Queer-phobic ICTs:

- A critical discourse analysis of the legal debates on platform governance in regard to anti queer hate speech online in the EU and German context -

Author: Finn Lennard Lenny Münch;

Programme: Public Governance across Borders

University of Twente, Enschede, Netherlands

Date of Submission: 28.06.2023

Date of Presentation: 29.06.2023

Wordcount: 11868

1st Supervisor: Dr. Azadeh Akbari Kharazi

2nd Supervisor: Dr. Guus Meershoek

Abstract

This paper investigates the negative implications of information and communication technologies (ICTs) on the queer community, focusing specifically on the issue of hate speech on social media platforms. Through a critical discourse analysis of relevant policy documents, this research examines the discourses surrounding the regulation of hate speech against queer people online, including aspects of platform governance and content moderation. The analysis draws upon legislative texts such as the Code of Conduct, the Digital Services Act and the NetzDG from the EU and Germany, as well as hate speech policies of major social media platforms Twitter, Facebook and YouTube. By examining these sources, the study reveals significant discrepancies between stated policy objectives and the real state of hate speech regulation. This thesis discusses the underlying value conflicts and power dynamics that shape the legal discourse on hate speech against queer people. Further, the findings highlight governance gaps, where the EU delegates responsibilities and authority to private intermediaries, resulting in an inadequate protection of the queer community. The thesis argues that the EU has a democratic duty to sufficiently address these issues. The research contributes to closing the research gap on anti queer discrimination online and points to policy recommendations.

1. Introduction

1.1 Research background

Information and communication technologies (ICTs) are never neutral. This thesis will discuss and reflect on how ICTs may have negative implications for the LGBTQIA+ community. More specifically, it attempts to determine how social media platforms are regulated or regulate themselves along the lines of EU and German law in regard to condemning, enabling and fighting hate speech against queer people.

New technologies such as the emergence of big social networks over the last decades have opened up a range of opportunities for the LGBTQIA+ community in terms of connecting with other queer people, gaining access to queer content and discussing with others to better grasp one's own sexual identity (Sybert, 2022). On the other hand, the digital world comes with numerous risks, dangers and areas lacking protection for queer people. There are countless examples of ICTs such as social media and other applications being used to expose, silence and persecute members of the queer community. One extreme example is the persecution of gay men in countries like Egypt, where authorities bait queer individuals using applications such as Grindr luring them into traps in order to arrest and punish them (Shihab-Eldin, 2023). In the EU, queer people on online platforms are regularly confronted with hateful comments, messages and content (Alkiviadou 2019). National governments as well as the EU are struggling to come up with adequate solutions to govern platforms to avoid or decrease hate speech against the queer community. The respective laws and intended policy changes face a lot of scrutiny by supporters and critics of more rigorous legislation alike.

The topic of anti queer hate speech is extremely relevant due to its real life implications. Scientific studies have repeatedly shown that cyber violence in the form of hate speech is directly connected to offline violence such as violent assaults, hate crimes and terrorist attacks (Gillespie, 2020). It is therefore essential to understand the problem of hate speech and especially the complexities and effectiveness of its counter measures and policies. Research in the field of platform governance and hate speech regulation is extremely contested and contains a variety of approaches that will be elaborated in the theoretical framework. The number of articles on forms of online discrimination against the queer community specifically and suitable measures to combat them are still very limited. Further, negative implications for the queer community by a lack of hate speech regulation or new tools such as automated content moderation are rarely discussed or analyzed. This thesis aims at pointing out the existing research gap in the field and contributing to closing it by analyzing and explaining potential lacks of sufficient protection of queer communities and individuals from hate speech by platforms and governments. As conceptualizations of hate speech in policy papers often disregard hateful language against queer people and its implications, this thesis can further contribute to a more extensive awareness about online discrimination (Alkiviadou 2019).

1.2 Research problem

Based on the identified research gap, the following main research question has been chosen:

"How is hate speech against the queer community online discussed, regulated and repressed by governments and social media platforms?"

This explanatory RQ promises to bring to light trends and changes in hate speech discourse and specifically anti queer hate speech regulation. Further, it will provide insights on the policy objectives of the EU, Germany and the platforms. Four sub questions have been developed to organize all of the interconnected concepts and issues at hand and to facilitate the answering of the main RQ.

"How is hate speech defined and addressed in legal documents and platform policies concerning queer people?"

This first SQ covers the discourses of hate speech regulations against queer people. It was chosen in order to understand the way hate speech is conceptualized in policies and with what level of urgency it is met. Further, answering this question can determine how much which emphasis anti queer hate speech specifically is treated with. Understanding and analyzing these discourses can illustrate what relevance is given to hate speech issues and what groups are usually named in its definitions. Lastly, the determination of the contexts the concept is mentioned in as well as the connections to other relevant concepts build the foundation for understanding the further SQs.

"What are the values, rights and freedoms typically upheld and discussed in relation to hate speech regulation?"

This next SQ is aimed at putting the discourses into a value perspective and uncovering the underlying disputes between core values of democratic societies such as freedom of speech and freedom from discrimination. It was designed to better grasp the roots and causes for contemporary hate speech policies and can also provide insights on the level of urgency of the topic of hate speech.

"How is automated content moderation discussed in EU and platform policies?"

This SQ covers an important aspect of content moderation and the fight against hate speech that is especially crucial to the protection of queer people online as mistakes in automated content moderation often negatively impact the queer community and infringe on their ability to express themselves freely. The way and manner this topic is discussed and emphasized in policy documents as well as a potential lack of discussion, can reveal more relevant insights that can help answer the main RQ.

"How can shortcomings in hate speech policies by the EU and the platforms be explained?"

The last SQ is meant to determine the reasons for shortcomings and governance gaps in the field of platform governance and hate speech regulation. It enables shining light on the discrepancies between the platforms

and EU's pledges and the current state of hate speech online. Based on an extensive theoretical framework and the discussions left out of the analyzed documents, reasons and explanations for lacks of effective governance in the field will be determined. The analyzed documents will therefore be connected to the findings of the conceptual framework. This question mirrors the explanatory nature of this thesis and aims at uncovering underlying causes and trends.

1.3 Research Approach

This paper will first provide a very extensive framework of the theories and concepts in question as well as the policies and legal measures by governments and companies. This is necessary in order to put the findings of the analysis into context and enable a more profound discussion. In order to adequately answer the RQ and SQs, a critical discourse analysis of the relevant policy document will be conducted. This qualitative research design is best suited to evaluate and explain current legal practices and reasons for shortcomings in governance. The policies used for this analysis represent the relevant legal framework in hate speech regulations as well as platform governance and content moderation legislation of the EU. Special emphasis is put on the EU's digital services act, which will be put in the context of its scientific reception (Turillazzi et al., 2022). Next to EU legislation, the German NetzDG will be used due to its comparably ambitious approach to combat hate speech. Further, the platform's self-set rules and standards will be analyzed. The specification on these documents promises to deliver the most relevant insights on discourse This paper thoroughly inspects previous research on platform governance and different counter measures to hate speech and their effectiveness as well as evaluations of the existing regulatory framework. On the basis of that, the relevant policies will be analyzed and put into the context of the current research insights in the fields of content moderation, hate speech and platform governance and deconstructed in terms of their underlying values and power relations. In the end, this paper attempts to deliver valuable insights on anti queer hate speech regulation and point to further research necessities and reforms.

2. Theoretical framework

2.1 Hate speech

An important challenge in fighting anti queer hate speech online are the vastly differing definitions throughout various policy papers, legislation and self regulating documents (Alkiviadou, 2019). Most of the definitions in EU policies don't feature queer people as victimized communities and focus on xenophobic and racist remarks. According to the UN, hate speech refers to "expressions that advocate incitement to harm based upon targets being identified with a certain social or demographic group" (Gagliardone et al., 2015, p. 7). According to the Minister Council of the EU, hate speech can be any form of expression that spreads, incites, promotes or justifies forms of hatred based on intolerance (Alkiviadou, 2019). In a framework decision to combat racism and xenophobia, the EU defined it as "public incitement to violence or hatred on the basis of certain characteristics, including race, color, religion, descent and national or ethnic origin" (Gorenc, 2022, p. 414). Essentially, hate speech targets individuals due to them belonging to a specific

marginalized and protected group. Other common features in current definitions of hate speech made by governments or organizations include intended harm or incitement of discrimination (Griffin, 2020). What makes debates about the clear definition of the phenomenon so crucial, is the fact that it can also be used to silence critical opposition or to support smear campaigns in elections (Gagliardone et al., 2015).

There are two groups addressed by hate speech; the targeted individuals and their respective social groups that hate speech aims to intimidate as well as the like minded spectators. Authors of hateful posts on the internet want to show like minded individuals that they are not alone with their hateful views and ideologies (Gagliardone et al., 2015). Hate speech is characterized by creating and reproducing tensions consequently leading to an *us versus them* rhetoric. It occurs in a variety of circumstances depending on means used, speaker, audience, social context and many other factors (Gagliardone et al., 2015). It can be categorized into three levels. The most severe form of hate speech is unlawful content such as incitement of violence that can be criminally prosecuted as it goes against national or international law. However, most hate speech posts belong to the other two categories. The second one is expression that can not be criminally prosecuted yet is eligible for civil restrictions such as removal or banning. Lastly, there is an enormous amount of hate speech not punishable by national or international law but still spreads hateful rhetoric and concern (Gagliardone et al., 2015). The cause for the difficulty to pinpoint one clear definition of hate speech is that it is not used for describing an objective truth but instead is mirroring normative social standards that can change and evolve over time (Griffin, 2020). As aforementioned, some disenfranchised communities were not or are not even featured in definitions proposed by influential institutions (Alkiviadou 2019). It is always a question of social norms and values that shape how scholars, policy makers or companies conceptualize the phenomena of hate speech. Some authors would categorize it as a form of cyber violence (Gorenc, 2022). Finally, all existing definitions are subjective results shaped by a process of social norms at the time and pressure from different groups and institutions (Gillespie, 2020).

2.2 Platform governance

Today's social media landscape is shaped by giant tech monopolies controlling the large part of the market. (Flew at al., 2019). There has been extensive criticism towards these platforms for not efficiently combating hate speech partly as existing media regulations imposed by governments over the last decades have been widely considered inadequate to deal with these issues (Flew at al., 2019). Scholars also pointed out how some features on social media platforms encourage and promote hate speech such as their algorithmic recommendations (Griffin, 2020). Through algorithms, the average user experience is characterized by individuals finding themselves amongst users and organizations that resonate the same beliefs leading to closed off bubbles (Gorenc, 2022). This can lead to users feeling confirmed in their ideologies and becoming desensitized for hateful and extremist content and thought (Gorenc, 2022). Within their niches, for example homophobic views can become reinforced and individuals can even be led to commit acts of violence (Laub, 2019). The clear connection between hate speech online and hate crimes offline has been shown repeatedly as well as social media's ability to catalyze these crimes (Laub, 2019). Ultimately, platforms are businesses that profit off of people finding their ideological echo chambers online. They benefit from controversial content as it increases the views and therefore the worth of advertisement space (Laub, 2019). Evidently, the business model of platforms is to maximize views and provoke attention (O'Regan, 2018). Features such as

autoplay add on that and can deepen extremist views by exposing users to successively worse content (Laub, 2019).

2.2.1 Accountability and liability of platforms

Regardless of the criticism, regulation of online spaces poses big challenges due to their decentralized character, unclear and conflicting jurisdictions as well as the content creation happening bottom up (Flew at al., 2019). Further, hate speech often happens anonymously and without clear jurisdictional responsibility (Gagliardone et al., 2015). Different government levels are involved in online regulation attempting to hold big platforms accountable. This is a governance challenge of itself as most existing media regulation policies exempt social media platforms from most responsibility and don't regard them as producers of the displayed content but instead as intermediaries and content distributors (O'Regan, 2018). Consequently, they can not be held accountable the same way as traditional media outlets would. Nevertheless, many scholars and policy makers argue that the process of organizing content is conducted by the platforms and they are therefore at least partly responsible (Caplan & Napoli, 2018). The platforms themselves argue that they do not generate content and mostly depend on their users to flag and control content (Gagliardone et al., 2015).

2.3 A value driven discourse over time

2.3.1 Freedom of expression or freedom from discrimination

The discourse on (anti queer) hate speech embodies an ideological debate about values. On the one hand, there is freedom of speech, which is held up high especially in Western democracies and mostly in policy debates of the US-American context. On the other hand, freedom from discrimination and the responsibility to protect human rights and disenfranchised groups and individuals is prioritized more in the EU context compared to the US-context (Gorenc, 2022). The different political cultures are underlined by policies and discrepancies in relevant legal texts. The European Court of Human Rights explicitly leaves room for the restriction of free speech as opposed to the US-constitution with its categorical promise of the supremacy of free speech (Bleich, 2014). The legal dilemma consists of determining if the damage created by online hate speech is big enough to make it necessary to intervene and infringe on the freedom of speech (Alkiviadou, 2019).

Freedom of speech is a fundamental and absolute right in international law illustrated in article 19 of the UN declaration of human rights in 1948 (Gorenc, 2022). The freedom from hate speech can not explicitly be found in the universal declaration of human rights. There is, however, the right to protection from discrimination that all people are entitled to. Therefore, if expression is discriminatory, it can be restricted (Gagliardone et al., 2015). Infringing on such a crucial right must always be justified by a legitimate goal and embedded in adequate legislation (Gagliardone et al., 2015). An example is Germany banning the distribution of Nazi literature and media as well as the outright denial of the holocaust (Gorenc, 2022). Despite different understandings of freedom of speech, most European countries have some form of anti hate speech regulation in place following their commitments to protecting human rights in light of the rupture of civilization during the second world war and the holocaust (Gorenc, 2022). This, however, is only one of

various turning points in the history and development of anti hate speech and anti discrimination legislation that has always been extraordinarily loaded as it relates to issues of democracy, dignity and liberty (Gagliardone et al., 2015).

Before the second world war, laws that claimed to regulate and combat discrimination and hate speech were for the most part employed to fight oppositions or gather support for reigning governments. During that time, hate speech legislation was seen as an authoritarian tool to silence political opponents (Gorenc, 2022). In the 1890s, Jewish activists in Germany exposed anti semitic hate speech and campaigned for legal reforms to receive legal protection, thereby causing a shift in paradigm. Subsequent hate speech regulation changed to focus on vulnerable groups in society, aiming for their protection. From then on, hate speech laws were used for more than repressive measures against authority critique (Gorenc, 2022). This legal and political shift was accompanied by societal changes and led to different kinds of legislation with the goal of protecting human rights and preventing identity based discrimination.

However, since the second world war there have also been occurrences of hate speech regulation being imposed to suppress opposition such as in South Africa, which attempted to censor voices criticizing its racist Apartheid regime (Gagliardone et al., 2015). Student and emancipation movements in the late 60s led to European countries employing new punishments and restrictions specifically against racist hate speech while the US held up and protected it as free speech, again illustrating the difference in values (Bleich, 2014).

2.3.2 Debating values in the digital age

Since the beginning of the digital realm gaining influence on public discourse, debates around these topics have radically changed. Suddenly, there was a parallel world that was widely accessible, affordable and enabled anonymous communication. This sphere quickly became one of the main spaces for public discourse and demanded regulation (Gorenc, 2022). This was accompanied by an increase in hateful language on big social media platforms contributing heavily to the polarization of societies (Flew et al., 2019). Hateful language can be understood as insulting rhetoric towards social groups and individuals belonging to these groups (Gorenc, 2022). The language used on platforms forms the thought, ideologies and public discourse offline. As there is no neutral use of language, research shows that even individuals not directly targeted but exposed to seeing hate speech, are heavily affected by it. These individuals as well as society as a whole become desensitized and tend to be more likely to adopt hateful ideologies themselves (Gorenc, 2022). In light of these findings, contemporary debates around hate speech regulation demand greater action by governments. An additional component to this debate is the spike in hate crimes and violent acts against different stigmatized and disenfranchised social groups that often constitute a direct result of hostile and hateful rhetoric (Gorenc, 2022). Therefore, many scholars argue for governments setting thorough legal limitations on freedom of speech to protect the foundations of democracy and social solidarity. Others point out, that extensive regulation could promote paternalistic attitudes towards discriminated groups (Alkiviadou, 2019). Similar to the 20th century, restrictive hate speech legislation could again help authoritarian regimes design regulation to silence opposition and activists (O'Regan, 2018). There are also

scholars calling for different approaches to tackle hate speech and claim that an infringement on freedom of speech would ultimately lead to hateful speech shifting to spaces that are even harder to regulate (Gorenc, 2022). Instead, they suggest engaging in discussion and providing counter claims online to fight the issue at its roots rather than fighting its symptoms (Gorenc, 2022). Oftentimes criticism of current legislation also regards only extreme instances of hate speech being eligible for criminal punishment while non-punishable cases fall through the legislative net (Gagliardone et al., 2015). Additionally, educating and sensitizing the public as well as private companies is regarded as necessary to combat hate speech for example by increasing media literacy (Gagliardone et al., 2015). Many scholars emphasize the role of civil society for this process (Gorenc, 2022). Further, some scholars argue for a complementary framework of government regulation and platform policies and point to less severe penalties, meaning that if platforms handle cases of hate speech themselves there is no need for criminal investigation (Alkiviadou, 2019). Others see a lack of effective persecution through law enforcement authorities to be the substantial flaw in anti hate speech efforts (Claussen, 2018). They therefore worry about the moderation and decision making power lying solely in the platform's hands as they could censor content that isn't favorable to their business interests (Claussen, 2018). This is why, most scholars see governments as having the duty to act in that regard as companies and intermediaries often don't sufficiently fulfill their legal obligations to protect human rights (Gagliardone et al., 2015).

2.4 Platform regulation

Next to shining light on legislative processes and regulations, it is worth looking into the standards and norms that shape most public discourse online. These are set by platforms themselves through their terms of service that go beyond national jurisdictions and build upon the company's own definitions of concepts (Gagliardone et al., 2015). Users are simultaneously subject to national law and service agreements, which illustrate the platform's efforts to avoid liability by for example featuring content warnings (Gagliardone et al., 2015). Further, the user agreements are considered to be opaque and partly unintelligible for users while benefiting elites over minorities and activists (Laub, 2019). As content curators and through their terms of service, platforms possess governance structures, which often fail to meet social responsibilities as platforms are not directly accountable for their users (Suzor, 2018). Some scholars therefore call for judging platforms by their degree of upholding rule of law standards. Accordingly, their style of governance would need to be transparent, equally applicable, consensual and stable (Suzor, 2018). So far, most platforms lack these parameters for good governance posing critics to demand standards for assessing the legitimacy of platform governance. One reason for these shortcomings is that the platform's rules do not equally apply to all users (Duffy & Meisner, 2023).

2.5 Content moderation

Platforms have been criticized for a long time for their lack of efficient content moderation (Gillespie, 2020). There is a clear connection between recent surges in hate crimes worldwide and false or hateful content online (Wilson et al., 2021). Currently the big platforms still lack consistent and transparent procedures to organize and govern the content on their networks, putting them under pressure by users, lawmakers and

shifting social norms towards hate speech (Morrow et al., 2022, Wilson et al., 2021). The question of content moderation is complex and connected to logistical issues and considerable costs for the platforms (Gillespie, 2020). A fitting definition of content moderation is the "organized governance of user-generated posts and behavior by information intermediaries and social media platforms, covering user activities of many kinds and encompassing an increasingly wide variety of interventions" (Morrow et al., 2022, p.1365). It can also be characterized as the "screening, evaluation, categorization, approval or removal/hiding of online content according to relevant communications and publishing policies" (Flew at al., 2019, p. 40). It's a process often outsourced by the companies to individuals in non-Western countries that carry the burden of reviewing huge amounts of traumatizing content (Laub, 2019). Another way of moderating content is labeling problematic, false or offensive content. Content labels are click-through barriers that can entail warnings, corrections and context (Morrow et al., 2022). Most of the big platforms already use this kind of moderation to minimize false information and hate speech. The platform's own content algorithms are considered moderation tools too as they shape and curate the content online (Oliva et al., 2021). The companies usually choose the tool that best protects their business interests rather than one for shaping democratic discourses in which minorities are protected (Gillespie, 2020). This tendency to resort to the easiest solutions as well as pressure by lawmakers and the public leads platforms to over-censor content as they do not regard the context in which it was posted in, for example when algorithmically sorting and removing content (Wilson et al., 2021).

2.6 Debating automated content moderation

Machine learning procedures used for automated content moderation work by comparing newly published content to previously deleted content also known as pattern matching (Gillespie, 2020). Furthermore, machine learning is employed to find new occurrences of hate speech. Therefore, automated content moderation still heavily relies on previous decisions made by humans. During the Covid19 pandemic, many platforms shifted their content moderation to automated processes (Gillespie, 2020). However, the trend of automatization has been going on for years (Oliva et al., 2021). The major benefit of automated content moderation is taking the burden off of the human moderators (Gillespie, 2020). Most big platforms outsource this job to people operating from countries in the Global South, that are faced with the most traumatizing online content imaginable and suffer from psychological distress and under poor working conditions (Steiger et al., 2021). Additionally, supporters of a fully automatized content moderation see AI as the only tool able to deal with the increasingly high amount of data and content published on social media platforms (Gillespie, 2020).

On the other hand, there are concerns about negative consequences. Indeed, machine learning can perceive occurrences in discourses differently through categorizing them into patterns. However, giving AI the power to police and sanction debates online can change our entire perception of discourse (Gillespie, 2020). Many scholars share the discomfort about AI making serious decisions such as banning users and consequently deciding about assertions of hateful and acceptable speech (Oliva et al., 2021). Connected to that, there is considerable concern about private companies policing speech and regulating discourse (Wilson et al., 2021). The big platforms are already influential curators and regulators of content and discourse and have immense power at their disposal (Oliva et al., 2021).

An essential underlying problem is that automated content moderation and the classification of hate speech are not mirroring objective truths. There is no true answer for the assertion of hate speech. Instead, all mechanisms and systems for eliminating hate speech and moderating content online are products of debates, varying definitions and determinations of what qualifies as hate speech. It's an ongoing process that needs to be reevaluated constantly. Machine learning will merely realize past instructions not considering changing values, attitudes and debates. Automated processes can not replace this social assertion (Gillespie, 2020). This debate again highlights the conflicting values in the field. The uncritical belief that AI is the allfitting solution to this very nuanced challenge is predominant in some company circles. This could be described as technological determinism in which companies, policy makers and scholars identify technological problems and regard them as solely solvable through technological means (Gillespie, 2020). These technocentric attitudes are often imprecise, disregarding possible negative implications of an automatization (Oliva et al., 2021). In light of this debate, few scholars and policy makers question the indisputable growth at all costs of the big companies. Consequently, debates about splitting up or limiting the size of powerful platforms rarely take place (Gillespie, 2020). But if the necessity to automize is merely a product of the sheer scale of content, antitrust efforts and limits to evergrowing platforms need to be considered (Gillespie, 2020).

Much criticism regards the fact that no contemporary AI is capable of conducting sufficient moderation that does not disproportionately benefit or hurt certain groups (Gillespie, 2020). The groups hurt by errors of statistically working automation are often marginalized minority communities such as the queer community caused by the automated systems struggling to identify sarcastic remarks and subculture specific expressions (Gillespie, 2020). Furthermore, they are essentially blind for context (Oliva et al., 2021). Queer and other disenfranchised communities, who use community specific speech including words and phrases that could be considered offensive in different contexts, are negatively impacted (Gillespie, 2020). Scholars have revealed the specialties in queer speech such as mock impoliteness and reclaiming slurs as well as the pro social function of both phenomena for the community (Oliva et al., 2021).

By employing context blind algorithms to protect vulnerable communities from hate speech, the exact opposite happens and these communities lose their voices online as well as their opportunity to interact and build each other up (Oliva et al., 2021). Ultimately, members of the community have to fear more censorship than white nationalists and other hateful accounts, who consciously avoid buzzwords known to the algorithms (Oliva et al., 2021). Consequentially, hateful content often prevails while queer people become invisible and obstructed from exercising their freedom of speech. All in all, automated content moderation has negative effects on the queer community as well as on many disenfranchised communities (Oliva et al., 2021). Some scholars therefore propose a complementary deployment of both human and automated moderators with the latter handling the "easy" cases while humans decide about less clear instances (Gillespie, 2020). This way the human moderators would be relieved of the burden of reviewing traumatizing content and automated systems could be provided with more contextual data or constant human oversight (Oliva et al., 2021).

2.7 Legal framework

-

Regardless of the posed challenges, the online sphere is not ungovernable (Flew at al., 2019). It is crucial to find policy solutions that work on an international level taking into regard how platforms operate worldwide (Flew at al., 2019). This layered governance challenge is currently regulated by an opaque net of legislation. The following will explore different approaches and legal attempts within the EU and Germany.

2.7.1 EU

Most of the legal framework on an EU level is mainly focused on xenophobic and racist hate speech online and offline (Alkiviadou, 2019). The EU has committed itself to fighting discrimination and taking appropriate action to reach that goal following its foundational commitment to protect human rights (Gorenc, 2022). In a framework decision of 2008 the EU defines hate speech as "public incitement to violence or hatred on the basis of certain characteristics, including race, colour, religion, descent and national or ethnic origin" (Gorenc, 2022, p.415). Since then, most member states have added sexual orientation and gender identity as well as disability to their list of characteristics (Gorenc, 2022). Regarding the fight against hate speech, the European Court of Human Rights (ECHR) has ruled that legislation infringing on the freedom of speech can only be justified by hindering and counteracting illegal content (Claussen, 2018). Freedom of speech is a uniform right thus its restriction can only happen to defend the integrity of democracy, which can include the protection of the freedom of others as well as preventing crime and unrest (Claussen, 2018). Freedom of expression as well as freedom from discrimination both are protected in the charter of fundamental rights of the EU's Lisbon treaty (Claussen, 2018). Legally, in line with the subsidiarity principle the EU has a competence to act on the matter of hate speech prevention as this problem can not be adequately tackled on a national level (Claussen, 2018).

2.7.2 E-Commerce directive

There have been multiple attempts within the EU to deal with problems resulting from the digital revolution. An early and significant step to achieve more legal certainty was the E-commerce directive of 2000. In that, the EU tried to clear up questions of liability surrounding online platforms (Claussen, 2018). Platforms are categorized into conduit providers that transmit information, caching providers that temporarily store information and hosting providers that permanently store information (Claussen, 2018). It was concluded that only platforms belonging to the latter can generally be held liable. This includes most social networks. However, it remains unclear if these platforms are liable for third party content and if they have the obligation to monitor all content. Further, exceptions were made for platforms if they did not know about illegal content or acted swiftly after discovering it (Claussen, 2018). Regardless of its ineffectiveness and the remaining open questions, the E-commerce directive is an influential predecessor of today's social media platform legislation in the EU.

2.7.3 Code of conduct

The code of conduct is an agreement between the European commission and a number of big social media platforms including Facebook, Youtube and Twitter. They agreed that hate speech has to be removed by the

platforms in the 24 hours after being posted (Alkiviadou, 2019). During the three monitoring periods, progress has been made in regard to what share of hate speech got removed in time. However, it became clear that platforms were more likely to remove content when it was reported by trusted users. Also the platforms knew about when the monitoring periods would take place. Even though the results of the code of conduct were mixed, it was still the most innovative approach of the time to successfully fight and eradicate hate speech (Alkiviadou, 2019). Another flaw is that platforms are only supposed to act if content gets reported so even if the code of conduct was a step in the right direction, it can not cover all areas of this problem.

2.7.4 Digital services act

The DSA went into effect in 2022 and is the successor of the E-commerce directive. Some scholars have titled the DSAs approach to hate speech as a delete first, think later approach (Turillazzi et al., 2022). Critics see the DSA failing to address the roots of hate speech and simply resorting to content removal (Malone, 2022). They fear that extremist users will be able to bypass penalties by avoiding the use of certain words and consequently still be able to spread hateful rhetoric (Malone, 2022). However, the DSA does move away from solely relying on self regulating measures by platforms regarding hate speech (Turillazzi et al., 2022). The DSA does not fundamentally change the liability issue as the platforms are still not directly responsible for content put out by the users up to the point where the content becomes illegal (Turillazzi et al., 2022). However, contrary to former legislation the companies do not get exempt from the responsibility and have to comply with certain obligations such as establishing functioning content moderation schemes and evaluating the risks of their algorithms. Overall, the DSA enables more transparent evaluation of platform algorithms. Nevertheless, it has been criticized for its vague definitions and essential point of critique is the outsourcing of regulatory duties to private entities (Cauffman & Goanta, 2021, Laub, 2019). The continued classification as intermediaries has various benefits for the platforms and has led many critics to accuse the DSA of emphasizing legal certainty for the platforms over user protection (Cauffman & Goanta, 2021). Nevertheless, the DSA can be understood as a milestone in platform governance as it is currently the most ambitious approach to holding non-EU companies accountable (Malone, 2022). It aims at providing stringent and consistent regulations eligible across all of the EU's organs and the member states (Malone, 2022).

2.7.5 Germany and the NetzDG

In light of its history, Germany has a special relationship towards hate speech (Laub, 2019). That is arguably one of the reasons why Germany introduced one of the most ambitious anti hate speech laws in 2017. The NetzDG (german: Netzwerkdurchsetzungsgesetz) entails that content that can be considered as manifestly illegal has to be removed in Germany within 24 hours on for-profit platforms with more than two million German users. This represents a more strict duty to ban content than in the EU's code of conduct (Alkiviadou, 2019). The removed content is saved for an additional ten weeks in case of a legal investigation. Reasons for removal of content are given to both the creating and the reporting user (Griffin, 2020). Failed compliance to these obligations leads to regulatory fines for the platforms (Claussen, 2018). In the German criminal code both insults violating human dignity and incitement of hatred based on nationality, race,

religion or ethnicity are criminalized (Griffin, 2020). Notably , hate speech against queer people is not per se criminalized in the criminal code. Freedom of expression is protected in article 5 of the German Grundgesetz with criteria for infringement on this fundamental right (Claussen, 2018). Following the law's introduction, debates and lawsuits in Germany about its legality and compatibility with EU law were held (Claussen, 2018).

Critics have expressed that suitable legislative answers to the problems posed by hate speech need to include more preventive and systemic approaches instead of merely deleting content (Griffin, 2020). Further criticism regards the platforms classification as intermediaries giving them the duty to compel users into following legal and community rules of conduct (Griffin, 2020). Some scholars view this as a disproportionate infringement of the freedom of speech as it enables overblocking of legal content by outsourcing legal decisions to private companies, who want to avoid being fined (Laub, 2019, Griffin, 2020). Further, existing and useful tools in the platform's infrastructure like banning or filtering, that could be used to achieve the objectives of the NetzDG, are not considered. Proposals for improvement suggest the assessment of content to be put in the hands of independent legal expert authorities instead of private platforms (Claussen, 2018). The NetzDG lacks in finding a systematic approach to social media governance. Its objectives merely cover the scope of deleting content instead of determining the underlying power structures of how platforms are exercising their power over users and how they shape discourses through algorithms and policies (Griffin, 2020). However, the overall reaction to the NetzDG has been mainly positive also due to it being one of the firsts of its kind. It is also worth noting that the law exceeded many expectations as it was originally designed to function as a code of conduct without any legal obligations (Claussen, 2018).

3 Methodology

In order to adequately answer the RQ and SQs, this paper resorts to conducting a Critical Discourse Analysis (henceforth:CDA). The discourses surrounding the topics discussed in the theory section are analyzed and examined in terms of the underlying power relations and checked for possible discrepancies in policy and reality.

3.1 case selection

Based on the research conducted for this paper, a selection of relevant documents has been made that present both the legal framework as well as policies and terms of the big social media platforms. This paper aims at looking specifically at the legal context of the EU as well as the closely connected German context. As the research showed, the EU has been developing ambitious regulations in the last years attempting to tackle the highly complex problems and challenges posed by big tech networks and their implications on society. Germany on the other hand has developed one of the most ambitious anti hate speech laws worldwide and is in the process of developing new legislation regarding public discourse online (Becker et al., 2023). Therefore, looking into relevant legal documents of the EU and its most influential member state Germany promises valuable insights that can help answer the RQ and SQ. The documents include the EU's 2000 E-

Commerce directive as well as its successor the 2022 Digital Services Act (DSA). Additionally, the EU's Code of Conduct and the Commission's respective 2019 Assessment report will be included. Lastly, the German Netzwerkdurchsetzungsgesetz (NetzDG) published in 2017. On the other hand, the most relevant hate speech conduct publications of the platforms Meta(Facebook), Twitter and YouTube have been selected for this analysis. These three companies represent a few of the biggest and most prominent platforms and have been essential in shaping public discourse online for many years. Specifically, Twitter's user agreement as well as their most recent publication on hateful conduct from April 2023 have been selected. For Meta, both its hate speech policy details as well as Facebook's community guidelines. Lastly, YouTubes hate speech policy is also included. These documents were selected in order to gain insights on the platform's official point of view and their declared commitment to fight hate speech.

3.2 method of data collection

The documents used for this analysis have been retrieved from the websites of the German government, the website of EU institutions as well as the websites of the platforms Meta, YouTube and Twitter (Appendix A). They consist of secondary data openly available on websites or in the case of the EU open source databases with all EU legal documents. The legislative texts amount to 140 pages while the platforms documents consist of 46 pages. The policy documents for the analysis were selected as they represent relevant cornerstones of the legal framework of the EU and Germany. Accordingly, they are the shaping elements of the legal realities in the field of platform governance, hate speech regulation and content moderation. In addition to the documents used in the analysis, further documents namely scientific and news articles as well as publications by governments and international organizations were consulted and analyzed in order to deepen the understanding about these discourses. The selected data consists of text documents and is therefore appropriate for a qualitative textual analysis in the form of a CDA.

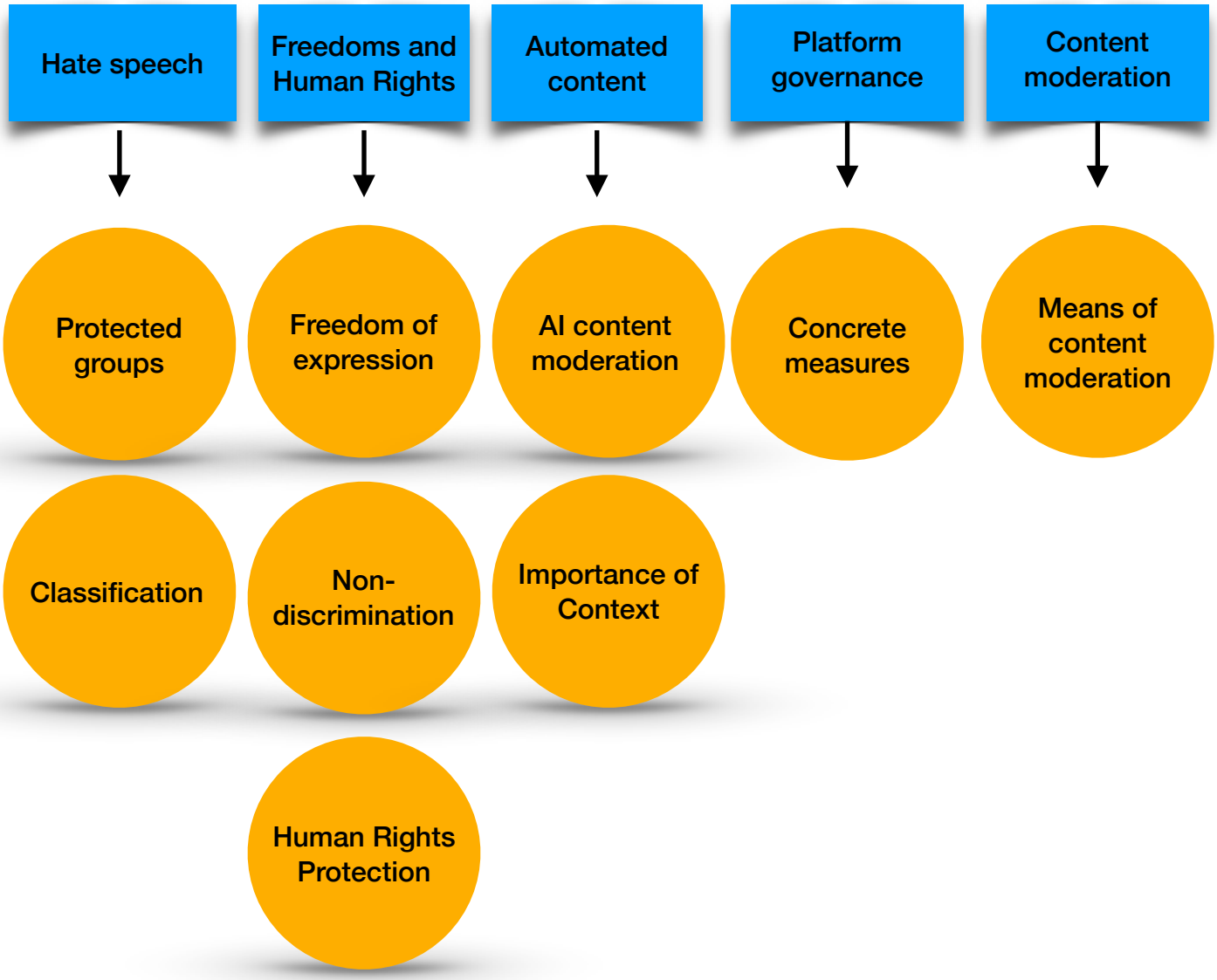
3.3 method of data analysis

This paper aims at determining shortcomings and discrepancies in hate speech and content moderation policies and identifying governance gaps. Further, the discourses in legal documents as well as in the platform's own policies regarding the topic of hate speech against queer people and content moderation are the central elements of this analysis. For these reasons, the conduction of a CDA is the most suitable research design to achieve productive results. Through a CDA, it is possible to uncover systemic biases against marginalized groups by analyzing and examining common themes and discourses in publications, policy papers or scientific literature. CDA analyzes any form of text document in and of itself while also considering its social context as well as the background of the producer of the text and the intended audience. It is deeply rooted in social critique and is determined to analyze the use of language and text in maintaining and reproducing these forms of inequality and oppression (Given, 2008). CDA scholars work on deconstructing discourses related to oppressive circumstances (Leipold & Winkel, 2017). For that, social practices and the societal contexts in which a given discourse takes place become the object of analysis (Jørgensen & Phillips, 2002). A CDA further intends to critically explain a given discourse in regard to the production and consumption of the text (Jørgensen & Phillips, 2002). Lastly, CDA research often points to

possibilities to change the oppressive nature of certain discourse and social practices, consolidating the critical character of this methodological tradition (Given, 2008).

Building on the theoretical framework derived from the research of these issues a coding scheme was developed. It is visualized in Figure 1. It first aims at determining what groups are usually protected and included in relevant hate speech mentions and definitions. Accordingly, it was identified if the queer community was specifically mentioned as a targeted group or not. Further, the classification of the concept of hate speech was examined meaning if it was prioritized and seen as particularly urgent or relevant. Additionally, the context in which hate speech as a concept appeared in the analyzed documents was identified including other concepts it was connected to. Next, the value driven discourse surrounding hate speech explored in the theoretical framework was operationalized in order to find out how certain rights and freedoms are valued in relevant documents. For that, mentions of human rights were identified as well as times where the freedom of expression was explicitly upheld. The same was done for the right to non-discrimination. Additionally, the remarks of these rights were explored in regard as to which context they were placed in. In order to answer the SQ regarding content moderation the counter measures that are proposed or imposed in the documents were inspected. Means of content moderation were examined as well as the context they were mentioned in. Additionally, concrete measures of platform governance were also identified and put into context. Lastly, the discourse on automated content moderation, usage of AI and possible mentions of the importance of context were checked and put into perspective. Based on the analysis discrepancies and governance gaps were identified as well as trends, changes and differences in the respective discourse. The qualitative data analysis tool atlas.ti was used in order to facilitate coding and analyzing the data.

Figure 1: Coding scheme



4. Analysis

4.1 Hate speech

4.1.1 Protected groups

Research suggests that hate speech regulations often do not feature queer people as protected groups. Indeed, they are nowhere to be found in most of the documents used for this analysis. The Code of conduct solely lists racist and xenophobic remarks as eligible for its hate speech definition that go against people on the basis of their race, color, religion, descent or national or ethnic origin. However, it could be argued that queer people are mentioned indirectly. The DSA for example wishes to

"guarantee different public policy objectives such as the safety and trust of the recipients of the service, including consumers, minors and users at particular risk of being subject to hate speech.."(European Parliament and The Council, 2022, p.11)

Queer people could arguably fall under the category of users at particular risk of being subject to hate speech. The most extensive mention of queer people can be found in some of the platform's own policy papers. Twitter's user agreement specifically mentions sexual orientation and gender identity as characteristics protected against hateful conduct. Both can be found in their policy on hateful conduct as well. Meta also names gender identity and sexual orientation as protected characteristics under their hate speech policy and further forbids any expression of homophobic contempt. Lastly, YouTube lists gender identity and expression as well as sexual orientation as two of the attributes that can not be discriminated against on their platform.

Overall, governmental hate speech policies and regulations still mostly reference racism and xenophobia as the only motives for hateful misconduct. The platform's definitions and conceptualizations of marginalized groups targeted by hate speech go considerably further and can be classified as generally more inclusive.

4.1.2 Classification of hate speech

The specific term hate speech can only be found in some of the EU's legislation on platform governance. Most of the time, the term hate speech is embedded into noble pledges for combatting it. EU policy papers feature multiple enthusiastic commitments to the fight against hate speech. It is evidently the stated goal of the EU to avoid the further spread of hateful conduct online.The Code of Conduct identifies hate speech as not only hurting affected and targeted groups and individuals but also the democracies of the member states as a whole. It is further committed to

"ensuring that online platforms do not offer opportunities for illegal online hate speech to spread virally." (European Commission, 2016, p.1).

The DSA states that offline legal rules should also apply online meaning that illegal hate speech is also illegal online. The concept of hate speech as well as the proclaimed efforts to fight it are often mentioned in connection to other forms of illegal content such as terrorism.

Meta conceptualizes hate speech as

"violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation" (Meta, 2023, p.1).

and further indicates how hate speech

"creates an environment of intimidation and exclusion, and in some cases may promote offline violence." (Meta, 2023, p.1).

Hateful conduct further gets classified as a misuse of the platform's services. Concurring with many scholars, the DSA regards hate speech as cyber violence (Gorenc, 2022). In some of the platform's policy papers, this classification is supported as well, for example by YouTubes hate speech policy.

4.2 Human rights and freedoms

The relevant policy papers for the most part cite the protection of human rights as their main objective and even in policies that do not specifically regard hate speech it is pointed out that the fundamental rights of users have to be guaranteed, respected and upheld. An interesting aspect is the comparison of the mentions of freedom of speech or expression versus the freedom from discrimination or the right to non-discrimination. Starting off on a superficial level and looking solely at the DSA, freedom of expression gets mentioned 18 times while the right to non-discrimination and its synonyms get mentioned six times.

Oftentimes the EU's pledges for defending human rights appear in connection with the prioritization of freedom of expression. This particular right is proudly upheld in the policy papers. In the Code of Conduct the EU proclaimed that they together with the social media platforms share both

"a collective responsibility and pride in promoting and facilitating freedom of expression throughout the online world" (European Commision, 2016, p.1).

Especially in the code of conduct, hate speech is rarely mentioned without an articulated commitment to the defense of free speech. The DSA on the other hand features many remarks about both freedom of expression and the right to non-discrimination. This mostly occurs while indicating the platforms to follow the EU's restrictions while guaranteeing fundamental rights at the same time.

On the platforms side, Meta proclaims that it is "committed to voice" and aims at creating
"a place for expression and give people a voice." (Meta, 2023b, p.1).

Later they mention that when limiting free speech they attempt to comply with international human rights standards. Meta themselves classify Facebook as a service that is primarily existent in order for people to freely express themselves. In their hateful conduct policies, Twitter points out that hate speech infringes on people's ability to express themselves creating an uncommon connection between hate speech and freedom of expression.
While most policy papers highlight the need to protect the freedom of speech instead of removing or banning a certain amount of content, this classification shines light on the threatened freedom of expression of the individuals that are being attacked by identity specific hate speech such as anti queer hate speech.

All relevant legislative documents reference the realization and protection of fundamental human rights as their prioritized objective. Over time, the right to non-discrimination has gained attention in legal texts and is now getting recognized and mentioned considerably more frequently than in the older documents. However, with only a few exceptions freedom of expression tends to occupy a more prominent spot in both legislative texts as well as in the platform's own policies. The US-American based platforms further tend to prioritize that right more vocally than the EU legislation. Discussions on hate speech are almost exclusively embedded in mentions of freedom of speech.

4.3 Measures of content moderation and platform governance

Many of the EU's relevant legal documents regard content moderation. The E-commerce directive remains relatively vague on giving platforms concrete instructions for content moderation or similar measures in platform governance. The term content moderation is not mentioned once but there are extensive explanations on liability limitations of the intermediary services. However, it is also pointed out that it is possible for courts to reach injunctions for removing or blocking illegal content. Although the E-commerce directive does not substantially hold platforms directly accountable and liable for the content on their sites, they open the door for future adaptation. The Commission is therefore instructed to provide regular reports about the progress and application of the directive. Further, these reports are supposed to specifically

"analyse the need for proposals concerning the liability of providers of hyperlinks and location tool services, 'notice and take down' procedures and the attribution of liability following the taking down of content." (European Parliament and The Council, 2000, p.15).

This represents the common approach of the time, namely to rely mostly on self regulation and for the most part exempting platforms from being accountable for example in media law. Regarding concrete measures to combat hate speech, suggestions and duties for content moderation come into play. In the DSA the EU compels platforms to resort to

"adapting content moderation processes, including the speed and quality of processing notices related to specific types of illegal content and, where appropriate, the expeditious removal of, or the disabling of access to, the content notified, in particular in respect of illegal hate speech or cyber violence, as well as adapting any relevant decision- making processes and dedicated resources for content moderation;" (European Parliament and The Council, 2022, p. 65).

The DSA is filled with regulations and recommendations for better content moderation not limited to the topic of hate speech. It repeatedly suggests the improvement, reform and adaptation of existing content moderation mechanisms, techniques as well as the decision-making processes around it. The DSA further calls on all networks to meet their existing obligations. Next to criticizing the platform's shortcomings, the DSA also commands them to conduct a thorough assessment of the risks of its technologies also in regard to illegal content such as hate speech. Building on that, the platforms are ought to develop adequate mitigation strategies. Further instructions include the duty to swiftly remove illegal content after it was reported. The German NetzDG strictly compels the companies to block content that occurs on their platforms and is considered illegal in regard to German law. Additionally, the NetzDG explains in detail that the platforms have to deliver regular extensive reports on the handling of complaints of users.

In the platform's policies some concrete steps against hate speech can be found as well. Twitter, for example, provides a list of potential enforcement options for content that violates their hateful conduct policies (Twitter, 2023, p.2). This list includes shadowbanning the content, meaning it can not be found through searches but is not blocked completely. Further, it incorporates the restriction of likes, comments and share

options as well as removing content altogether and suspending violating accounts. YouTube declares that it will remove content or issue penalties if creators promote hostility and violence or repeatedly target individuals and groups protected by their hate speech definition. Specifically, they will delete channels if they received three strikes in the course of 90 days meaning that they violated the community standards and hate speech policies of the platform. If a severe misconduct occurs, YouTube aims at deleting a violating channel right away (YouTube, 2023, p.1).

All of the documents vow to different degrees to fight hate speech and other types of illegal content. Some texts such as the E-commerce directive remain extremely vague when it comes to concrete instructions and binding provisions citing the platform's limited liability as a reason. More recent legislation extends the possibilities to hold platforms liable and repeatedly point out their responsibility to act on issues like hate speech. This presents a trend to more liability of the platforms and more detailed guidelines for content moderation. The platforms themselves elaborate on the sanctions and penalties that they are able to impose on users that violate laws or terms of service but don't directly address issues of liability.

4.4 Discourse on automated content moderation

As the discussions about automated content moderation is relatively new, older legislation such as the E-commerce directive do not feature them. The DSA on the other hand mentions automated tools repeatedly. Most of the time in connection with instances in which they could complement other content moderation practices. However, the DSA also points to potential shortcomings and states that when taking steps against illegal content

“providers concerned should, for example, take reasonable measures to ensure that, where automated tools are used to conduct such activities, the relevant technology is sufficiently reliable to limit to the maximum extent possible the rate of errors.” (European Parliament and The Council, 2022, p.7).

Further, it compels platforms to provide justifications for content removal with respect to reversal possibilities of the decision even if the content was removed through automated content moderation. Additionally, the assessment of the Code of conduct includes a brief inventory of the use of technology and automatic detection systems by the platforms and concludes that the big platforms have considerably increased their usage of these kinds of tools (European Commission, 2019, p.6). Nevertheless, a discussion on the benefits and risks of that technology is not featured.

Twitters and Facebook's hateful conduct policies are the only documents featured in this analysis that broadly recognize the importance of context when moderating content. Twitter explains that certain terms that would normally be understood as slurs can present an empowering act of reclaiming these words when used within a community context and admits

“When we review this type of content, it may not be clear whether the context is to abuse an individual on the basis of their protected status, or if it is part of a consensual conversation.”(Twitter, 2023, p.2).

They continue by asking users for help in identifying such cases in order to avoid unnecessary removal or penalties. Facebook also recognizes the fact that some offensive terms that would violate community standards can be used in a self-referential or empowering way by members of the affected community. Facebook states that their policies

“are designed to allow room for these types of speech, but we require people to clearly indicate their intent. If the intention is unclear, we may remove content.” (Meta, 2023, p.1)

With that, they essentially admit that overblocking occurs sometimes and shift the responsibility for justifying and explaining this kind of content partly to the users belonging to the affected communities. Although the platform's relevant hate speech policies broadly cover the question of context in content moderation, they do not discuss what implications automated decision making in content moderation as well as technological content moderation tools may have on disenfranchised communities.

4.5 Results and key takeaways

A key finding is the lack of sexual orientation and gender identity in the definitions of the EU. However, there is a trend of queer people being mentioned more now than in earlier legislation. The platforms on the other hand provide extensive lists of vulnerable groups including queer people in their hate speech policies. Interestingly, what comes to light in this analysis are the immense discrepancies between the ambiguous pledges of improvement by the platforms to reform and by the legislature to regulate on the one hand and the evident shortcomings in content moderation and combating hate speech on the other hand.

Twitter for example is experiencing an “unprecedented rise in hate speech” regardless of their proclaimed commitment to eradicate it and protect its users (Frenkel & Conger, 2022, p.1). The so-called Facebook files brought to light that Facebook exempted millions of accounts from their rules of service and hate speech policies due to their prominence (Hurtz et al., 2021). This means that the rules Facebook gave itself to combat hateful conduct do not apply to all users. Only recently, it became public that Twitter is systematically not flagging or removing content by “Twitter Blue” users and that its algorithms are actively promoting their content (Center for Countering Digital Hate, 2023).

Additional shortcomings become apparent when inspecting the topic of automated content moderation. Admittedly, it has gained attention in legislative texts and especially in the platform's own policies. There is however little to no mention of the risks of its employment. Further, there is a lack of regulation in that field constituting another governance gap. The Facebook files also revealed that the flaws in their automated content moderation tools are well known to the company (Hurtz et al., 2021). There needs to be a concrete set of rules and definitions to regulate these automated tools in order to avoid overblocking of content, which oftentimes affects disenfranchised groups and hinders them in realizing their freedom of speech (Oliva et al., 2021).

The platforms are well aware of these shortcomings as well as of the dangerous implications that (anti queer) hate speech has on individuals, groups and society as a whole. In their policies they provide definitions, risks and counter measures that exceed some of the legislative texts in terms of volume. Further, the Facebook Files revealed that Meta is extremely well informed about the real life effects of hate speech (Hurtz et al., 2021). Still, hateful content can be found across all of the platforms with serious effects. After all, there is plenty of research revealing a clear link between online hate and offline violence (Laub, 2019).
Legislators and companies attempt to shift the responsibility to each other but finally it lies with the legal authorities to impose adequate and consequential restrictions on the power of these platforms and the negative effects that their shortcomings in regard to countering hate speech have on society. From the platform's perspective, it is logical to not necessarily go further with their content moderation measures than they are legally obligated to do. Naturally, their main interest is not the protection of democratic principles, human rights or disenfranchised groups but to maintain and expand their influence and financial revenue (O'Regan, 2018). In order to achieve that, the platforms change and develop their algorithms so that the users stay on their sites as long as possible to create advertisement revenue (Laub, 2019). The threatening effects of their policies on society are only of secondary concern to these companies. The immense challenges in platform governance that were introduced in the theoretical framework showed the considerable difficulties posed for European lawmakers to regulate online tech giants that avoid falling into a clear jurisdiction. However, it is not impossible (Flew at al., 2019).

Regulations such as the DSA constitute the increased efforts made by the EU and others to tackle that issue and to hold platforms more accountable. Although the DSA presents a milestone in platform governance policy, it does feature less strict obligations for the companies than anticipated (Cauffman & Goanta, 2021). This is also due to extensive lobbying by the platforms in Brussels documented by various NGOs (Bank et al, 2021). Consequently, the DSAs main purpose is providing legal certainty for the platforms rather than protecting users rights according to many scholars (Cauffman & Goanta, 2021). This perfectly illustrates how adequate and effective legislation on content moderation also relies on legal improvements in lobby control in the EU. Official legislation also often lacks instructions or suggestions for how to deal with hateful content that is not illegal but still comes with severe harm for the targeted users. This is especially striking as scholars have been proposing a considerable number of measures that would for example increase media literacy or lead to more social empowerment (Gorenc, 2022). Even the DSA, which is arguably the most progressive and ambitious legislation in this analysis, lacks substantial measures to include civil society (Cauffman & Goanta, 2021). This aligns with the fact that the companies in question were significantly more involved in the process of developing the act than civil society actors such as NGOs (Cauffman & Goanta, 2021, Bank et al, 2021).

Achieving adequate legislation that enables hate speech reduction, also requires a reconsideration of the continued classification of platforms as intermediaries as this is widely regarded as problematic from a rule of law standards perspective (Cauffman & Goanta, 2021). Nevertheless, in the analyzed documents the EU regularly subscribes to that classification, leaving the practical implementation of anti hate speech measures to the platforms and thereby leaving them with a wide scope of action. In essence, fundamentally important regulatory questions about hate speech regulation and content moderation are hereby being outsourced.

Rather than discussing tougher standards or breaking up the power of the big platforms, many of the legal documents frame taking action on regulatory issues like hate speech as the platform's responsibility.
The reasons for the identified governance gaps can not solely be found in the boundary spanning nature, unclear jurisdiction and the sheer complexity of the issue. It is important to also consider the interests of the platforms and the underlying power relations. A big takeaway from this discourse is that the legal debates essentially represent an underlying struggle between different sets of values. This is reflected in the emphasized rights and freedoms. Even though the freedom from discrimination has gained more attention and prominence in the relevant legal texts, it is still mentioned less than the freedom of speech. Especially, the US-based social media platforms emphasize their commitment to free speech (Meta, 2023). TheUS-EU divide in prioritized freedoms is clearly distinguishable. However, the questions of values do not stop there.
The very core of the issue of content moderation, platform governance and hate speech regards democratic values such as legitimacy of the tech giants power. Big platform companies such as the ones used in this analysis nowadays hold more power than entire countries. The impact they have on shaping public discourse and therefore also the political realm can hardly be overstated. This overarching power of the big platforms is rarely questioned and there seems to be little legal initiative to break up that power.

Arising from the EU's principle of subsidiarity, it can take action in matters where member states can merely come up with less effective measures (Claussen, 2018). The nature of hate speech is boundary spanning and multilayered. Therefore, the EU must act on its responsibility to protect the rights and freedoms of its citizens online. With countries such as Germany that have developed their own legislation on the topic of hate speech, effective EU regulations could work in a complementary way. Matters of platform governance and the fight against hate speech are crucial to the safety of citizens as well as the stability of the EU's democratic societies. Consequently, they are best dealt with by the democratically legitimized legislative bodies of EU and national governments rather than by private companies with mainly commercial and power interests. The EU needs to act upon its democratic responsibilities and ensure that the immense power of the platforms is restrained and fundamental rights and freedoms are not threatened online or offline. Therefore, instead of outsourcing the power to regulate public discourse to profit driven entities, the EU should develop more strict regulation or establish neutral review bodies or employ a few of the many mechanisms that scholars have been suggesting for years (Flew at al., 2019).

5. Conclusion

Concluding, hate speech against the queer community online has gained attention in the last decades in governmental and self regulating policy documents. Nevertheless, answering the research question, the analysis has revealed that anti queer hate speech is regulated inadequately. The relevant institutions are not sufficiently fighting hate speech and protecting the queer community and other marginalized groups. One key finding of the analysis is the lack of specific mentions of the queer community in that regard. Solely the social media platforms specifically state gender identity and sexual orientation as protected groups while these groups are merely implied to be included in the EU's legislation.

Recent legislation has led to more transparency about the platform's measures of content moderation.

However, extensive lobbying, too few binding measures and the outsourcing of the authority to act have so far prevented strict and more effective legislation. There are a variety of measures employed to repress hate speech. These are mainly different forms of content moderation. The lawmakers provide obligations for the platforms to delete illegal content and come up with risk assessment and mitigation strategies. However, the implementation of content moderation obligations, the choice of which measures to use to fight hate speech and the employment of algorithms are left to the platforms. This constitutes a gap in governance due to the EU not taking appropriate action in this field in which member states alone can not sufficiently legislate. Regarding automated content moderation, the analysis has revealed a lack of discussion on this topic even though one is urgently needed especially for queer and other disenfranchised communities as they are the ones suffering under its negative implications. Even though the freedom of expression is seemably prioritized over the freedom from discrimination in policy documents, the latter has gained traction. A clear discrepancy between the US-companies and the EU-legislators can be distinguished.

The most significant findings are the discrepancies between the content of the documents and the reality of hate speech in general and against queer people today. The end of the analysis describes the lack of protection of disenfranchised communities online by the platforms despite their evident knowledge of the severity of these issues. It further points out the wrongdoings and shortcomings of EU and governments to properly regulate the companies. When putting the lawmakers' promises and the findings of the analysis in relation to the gathered understandings derived from the theoretical framework the immense divide becomes visible. As mentioned before, lobbying is one factor contributing to the inadequate legislation as well as the EU outsourcing the tasks and implementation procedures to the profit oriented companies. Consequently, this paper argues that the EU does not fulfill its democratic responsibility to act on these issues in order to ensure the protection of marginalized communities and individuals such as the queer community. Instead, the EU continues to allow a small number of companies, whose main interest is maximizing their profit rather than protecting democratic principles, to govern discourses online. These platforms evidently have an interest in generating controversy and attention on their sites as this drives up their advertisement revenue (Laub, 2019, O'Reagan, 2018). Due to a lack of clear, concrete and binding regulations, the power to decide what qualifies as hate speech and what content gets removed often lies in the hands of these private companies. Many scholars have pointed out that predetermined standards developed by private companies on what hate speech is pose a threat to democratic principles and go against the nature of this developing societal assertion (Gillespie, 2020). The value-related variety of hate speech definitions as well as the different levels of urgency this problem is met with will remain a considerable challenge to governance efforts. These issues and assertions, however, have to be tackled by democratically legitimized entities instead of private actors.

In terms of practical implications, future legislation must feature more inclusive and clear hate speech definitions, more binding obligations and explore limiting the power the platforms have over speech, discourse and debate online. Scholars' suggestions include the creation of independent legal expert authorities for content moderation purposes (Claussen, 2018). These questions are deeply connected to good governance and present one of the biggest challenges for public administrators. As mentioned before, discussions and clear effective legislation must to be developed regarding the opportunities and risks of

automated content moderation procedures. This is especially vital to queer communities and individuals as errors disproportionally affect them and lead to an infringement of their freedom of expression.

This paper aimed at providing insights on multiple aspects connected to anti queer hate speech and platform governance policies. Due to the limited amount of words, not all concepts could be explored in detail. More research needs to be conducted to determine the risks and chances of automated content moderation for queer and other disenfranchised communities. The topic queerphobic ICTs and the negative implications of technology on queer communities and individuals specifically is relatively unexplored and new. This paper merely covers a fraction of what queerphobic ICTs and their implications entail. Future research should therefore also investigate different aspects of these issues for example facial recognition systems identifying queer people or the usage of applications and social media for luring in and persecuting queer people in authoritarian regimes.

Additional remarks on the use of language:
In this paper the terms queer and LGBTQIA+ have been used interchangeably and are meant to include all persons that are not heterosexual or cisgender.

References

Alkiviadou, N. (2019). Hate speech on social media networks: towards a regulatory framework?. *Information & Communications Technology Law*, *28*(1), 19-35.

Bank, M., Duffy, F., Leyendecker, V., & Silva, M. (2021). The Lobby Network: Big Tech's Web of Influence in the Eu. Retrieved from https://corporateeurope.org/sites/default/files/2021-08/The%20lobby%20network%20-%20Big%20Tech%27s%20web%20of%20influence%20in%20the%20EU.pdf

Becker, K., Clement, K. & Berlin, V. W. A. (2023).Gesetz gegen digitale Gewalt: Mit Accountsperren gegen Hass im Netz. *tagesschau.de*. https://www.tagesschau.de/inland/innenpolitik/eckpunktepapier-digitale-gewalt-101.html

Bleich, E. (2014). Freedom of expression versus racist hate speech: Explaining differences between high court regulations in the USA and Europe. *Journal of Ethnic and Migration Studies*, *40*(2), 283-300.

Caplan, R., & Napoli, P. M. (2018). Why media companies insist they're not media companies, why they're wrong, and why it matters. *In Medias Res*, 60.

Cauffman, C., & Goanta, C. (2021). A new order: the digital services act and consumer protection. *European Journal of Risk Regulation*, *12*(4), 758-774.

Center for Countering Digital Hate. (2023). *Twitter fails to act on 99% of Twitter Blue accounts tweeting hate — Center for Countering Digital Hate | CCDH*. Center for Countering Digital Hate | CCDH. https://counterhate.com/research/twitter-fails-to-act-on-twitter-blue-accounts-tweeting-hate/

Claussen, V. (2018). Fighting hate speech and fake news. The Network Enforcement Act (NetzDG) in Germany in the context of European legislation. *Media Laws*, *3*(3), 110-136.

Duffy, B. E., & Meisner, C. (2023). Platform governance at the margins: Social media creators' experiences with algorithmic (in) visibility. *Media, Culture & Society*, *45*(2), 285-304.

Flew, T., Martin, F., & Suzor, N. (2019). Internet regulation as media policy: Rethinking the question of digital communication platform governance. *Journal of Digital Media & Policy*, *10*(1), 33-50.

Frenkel, S. & Conger, K. (2022). Hate Speech's Rise on Twitter Under Elon Musk Is Unprecedented, Researchers Find. *The New York Times*. https://www.nytimes.com/2022/12/02/technology/twitter-hate-speech.html

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. Unesco Publishing.

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, *7*(2), 2053951720943234.

Given, L. M. (Ed.). (2008). The Sage encyclopedia of qualitative research methods. Sage publications.

Gorenc, N. (2022). Hate speech or free speech: an ethical dilemma?. *International Review of Sociology*, 1-13.

Griffin, R. (2022). New school speech regulation as a regulatory strategy against hate speech on social media: The case of Germany's NetzDG. *Telecommunications Policy*, *46*(9), 102411.

Hurtz, S., Kreye, A., Mascolo, G. & Obermaier, F. (2021). Facebook Files: Die Erkenntnisse aus den internen Dokumenten. *Süddeutsche.de*. https://www.sueddeutsche.de/kultur/facebook-files-mark-zuckerberg-1.5448206

Jørgensen, M. W., & Phillips, L. J. (2002). *Discourse analysis as theory and method*. Sage.Laub, Z. (2019). Hate speech on social media: Global comparisons. *Council on foreign relations*, *7*.

Leipold, S., & Winkel, G. (2017). Discursive agency:(re−) conceptualizing actors and practices in the analysis of discursive policymaking. *Policy Studies Journal*, *45*(3), 510-534.

Malone, I. (2022) Will the EU's Digital Services Act Reduce Online Extremism? Just Security. https://www.justsecurity.org/81534/will-the-eus-digital-service-act-reduce-online-extremism/

Morrow, G., Swire−Thompson, B., Polny, J. M., Kopec, M., & Wihbey, J. P. (2022). The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology*, *73*(10), 1365-1386.

Oliva, T. D., Antonialli, D. M., & Gomes, A. (2021). Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to LGBTQ voices online. *Sexuality & Culture*, *25*, 700-732.

O'Regan, C. (2018). Hate speech online: an (intractable) contemporary challenge?. *Current Legal Problems*, *71*(1), 403-429.

Shihab-Eldin, B. A. (2023). How Egyptian police hunt LGBT people on dating apps. *BBC News*. https://www.bbc.com/news/world-middle-east-64390817

Siegel, A. A. (2020). Online hate speech. Social media and democracy: The state of the field, prospects for reform, 56-88.

Steiger, M., Bharucha, T. J., Venkatagiri, S., Riedl, M. J., & Lease, M. (2021). The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1-14).

Suzor, N. (2018). Digital constitutionalism: Using the rule of law to evaluate the legitimacy of governance by platforms. *Social Media+ Society*, *4*(3), 2056305118787812.

Sybert, J. (2022). The demise of# NSFW: Contested platform governance and Tumblr's 2018 adult content ban. *new media & society*, *24*(10), 2311-2331.

Turillazzi, A., Casolari, F., Taddeo, M., & Floridi, L. (2022). The Digital Services Act: an analysis of its ethical, legal, and social implications. *Legal, and Social Implications (January 12, 2022)*.

Wilson, R. A., & Land, M. K. (2020). Hate speech on social media: Content moderation in context. *Conn. L. Rev.*, *52*, 1029.

Appendix A: data collection

Bundesministerium der Justiz (2017) Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz- NetzDG), BGBI. I S: 3352

European Commission (2016) Code of Conduct on countering illegal hate speech online

European Commission (2019) Assessment of the Code of Conduct on Hate Speech online 2016-2019, Information Note

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

European Parliament and The Council (2022) Digital Services Act, Regulation (EU) 2022/2065 on a Single Market For Digital Services

Meta (2023) Hate Speech I Transparency Center. Retrieved from https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/

Meta (2023b) Facebook Community Standards I Transparency Center. Retrieved from https://transparency.fb.com/en-gb/policies/community-standards/

Twitter (2023) *Twitter's policy on hateful conduct | Twitter Help*. Retrieved from https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

Twitter (2023b) Twitter User Agreement Retrieved from https://cdn.cms-twdigitalassets.com/content/dam/legal-twitter/site-assets/privacy-policy-new/Privacy-Policy-Terms-of-Service_EN.pdf

YouTube (2023) Hate Speech Policy I YouTube Help I Google. Retrieved from https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=9282436&sjid=16720524244747132480-EU

Appendix B: Coding scheme

Concept	Codes	Key words	Examples
Hate speech	Protected groups	Sexual orientation, gender identity, queer, LGBTQIA+	„protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease.“ (Meta, 2023, p.1)
	Classification	Illegal, criminal, misuse, terrorist, objectives, preventing spread	illegal hate speech or other types of misuse of their services for criminal offences“ (European Parliament and The Council, 2022, p 22)
Content moderation	Means of content moderation	Measures of content moderation	„adapting content moderation processes, including the speed and quality of processing notices related to specific types of illegal content and, where appropriate, the expeditious removal of, or the disabling of access to, the content notified, in particular in respect of illegal hate speech or cyber violence…“ (European Parliament and the Council, 2022, p. 65)
Platform governance	Concrete measures of platform governance	Training, adapting, employing, obligations, penalties	„…with actions geared at ensuring that illegal hate speech online is expeditiously acted upon by online intermediaries and social media platforms“ (European Commission, 2016, p.1)

Concept	Codes	Key words	Examples
Human Rights and freedoms	Freedom of expression upheld	Freedom of expression, defend, responsibility, fundamental, free speech,	„The IT Companies and the European Commission also stress the need to defend the right to freedom of expression,…“ (European Commission, 2016, p.1)
	Non-discrimination upheld	Right to non-discrimination, freedom from discrimination, safety, dignity	When designing, applying and enforcing those restrictions, providers of intermediary services should act in a non-arbitrary and non-discriminatory manner“ (European Parliament and The Council, 2022, p. 12)
	Human rights protection	Protection, fundamental , Charta	„we look to international human rights standards to make these judgments.“ (Meta, 2023, p 1)
Automated content moderation	AI content moderation	AI, technical tools, automized	As part of the efforts to improve the way hate speech content is detected and removed, IT Companies are making an increasing use of technology and automatic detection system.“ (European Commission, 2019, p.6)
	Importance of Context	Context, reclaiming, slurs,	„As „When we review this type of content, it may not be clear whether the context is to abuse an individual on the basis of their protected status, or if it is part of a consensual conversation.“ (Twitter, 2023, p.1)