to interfere** with their users' fundamental rights. To do so, it would be important to retain the prohibition of a general monitoring obligation, and the concepts of reasonableness, proportionality, and feasibility would need to be interpreted in a good faith manner by competent authorities and courts (p 2-3)
	In this way the new approach can co-exist with current rules and provide an overarching framework for responsibility online, while also making it possible to adapt quickly to address emerging concerns in the online space where there is concrete evidence that more specific vertical measures are needed (p 3)
	The scope of the new framework should be **broadly defined, technology-neutral, and principles-based, applying proportionately to a variety of different online services rather than a specific list** – which can become outdated or inapplicable in time. This is preferable to a patchwork approach which includes certain services in scope on the basis of criteria such as the size of the company or the type of content involved, as digital services are dynamic by design. The principles-based approach would establish a sliding scale of different measures that **allows service providers to react appropriately to the concerns that are specific to their services and in a manner that is commensurate with their unique situations and abilities**. The concepts of proportionality and feasibility would then take account of situations where the nature of the service requires a different approach. For example, services such as electronic communications service providers and cloud infrastructure providers are more limited in what they can do to address illegal content uploaded or shared by their users, given the technical architecture of their services and the contractual relationships they hold with users. To expect the same content management efforts from their services as that requested of public-facing content sharing services belies their technical and operational nature, and would give rise to unjustified privacy, security, and commercial interferences (p 3)
	Illegal content is more easily defined and more consistent across the EU than content which is "harmful" but not illegal - the concept of "harmful" is subjective, depends greatly on context and can vary considerably between Member States when differences in culture and language are taken into consideration. The clearer definition of illegal content in national law would permit quicker action on tackling this content. Because the management of harmful content or activity requires nuance, a specific focus on the management of illegal content and activity at EU level will help to avoid infringing on fundamental rights for more context-specific cases (p 3)
	Crucially, the focus of an **oversight body's work should be restricted to the broad measures** which service providers are taking – it should not have the power to assess the legality of individual pieces of content and it should not be empowered to issue takedown notices, which is the remit of the courts. Such competences call into play multiple critical constitutional and procedural questions, which are best left to the courts (p 4).
	Zuckerberg, Mark. 2019. "Standing For Voice and Free Expression" [42]
	We're increasingly seeing **laws and regulations around the world that undermine free expression and people's human rights**. These local laws are each individually troubling, especially when they shut down speech in places where there isn't democracy or freedom of the press. But it's even worse when countries try to impose their speech restrictions on the rest of the world (p 9)

[42] Zuckerberg, Mark. 2019. "Standing For Voice and Free Expression" Zuckerberg Transcripts . 1022. https://dc.uwm.edu/zuckerberg_files_transcripts/1022

	Increasingly, we're seeing people try to define more speech as dangerous because it may lead to political outcomes they see as unacceptable. Some hold the view that since the stakes are so high, they can no longer trust their fellow citizens with the power to communicate and decide what to believe for themselves. I personally believe this is more dangerous for democracy over the long term than almost any speech. Democracy depends on the idea that we hold each others' right to express ourselves and be heard above our own desire to always get the outcomes we want. You **can't impose tolerance top-down**. It has to come from people opening up, sharing experiences, and developing a shared story for society that we all feel we're a part of. That's how we make progress together (p 10)
	Zuckerberg, Mark. 2019. "Interview with Washington Post on Free Speech" [43]
	On Thursday, though, Zuckerberg's message served more as a warning that **overreaction could stifle the very sort of speech that many regulators seek to protect** (p 4)

Human agency and techno-determinism

Human agency	**OHCHR. 2018. Report on Content Moderation. A/HRC/38/35**[44]
	"The development of content moderation policies typically involves **legal counsel, public policy and product managers, and senior executives**" (p 10).
	Companies should articulate the **bases for such restrictions**, however, and demonstrate the necessity and proportionality of any content actions (such as removals or account suspensions) (p 11)
	"Meaningful examination of context may be thwarted by time and resource constraints **on human moderators**, overdependence on automation or insufficient understanding of linguistic and cultural nuance. Companies have urged users to supplement controversial content with contextual details, but the feasibility and effectiveness of this guidance are unclear" (p 11).
	"**User flags give individuals** the ability to log complaints of inappropriate content with **content moderators**. Flags typically do not enable nuanced discussions about appropriate boundaries (e.g., why content may be offensive but, on balance, better left up). They have also been "gamed" to heighten pressure on platforms to remove content supportive of sexual minorities and Muslims. Many companies have developed specialized rosters of **"trusted" flaggers, typically experts, high-impact users and, reportedly, sometimes government flaggers**. There is little or no public information explaining the selection of specialized **flaggers**, their interpretations of legal or community standards or their influence over company decisions. (p12-13)
	"The massive scale of user-generated content has led the largest companies to develop automated moderation tools. Automation has been employed primarily to flag content for **human review**, and sometimes to remove it." (p 12)
	"Automation may provide value for companies assessing huge volumes of usergenerated content, with tools ranging from keyword filters and spam detection to hashmatching algorithms and natural language processing. Hash matching is widely used to identify child sexual abuse images, but its

[43] Zuckerberg, Mark and Romm, Tony. 2019. "Interview with Washington Post on Free Speech". Zuckerberg Transcripts . 1023. https://dc.uwm.edu/zuckerberg_files_transcripts/1023 ashington Post

[44] United Nations General Assembly / Human Rights Council. 2018. "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression". A/HRC/38/35. 6 April 2018.

	application to “extremist” content — which typically requires assessment of context — is difficult to accomplish without clear rules regarding “extremism” or **human evaluation**. The same is true with natural language processing" (p 12)
	Automation often will be supplemented **by human review**, with the biggest social media companies developing **large teams of content moderators** to review flagged content. Flagged content may be routed to **content moderators**, which will typically be authorized to make a decision — often within minutes — about the appropriateness of the content and to remove or permit it. In situations where the appropriateness of particular content is difficult to determine, **moderators** may escalate its review to content teams at company headquarters. In turn, company officials — typically public policy or “trust and safety” **teams** with the engagement of general counsel — will make decisions on removals. Company disclosure about removal discussions, in aggregate or specific cases, is limited (p 13)
	Automated content moderation, a function of the massive scale and scope of user-generated content, poses distinct risks of content actions that are inconsistent with **human rights law**. Company responsibilities to prevent and mitigate **human rights impacts** should take into account the significant limitations of automation, such as difficulties with addressing context, widespread variation of language cues and meaning and linguistic and cultural particularities. Automation derived from understandings developed within the home country of the company risks serious discrimination across global user bases. At a minimum, technology developed to deal with considerations of scale should be rigorously audited and developed with broad user and civil society input. The responsibility to foster accurate and context-sensitive content moderation practices that respect freedom of expression also requires companies to strengthen and ensure professionalization of their **human evaluation** of flagged content. This strengthening should involve protections for **human moderators** consistent with human rights norms applicable **to labour rights** and a serious commitment to involve cultural, linguistic and other forms of expertise in every market where they operate. Company leadership and policy teams should also diversify to enable the application of local or subject-matter expertise to content issues. (p18)
Human agency	**OHCHR. 2018. Report on AI. A/73/348** [45]
	In all circumstances, **humans play a critical role in designing and disseminating AI systems**, defining the objectives of an AI application and, depending on the type of application, selecting and labelling datasets and classifying outputs. **Humans** always **determine** the application and use of AI outputs, including the extent to which they complement or **replace human decision-making** (p 4)
	At the foundation of AI are algorithms, computer code **designed and written by humans**, carrying instructions to translate data into conclusions, information or outputs (p 4)
	Human agency is integral to AI, but the distinctive characteristics of **AI deserve human rights scrutiny with** respect to at least three of its aspects: automation, data analysis and adaptability (p 5)
	Social media companies use AI to filter content across the range of their rules (from nudity to harassment to hate speech and so on), although the extent to which such companies rely on **automation without human input** on specific cases is not known (p 8)
	Support and pressure for increasing the role of AI come from both the private and public sectors. Companies claim that the volume of illegal, inappropriate and harmful content online far exceeds the capabilities of **human moderation** and argue that AI is one tool that can assist in better tackling this challenge. According to some platforms, AI is not only more efficient in identifying inappropriate (according to their rules) and illegal content for removal (usually by a **human moderator**) but also has a higher accuracy rate than **human decision-making**. States, meanwhile, are pressing for efficient, speedy automated moderation across a range of separate challenges, from child sexual abuse and terrorist content, where AI is already extensively deployed, to copyright and the removal of ‘“extremist” and “hateful” content. The European Commission Recommendation on measures to further improve the effectiveness of the fight against illegal content online of March 2018 calls upon Internet platforms to use automatic

[45] United Nations General Assembly. 2018. Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. A/73/348. 29 August 2018

	filters to detect and remove terrorist content, **with human review** in some cases suggested as a necessary counterweight to the inevitable errors caused by the automated systems. (p 8)
	Efforts to automate content moderation may come at a cost to **human rights** (see A/HRC/38/35, para. 56). AI-driven content moderation has several limitations, including the challenge of assessing context and taking into account widespread variation of language cues, meaning and linguistic and cultural particularities. Because AI applications are often grounded in datasets that **incorporate discriminatory assumptions**, and under circumstances in which the cost of overmoderation is low, there is a high risk that such systems will default to the removal of online content or suspension of accounts that are not problematic and that content will be removed in accordance with biased or discriminatory concepts. As a result, vulnerable groups are the most likely to be disadvantaged by AI content moderation systems. For example, Instagram's DeepText identified "Mexican" as a slur because its datasets were populated with data in which "Mexican" was associated with "illegal", a negatively coded term baked into the algorithm (p 8-9)
	AI makes it difficult to scrutinize the logic behind content actions. Even when algorithmic content moderation is complemented by **human review** — an arrangement that large social media platforms argue is increasingly infeasible on the scale at which they operate — a tendency to defer to machine-made decisions (on the assumptions of objectivity noted above) impedes interrogation of content moderation outcomes, especially when the system's technical design occludes that kind of transparency (p 9)
	The complexity of decision-making inherent in content moderation may be exacerbated by the introduction of automated processes. Unlike **humans**, algorithms are today not capable of evaluating cultural context, detecting irony or conducting the critical analysis necessary to accurately identify, for example, "extremist" content or hate speech and are thus more likely to default to content blocking and restriction, undermining the rights of individual users to be heard as well as their right to access information without restriction or censorship. (…) In an AI-governed system, the dissemination of information and ideas is governed by opaque forces with priorities that may be at odds with an enabling environment for media diversity and independent voices. Relevantly, the **Human Rights Committee** has found that States should "take appropriate action … to prevent undue media dominance or concentration by privately controlled media groups in monopolistic situations that may be harmful to a diversity of sources and views" (p 12)
Human agency	**OHCHR. 2019. Report on hate speech and regulatory solutions. A/74/486** [46]
	Finally, the companies must also train their **content policy teams, general counsel and especially content moderators** in the field, that is, those conducting the actual work of restriction (principle 16, commentary). As part of the training, the norms **of human rights law** that the content moderation is aimed at protecting and promoting should be identified. (p 17)
	Companies may find that detailed contextual analysis is difficult and **resourceintensive**. The largest companies rely heavily on automation in order to do at least the first-layer work of identifying hate speech, which requires having rules that divide content into either one category (ignore) or another (delete). They use the power of artificial intelligence to drive these systems, but the systems are notoriously bad at evaluating context (see A/73/348). However, if the companies are serious about protecting **human rights** on their platforms, they must ensure that they define the rules clearly and require **human evaluation**. Human evaluation, moreover, must be more than an assessment of whether particular words fall into a particular category. It must be based on real learning from the communities in which hate speech may be found, that is, **people who can understand the "code"** that language sometimes deploys to hide incitement to violence, evaluate the speaker's intent, consider the nature of the speaker and audience and evaluate the environment in which hate speech can lead to violent acts. **None of these things are possible with artificial intelligence alone**, and the definitions and strategies should reflect the nuances of the problem. The largest companies should bear the

[46] United Nations General Assembly. 2019. "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression". A/74/486. 9 October 2019.

	burden of these resources and share their knowledge and tools widely, as open source, to ensure that smaller companies, and smaller markets, have access to such technology (p 19-20)
	Ensure that any enforcement of hate speech rules involves an **evaluation of context** and the harm that the content imposes on **users and the public**, including by ensuring that any use of automation or artificial intelligence tools involve **human-in-the-loop**; (p 23)
Human agency	**European Commission. 2020. "Shaping Europe's digital future"[47]**
	"Initiative to improve **labour conditions of platform workers** 2021" (p 7)
	European Commission. 2020 "Illegal content on online platforms". Communication on Website[48]
	Online platforms should set out easy and transparent rules for notifying illegal content, including fast-track procedures for **'trusted flaggers'**. Content providers should be informed about such decisions and have the opportunity to contest them in order to avoid unintended removal of legal content
	Decisions to remove content are accurate and well-founded, especially when automated tools are used, companies should put in place effective and appropriate safeguards, including **human oversight and verification**, in full respect of fundamental rights, freedom of expression and data protection rules.
	European Commission. 2019. "Assessment of the Code of Conduct on Hate Speech online"[49]
	All platforms have also significantly increased **the number of employees monitoring** and reviewing the content. Facebook reports having a global network of about 15,000 **people working** on all types of content review and across Google and YouTube there are more than 10,000 **people working** to address content that may violate the company's policies (p 3)
	"**human moderation of hate speech**" (p 4)
	Since the signature of the Code, Facebook/Instagram have organised a total of 51 **training sessions** on its community standards in relation to hate speech, for up to 130 civil society organisations operating **as trusted flaggers**. Out of 38 training sessions provided in 2018 by YouTube to NGOs on their content policy and trusted flagger programme, 18 were focused on hate speech and abusive content. (p 4)
	It should be noted that all content surfaced by automatic detection system is assessed by the team of reviewers before being actioned (**human in-the-loop**). (p 7)
	European Commission. 2016. "The EU Code of conduct on countering illegal hate speech online"[50]
	Upon receipt of a valid removal notification, the IT Companies to review such requests against their rules and community guidelines and where necessary national laws transposing the Framework Decision 2008/913/JHA, with **dedicated teams reviewing** requests (p 2)

Techno-deter	**European Commission. 2016. "The EU Code of conduct on countering illegal hate speech online"[51]**
	The IT Companies to review the majority of valid notifications for removal of illegal hate (p 2)

[47] European Commission. 2020. "Shaping Europe's digital future". COM(2020) 67 final. 19.2.2020.
[48] European Commission. 2020. "Illegal content on online platforms. Policy". ec.europa.eu, accessed on March, 23 2020. https://ec.europa.eu/digital-single-market/en/illegal-content-online-platforms
[49] European Commission. 2019. "Assessment of the Code of Conduct on Hate Speech online". 27 September 2019.
[50] European Commission. 2016. "The EU Code of conduct on countering illegal hate speech online". May 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
[51] European Commission. 2016. "The EU Code of conduct on countering illegal hate speech online". May 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

minism	**European Commission / Directorate-General Connect. 2019. "Leaked Document on Digital Services Act"[52]**
	The lack of legal clarity also entails a **regulatory disincentive for platforms and other intermediaries to act proactively to tackle illegal content**, as well as to **adequately address harmful content online**, especially when combined with the issue of fragmentation of rules addressed above. As a consequence, many digital services avoid **taking on more responsibility in tackling illegal content**, for fear of becoming liable for content they intermediate. (p 2)
	Finally, a binding "Good Samaritan provision" would **encourage and incentivise proactive measures**, by clarifying the lack of liability as a result of Such measures, on the basis of the notions already included in the Illegal Content Communication. (p 5)
	General monitoring and automated filtering. While the prohibition of general monitoring Obligations should be maintained as another foundational cornerstone of Internet regulation, sp**ecific provisions governing algorithms for automated filtering technologies - where these are used- should be considered,** to provide the necessary transparency and accountability of automated content moderation Systems (p 5)
	European Commission. 2019. "Assessment of the Code of Conduct on Hate Speech on line"[53]
	The removal rate is now stable at more than 70% on average. In 2016, after the first monitoring exercise on the implementation of the Code of conduct only 28% of the content flagged was removed. The **current average removal rate can be considered as satisfactory** in an area such as hate speech, given that the line against speech that is protected by the right to freedom of expression is not always easy to draw and is highly dependent on the context in which the content was placed (p 3)
	European Commission. 2020 "Illegal content on online platforms". Communication on Website[54]
	More efficient tools and proactive technologies: Companies should set out clear notification systems for users. They should have **proactive tools to detect** and remove illegal content, in particular for terrorism content and for content which does not need contextualisation to be deemed illegal, such as child sexual abuse material or counterfeited goods.
	The Commission considers that online intermediaries can put in place **proactive measures** without losing the liability exemption the the e-Commerce Directive.
	Zuckerberg, Mark. 2019. "Standing For Voice and Free Expression" [55]
	Our **AI systems have also gotten more advanced at detecting** clusters of fake accounts that aren't behaving like humans. We now remove billions of fake accounts a year — most within minutes of registering and before they do much. **Focusing on authenticity and verifying accounts is a much better solution than an ever-expanding definition of what speech is harmful**. (p 5)
	We're particularly focused on well-being, especially for young people. We built a team of thousands of people and **AI systems that can detect** risks of self-harm within minutes so we can reach out when people need help most. In the last year, we've helped first responders reach people who needed help thousands of times. For each of these issues, I believe we have two responsibilities: to remove content when it could cause real danger as effectively as we can, and to fight to uphold as wide a definition of freedom of expression as possible — and not allow the definition of what is considered dangerous to expand beyond what is absolutely necessary. That's what I'm committed to. (p 5)
	EDiMA. 2020. "EDiMA calls for a new Online Responsibility Framework"[56]

[52] European Commission / Directorate-General Connect. 2019. "Leaked Document on Digital Services Act". Retrieved via "Netzpolitik.org", 16.7.2019. https://netzpolitik.org/2019/leaked-document-eu-commission-mulls-new-law-to-regulate-online-platforms/

[53] European Commission. 2019. "Assessment of the Code of Conduct on Hate Speech on line". 27 September 2019.

[54] European Commission. 2020. "Illegal content on online platforms. Policy". ec.europa.eu, accessed on March, 23 2020. https://ec.europa.eu/digital-single-market/en/illegal-content-online-platforms

[55] Zuckerberg, Mark. 2019. "Standing For Voice and Free Expression" Zuckerberg Transcripts. 1022. https://dc.uwm.edu/zuckerberg_files_transcripts/1022

[56] EDiMA. 2020. "EDiMA calls for a new Online Responsibility Framework". January 20th 2020. https://edima-eu.org/news/edima-calls-for-a-new-online-responsibility-framework/

	EDiMA envisages a new 'Online Responsibility Framework', that would enable and incentivise online service providers **to do more to protect consumers from illegal content.** Such a system can only work if online service providers know they won't be punished for **taking additional measures**, so limited liability must be reaffirmed as part of any new framework (p 1)
	"Our members understand and share the concern that people have about illegal and harmful content online and we want to do more to tackle this problem. We need rules that allow us to take more responsibility online and these rules should encourage, not discourage further action" – Siada El Ramly, Director General of EDiMA (p 1)
	A new framework of responsibility could then set out roles and responsibilities for online service providers to tackle illegal content while respecting the unique features of the services. Responsibility in this sense would mean systemic steps, processes and procedures which a service provider can put in place to address illegal content or activity more proactively (p 2)
	Zuckerberg, Mark. 2020. Munich Security Conference [57]
	But- but **it is evolving**, um, and, and like say they are **improving**. So there are different kinds of threats that we see. So moving on from elections for a second, um, that's certainly one kind of content issue, but **we also have issues around things like hate speech**. (09:05): And one of the differences between hate speech in elections is that the people who go out who- **who say hateful things aren't necessarily getting smarter at saying hateful things**. So as **the AI systems get better**, we generally are just catching more and more of the hate speech, um, and are able to take it down. (09:24): And it's- **it's not like, um, like that... like hate speech is getting more sophisticated**. So are there... it's... I think that as the systems get better, we will get closer and closer to having a lower prevalence of that on the systems worse than something that is adversarial, like elections, um, or- or election interference, it- we just need to stay on top of it. (09:42): And I- I think we can't take for granted. **This isn't a problem that you ever fully solve.** Um, we i- we'll- we'll keep on needing to work on the- the defenses. Um, but at this point I do feel like we're improving faster, uh, than- than the adversaries. And there've been a track record, um, since 2016 of a number of very important high profile major elections, which I think there have been relatively clean results, um, and- and online discourse in that I think can give us some confidence going forward." (08:52-10:09)
	"So just to d- kind of elucidate on what I mean by that. When- when I got started, you know, I- I started the company in my dorm room, um, and you know, back then, obviously we could not have 35,000 people doing content and security review. Um, **the AI 16 years ago did not exist at the same level that it does today to, um, to identify this type of harmful stuff**. (12:37): So basically the way that the company ran for the first 12 years, um, was that people in the community, if they saw something that they thought was harmful, they would flag it for us and we would look at it reactively. And I... for a while, I- that was reasonable, but then, you know, we got to a point where, you know, we're a large enough scale company that we should be able to have a multibillion-dollar effort on content and security review. (13:00): **The AI technology evolved to the point** where now we can **proactively identify a lot of different types of content** so we have a responsibility to do that. But going from reactive to proactive on this was a multiyear journey. There are, you know, elections is one type of... th- is- is one type of the area that we're worried about, but there were about 20 different areas of- of dangerous and harmful content that we track everything from terrorist propaganda, to child exploitation, to incitement of violence, to hate speech, to... just go down the list. There are about 20 different types of categories. And the way that we judge ourselves is every six months we issue a transparency report of how much of this type of content are we finding on the service and what percent of the content on the service are our AI and other systems identifying and taking down before it's reported to us, uh, by someone else. " (12:18-13:47)
	"Um, some are harder than others. So for example, hate speeches is a **particularly challenging one because we have to be able to train AI systems to detect really small nuances, right?** If someone posting a video of a racist attack because they're condemning it, which probably means they should be able to say that or are they, um, subtly encouraging other people to copy that attack. Um, and that, you know, multiply that

[57] Zuckerberg, Mark. 2020. „Mark Zuckerberg at Munich Security Conference". Zuckerberg Transcripts. 1091 https://dc.uwm.edu/zuckerberg_files_transcripts

	challenge of kind of that subtlety linguistically by, you know, 150 languages around the world where we operate and the- the ability to make mistakes we're taking down the wrong kind of thing, um, but we're making progress. (14:44): 24 months ago on hate speech, we were at 0%. We **were taking down proactively, and I think today we're at around 80%.** So it's, um, so it's- it's accelerating. Um, it is- it's- it's a hard problem. I don't know if we'll get that one to 99% anytime soon, but as **AI continues improving,** I think we're- we're gonna... tha- that's a tailwind and as we keep on investing in the technology, we'll be able to keep on doing better and better on this, but it's a- it's a long-term investment. " (14:09-15:10)
	"Now, **as AI gets better**, we'll be able to more **efficiently filter** out more of the bad stuff. Um, and- and- and I think we have a responsibility to do that better and with increasing precision. And I- I- I think that companies should have to publish transparency reports like we do on the volume of content that they find or is reported to them, um, have to publish what percent they're able to **identify proactively** and should have to show good faith and- and ability to improve on funding more, uh, over time" (26:04-26:30)

More and less corporate governance

More corporate govern-ance	**EDiMA. 2020. "EDiMA calls for a new Online Responsibility Framework"[58]**
	EDiMA envisages a new 'Online Responsibility Framework', that would enable and incentivise online service providers **to do more to protect consumers from illegal content**. Such a system can only work if **online service providers know they won't be** punished for taking additional measures, so limited liability must be reaffirmed as part of any new framework (p1)
	"Our members understand and share the concern that people have about illegal and harmful content online and we want to **do more to tackle this problem. We need rules that allow us to take more responsibility online and these rules should encourage, not discourage further action**" – Siada El Ramly, Director General of EDiMA.
	Under this framework, a service provider within the scope would then be in a position to **take reasonable, proportionate and feasible actions** to mitigate observed issues arising from the presence of illegal content or activity on their services. **Service providers would define the kind of measures** which best suit their unique situation, and which are the least intrusive for users (p 2)
	Built-in safeguards would be required to ensure that measures taken under this framework of responsibility **would not compromise service providers' limited liability**. This would reconcile responsibility with online service providers' freedom to conduct a business, the need for legal certainty for both private sectors and competent authorities, and ensure that service providers are **not perversely incentivised to interfere** with their users' fundamental rights (p 2)
	Ultimately, this new framework for responsibility would incentivise and give confidence to online service providers **to take additional effective action against illegal content** and activity on their services, in a manner that preserves the **foundational legal principles of the open internet**. This framework should also be complimented by analysis and action on the responsibilities that other actors in the online ecosystem can take to meet the objectives of tackling illegal content.(p 3)
	Zuckerberg, Mark. 2020, "MZ post about Quarterly Community Update"[59]
	And until we get clearer rules or establish other mechanisms of governance, I expect we and our whole industry will continue to face a very high level of scrutiny. During this, our job is to keep doing what we think is right on the social issues, and to stay focused on continuing to deliver product improvements and better experiences for our community. (p 1-2)

[58] EDiMA. 2020. "EDiMA calls for a new Online Responsibility Framework". January 20th, 2020. https://edima-eu.org/news/edima-calls-for-a-new-online-responsibility-framework/
[59] Zuckerberg, Mark. 2020. "MZ post about Quarterly Community Update". Zuckerberg Transcripts . 1081. https://dc.uwm.edu/zuckerberg_files_transcripts/1081

	Zuckerberg, Mark. 2020. "Financial Times editorial: Big Tech Needs More Regulation"[60]
	To be clear, this isn't about passing off responsibility. **Facebook is not waiting for regulation; we're continuing to make progress** on these issues ourselves (p 3)
	Zuckerberg, Mark. 2019. "Standing For Voice and Free Expression" [61]
	We build specific systems to address each type of harmful content — from incitement of violence to child exploitation to other harms like intellectual property violations — about 20 categories in total. **We judge ourselves** by the prevalence of harmful content and what percent we find proactively before anyone reports it to us. For example, our AI systems identify 99% of the terrorist content we take down before anyone even sees it. This is a massive investment. We now have over 35,000 people working on security, and our security budget today is greater than the entire revenue of our company at the time of our IPO earlier this decade. **All of this work is about enforcing our existing policies**, not broadening our definition of what is dangerous. If we do this well, we should be able to stop a lot of harm while fighting back against putting additional restrictions on speech (p 5)
	Or take hate speech, which we define as someone directly attacking a person or group based on a characteristic like race, gender or religion. **We take down content that could lead to real world violence.** In countries at risk of conflict, that includes anything that could lead to imminent violence or genocide. And we know from history that dehumanizing people is the first step towards inciting violence. If you say immigrants are vermin, or all Muslims are terrorists — that makes others feel they can escalate and attack that group without consequences. So we don't allow that. I take this incredibly seriously, and we work hard to get this off our platform (p 8)
	As long as our governments respect people's right to express themselves, as long as our platforms **live up to their responsibilities to support** expression and prevent harm, and as long as we all commit to being open and making space for more perspectives, I think we'll make progress (p 10-11)
	Zuckerberg, Mark. 2020. Munich Security Conference [62]
	Um, in the **absence of that kind of regulation, we will continue doing our best**. We're gonna build up the muscle to do it, um, to- to basically find stuff as practically as possible (21:27-22:14, p 9).
	European Commission / Directorate-General Connect. 2019. "Leaked Document on Digital Services Act"[63]
	The lack of legal clarity also entails a **regulatory disincentive for platforms and other intermediaries to act proactively to tackle illegal content**, as weil as to **adequately address harmful content online**, especially when combined with the issue of fragmentation of rules addressed above. As a consequence, many digital services avoid taking on **more responsibility** in tackling illegal content, for fear of becoming liable for content they intermediate. This leads to an environment in which especially small and medium-sized platforms face a regulatory risk which is unhelpful in the fight against online harms in the broad sense. At the same time, when companies do take measures against potentially illegal content, they have limited legal incentives for taking appropriate measures to protect legal content (p 2)
	Finally, a binding "Good Samaritan provision" would encourage and **incentivise proactive measures**, by clarifying the lack of liability as a result of Such measures, on the basis of the notions already included in the Illegal Content Communication (p 5)

[60] Zuckerberg, Mark. 2020. "Financial Times editorial: Big Tech Needs More Regulation". Zuckerberg Transcripts. 1097 https://dc.uwm.edu/zuckerberg_files_transcripts
[61] Zuckerberg, Mark. 2019. "Standing For Voice and Free Expression" Zuckerberg Transcripts. 1022. https://dc.uwm.edu/zuckerberg_files_transcripts/1022
[62] Zuckerberg, Mark. 2020. „Mark Zuckerberg at Munich Security Conference". Zuckerberg Transcripts. 1091 https://dc.uwm.edu/zuckerberg_files_transcripts
[63] European Commission / Directorate-General Connect. 2019. "Leaked Document on Digital Services Act". Retrieved via "Netzpolitik.org", 16.7.2019. https://netzpolitik.org/2019/leaked-document-eu-commission-mulls-new-law-to-regulate-online-platforms/

	European Commission. 2020 "Illegal content on online platforms". Communication on Website[64]
	Online platforms need to **be more responsible in content governance**. The recommendation proposes a **common approach to quickly and proactively detect, remove and prevent the reappearance of content**
	European Commission. 2016. "The EU Code of conduct on countering illegal hate speech online"[65]
	The IT Companies to have in place **clear and effective processes to review notifications** regarding illegal hate speech on their services so they can **remove or disable access to such content**. The IT companies to have in place Rules or Community Guidelines clarifying that they prohibit the promotion of incitement to violence and hateful conduct (p 2)

Less Corp Gover nance	**OHCHR. 2018. Report on Content Moderation. A/HRC/38/35[66]**
	In the light of legitimate State concerns such as privacy and national security, the appeal of regulation is understandable. However, such rules involve **risks to freedom of expression, putting significant pressure on companies such that they may remove lawful content in a broad effort to avoid liability**. **They also involve the delegation of regulatory functions to private actors that lack basic tools of accountability**. **Demands for quick, automatic removals risk new forms of prior restraint** that already threaten creative endeavours in the context of copyright. **Complex questions of fact and law should generally be adjudicated by public institutions**, not private actors whose current processes may be inconsistent with due process standards and whose motives are principally economic (p 7)
	Company policies on hate, harassment and abuse also do not clearly indicate what constitutes an offence. Twitter's prohibition of "behavior that incites fear about a protected group" and Facebook's distinction between "direct attacks" on protected characteristics and merely "distasteful or offensive content" are subjective and unstable bases for content moderation (p 10)
	The **vagueness of hate speech and harassment policies has triggered complaints of inconsistent policy enforcement that penalizes minorities while reinforcing the status of dominant or powerful groups**. Users and civil society report violence and abuse against women, including physical threats, misogynist comments, the posting of non-consensual or fake intimate images and doxing; threats of harm against the politically disenfranchised, minority races and castes and ethnic groups suffering from violent persecution; and abuse directed at refugees, migrants and asylum seekers (p 10)
	However, as the leading review of Internet transparency concludes, companies disclose "the least amount of information about how private rules and mechanisms for self- and co-regulation are formulated and carried out". In particular, **disclosure concerning actions taken pursuant to private removal requests under terms of service is "incredibly low". Content standards are drafted in broad terms, leaving room for platform discretion that companies do not sufficiently illuminate**. (p 14)
	Company rules routinely **lack the clarity and specificity** that would enable users to predict with reasonable certainty what content places them on the wrong side of the line. This is particularly evident in the context of "extremism" and **hate speech**, **areas of restriction easily susceptible to excessive removals in the absence of rigorous human evaluation of context**. (p 15)

[64] European Commission. 2020. "Illegal content on online platforms. Policy". ec.europa.eu, accessed on March, 23 2020. https://ec.europa.eu/digital-single-market/en/illegal-content-online-platforms

[65] European Commission. 2016. "The EU Code of conduct on countering illegal hate speech online". May 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

[66] United Nations General Assembly / Human Rights Council. 2018. "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression". A/HRC/38/35. 6 April 2018.

	OHCHR. 2018. Report on AI. A/73/348 [67]
	Support and pressure for increasing the role of AI come from both the private and public sectors. Companies claim that the volume of illegal, inappropriate and harmful content online far exceeds the capabilities of human moderation and argue that AI is one tool that can assist in better tackling this challenge. According to some platforms, AI is not only more efficient in identifying inappropriate (according to their rules) and illegal content for removal (usually by a human moderator) but also has a higher accuracy rate than human decision-making. States, meanwhile, are pressing for efficient, speedy automated moderation across a range of separate challenges, from child sexual abuse and terrorist content, where AI is already extensively deployed, to copyright and the removal of '"extremist" and "hateful" content.(p8)
	OHCHR. 2019. Report on hate speech and regulatory solutions. A/74/486[68]
	"The pressure is for automated tools that would serve as a form of **pre-publication censorship**" (p14)
	"Because such filters are notoriously unable to address the kind of natural language that typically constitutes hateful content, they can cause significant disproportionate outcomes. Furthermore, there is research suggesting that such filters disproportionately harm historically underrepresented communities" (p14)
	EDRi. 2019. "A privately managed public space?"[69]
	"In its letter to France criticising the draft legislation on hateful content, the European Commission itself acknowledged that simply pushing companies to **remove excessive amounts of content is undesirable** and that the use of automatic filters is ineffective in front of such a complex issue" (p1)
	EDRi. 2019. "Interoperability: A way to escape toxic online environments"[70]
	"However, **obliging the platforms to remove contents is not going to solve the problems of online hate speech**, violence, or polarisation of our societies. Rather than fiddling around trying to treat the **symptoms, the focus should be on addressing the underlying societal problems**" (p 1)
	EDRi. 2019 Leaked Commission document. "More responsibility to online platforms– but at what cost?"[71]
	"At the same time, the note proposes that harmful content should best be dealt with through voluntary codes of conduct, which shifts the censorship burden to the platform companies. However, companies' **terms of service are often a convenient way of removing legal content** as they are **vague and redress mechanisms are often ineffective**" (p 2).
	Zuckerberg, Mark. 2019. "Standing For Voice and Free Expression" [72]
	"And while I worry about an erosion of truth, I don't think most people want to live in a world where you can only post things that **tech companies judge to be 100% true**" (p 7)
	Zuckerberg, Mark. 2020. Munich Security Conference [73]

[67] United Nations General Assembly. 2018. Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. A/73/348. 29 August 2018

[68] United Nations General Assembly. 2019. "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression". A/74/486. 9 October 2019.

[69] Järvinen, Heini. 2019. „A privaetly managed public space?" EDRi.org. November 20th, 2019. https://edri.org/online-content-moderation-privately-managed-public-space/

[70] Berthélémy, Chloé. 2019. "Interoperability: A way to escape toxic online environments". EDRi.org. December 4th 2019. https://edri.org/interoperability-way-to-escape-toxic-online-environments/

[71] EDRi. 2019. "More responsibility to online platforms- but at what cost?" EDRi.org. June 19th 2019. https://edri.org/more-responsibility-to-online-platforms-but-at-what-cost/

[72] Zuckerberg, Mark. 2019. "Standing For Voice and Free Expression" Zuckerberg Transcripts. 1022. https://dc.uwm.edu/zuckerberg_files_transcripts/1022

[73] Zuckerberg, Mark. 2020. „Mark Zuckerberg at Munich Security Conference". Zuckerberg Transcripts. 1091 https://dc.uwm.edu/zuckerberg_files_transcripts

	"Um, but at some level, I- I do think that **we don't want private companies making so many decisions** about how to balance social equities without a more democratic process" (p 9, 21:27)
	European Commission / Directorate-General Connect. 2019. "Leaked Document on Digital Services Act"[74]
	Besides the costly, slow and potentially Contradictory oversight exercised by different sectoral regulators, one consequence is that **many public interest decisions that should be taken by independent public authorities are now delegated to online platforms**, **making them de-facto regulators without adequate and necessary oversight, even in areas where fundamental rights are at stake** (p 3)

[74] European Commission / Directorate-General Connect. 2019. "Leaked Document on Digital Services Act". Retrieved via "Netzpolitik.org", 16.7.2019. https://netzpolitik.org/2019/leaked-document-eu-commission-mulls-new-law-to-regulate-online-platforms/