Vilnius University Faculty of Law

Department of Private Law

Ana Sutidze

Il study year, International and EU Law programme Student

Master's Thesis

Safeguarding the Right to be Free from Discrimination in the Digital Domain

Teisės nebūti diskriminuojamam užtikrinimas skaitmeninėje aplinkoje

Supervisor: Vygantė Milašiūtė

Reviewer: Lekt. dr. Paulius Jurčys

Vilnius

2023

Abstract

This Master Thesis is dedicated to safeguarding the right to be free from discrimination in digital domain, more specifically hate speech being a form of discrimination. Emphasis is placed on the challenges arising from the digital realm within the context of social media platforms and controversies regarding safeguarding fundamental human rights. The study examines the measures taken by different states and entities in addressing emerging threat - hate speech. The research underscores the global significance of these issues, with a specific focus on NetzDG as a pioneer, especially on the most used platform worldwide - Facebook. This master thesis aims to contribute to the discourse on hate speech in the digital domain while offering practical implications for NetzDG's legislative landscape and Facebook regulations.

Keywords: Hate speech; Digital Domain; Over-Removal; Oversight Board.

Contents

Introduction

In recent years, this pervasive growth of digital technologies and landscape, more specifically, the widespread adoption of social media platforms and society's dependence on those platforms, has become an inherent aspect of global society. With more than 3 billion monthly users, the social media giant/the lord of the digital realm - Facebook has transcended its role as a mere platform to be compared to a cyber-sphere that is almost complete with its own legal space governed by the terms of service. This shift once again signifies the unavoided emergence of a distinct legal realm that operates in cooperation with the physical world that we live in, however in this case we cannot deny that cyberspace is uniquely managed by private entities. Within this virtual space, crucial decisions pertaining to freedom of speech and expression are exclusively dictated by these corporate entities.

While the acknowledged purpose of social media is to facilitate human interaction and uphold freedom of speech, the exponential growth of these platforms has led to noteworthy threats, particularly the unrestrained and violent proliferation of hate speech. What was initially conceived as a sanctuary for free expression has now metamorphosed into a space where hate speech thrives, posing substantial risks to the protection of basic human rights. Hence, mitigating these challenges has evolved into a formidable global task.

Content moderation on these platforms is intricately tied to corporate or/and non-public interests, that is why it is impossible not to draw scrutiny and elicit opinions from scholars and human rights organizations alike.

The scientific novelty which is inherent in this study lies in its exploration of the multifaceted challenges and controversies associated with managing hate speech in the digital domain. With the surge in digital growth and society's reliance on all the new platforms, the proliferation of hate speech has prompted a critical examination of whether regulatory frameworks or individualized approaches by social networks prove more effective in addressing this pervasive issue. Considering the effectiveness of regulations, this research paper will compare both legal worlds – having the state be the authority and having a social network as a regulator.

This study aims to address hate speech discrimination as the language of power in the digital realm. It encompasses a comprehensive comparative analysis of online platform regulations, focusing particularly on European countries that have enacted pertinent legislation, shedding light on NetzDG as a first step towards the regulation of hate speech in digital realm on by the state. Moreover, the Research seeks to delineate the responsibility of social media companies in ensuring that their platforms safeguard the fundamental rights of citizens (which, in this case, are users of this particular platform) as outlined in national and international law.

An integral component of this study involves an evaluative lens on the efficacy of self-regulation mechanisms employed by social media companies, Facebook being the primary topic of discussion and Germany’s national regulation NetzDG as one of the main research objects trying to tackle hate speech.

The literature underpinning this study spans theoretical discussions and practical insights from scholars, law practitioners, and theorists. By harnessing data that was derived from social media, the study endeavors to assess the gravity and influence of hate speech. It systematically questions the efficacy of regulatory measures and investigates and analyzes the established practices of Facebook in combatting hate speech.

Employing a diverse array of research methodologies, including quantitative, qualitative, comparative, and analytical methods, this study not only offers a terminological definition of hate speech but also navigates the expansive landscape of cyberspace. Delving into the various perspectives and challenges associated with cyberspace, from its conceptualization to judicial practice, the study also explores the approaches of multiple countries and relevant court practices concerning hate speech on social media. It critically dissects Germany's network regulation and appraises the mechanisms and processes implemented by Facebook in tackling hate speech. Through this comprehensive approach, the study aims to contribute nuanced insights to the discourse surrounding the intersection of technology, constitutional values, and the imperative to combat hate speech in the digital age.

Deciphering Foundational Elements

Discrimination, which is mainly perceived as unjust treatment based on certain characteristics, is a highly relevant aspect and a topic of discussion within the realm of law. It extends its influence all over various social domains, and the significance of this topic is inherently spread across different fields.

Could you imagine that discrimination might occur any step we take? In various forms and environments, such as kindergartens, schools, universities- studying facilities, after that working places, and so on, it is following us through our whole timeline of life, and this life of ours, especially in the 21st century also includes cyberspace.

In this contemporary era, which is dominated by all the digital technologies, our daily lives are automatically shifted to digital space. The digital world did offer immense opportunities for communication, education, entertainment and so on, but it also brought discrimination and other challenges as well. Instances of discrimination can easily be found in cyberspace. Individuals might just either encounter discriminatory treatment or they might themselves become the victims of such behavior, with hate speech emerging as a prevalent form of discrimination. That is why hate speech should get its own recognition.

Defining Hate speech

“Tolerance of hate speech is not tolerance borne by the community at large. Rather, it is a psychic tax imposed on those least able to pay” (Matsuda, 1989). “Real people” encounter “real harms” in the “real world” and not some hypothetical scenarios; that is exactly why Matsuda listed what could be considered as the “real harms,” and this list included racial slurs, pranks, “I am just kidding” type of attitude, where most of them only serve either degrading or straightaway attacking other people. Another interesting approach about hate speech belongs to Delgado (1982), he did emphasize it quite clearly, that every citizen has a right to live a life that is free from others attacking their dignity.

The way that we decide to label a phenomenon reveals our point of view, whether we acknowledge it as a challenge, and what the possible responses could be. For example, this is true for the concept of "hate speech" in the United States, which covers everything from "cross-burnings, racial epithets, insults to religion, bestial and other offensive depictions of vulnerable minorities in leaflets, posters, or on the internet, broad-brush ascriptions of criminality or dangerousness, calls to unite against the members of a hated group, and neo-Nazis marching in American suburbs with swastikas and placards saying "Hitler should have finished the job." The complications of encountering hate speech are toughened by its diversity, presenting possible challenges and difficulties in our reactions to these utterances (Waldron, 2010).

That is why it is important to somehow classify what might be considered as hate speech. Lewis (2007) spoon-feeds the readers with the easiest way to grasp the idea of hate speech by describing it as any speech which causes some offense to others. Obviously, this short “definition” is not enough, or we wouldn’t need any other research regarding this topic, that’s why it’s better to also consider other attempts to break down hate speech into smaller pieces. Slightly earlier Cohen-Almagor (2011) characterized Hate speech as hostile, harmful speech that is motivated by bias and directed towards an individual or group of individuals based on their characters and inherent features. It conveys sentiments that are aggressive, condemning, frightening, discriminating, and/or biased toward certain qualities, which can include either national origin, gender, race, religion, ethnicity, color, or sexual orientation.

According to Gagliardone et al. (2014) There are different views of what can actually constitute "hate speech" because there is not a definition for this phrase that has been agreed-upon. Still, there are two broad patterns that are present in both policy debates and scholarly writing. On one side of the spectrum, we have comprehensive definitions that aim to cover the various ways in which hate speech can appear in all of its complexity. Gagliardone uses pre-mentioned Cohen-Almagor’s definition to better demonstrate the “broad and comprehensive” approach. On the other hand, definitions on the opposite end of the spectrum oppose to this notion of hate speech's expansiveness, stating that this results in being susceptible to be manipulated easily (above-mentioned Lewis’s definition can be a good example of this).

Several authors have been mentioned already, and it seems that almost all of them are simply trying to approach their characterization regarding the type of hate speech either directly or indirectly (Brown, 2015).

All of these make the pattern quite noticeable, even though there are tons of different opinions regarding Hate Speech itself, there’s still no consensus about its definition. So, it might be better if we try to break it down into smaller pieces to understand the fundaments.

The main ambiguity of the legal definition of "hate speech" is primarily based on how the word "hate" is understood. First of all, this is a vague feeling that comes about due to the possibility that the speaker/author is referring to their own hateful characteristics. In this case, their desire can be promoting hatred, making the targets feel disregarded, or even just magnifying already existing hateful sentiments. Waldron (2012) refers to Robert Post's essay in the collection Extreme Speech and Democracy, according to which Post is referring to hatred as an extreme form of dislike. Additionally, Waldron brushes upon the topic whether hatred is more about the main aim or result, or as just the primary motivation of the speech.

When it comes to the term "speech" – there are several things that need to be underlined. Oral/verbal expressions do have the potential to be harmful (and they also tend to be like that in a lot of cases, but we will talk about this later). Nevertheless, the kinds of attacks on aforementioned groups that give rise to initiatives to somehow control and counteract "hate speech" do not only involve verbal attacks but also attacks that might occur through either print, in publications, in displays, or most importantly, on the internet (Waldron, 2012).

Evidently, it is impossible to have a definition of hate speech when it is even more complicated to perceive two ambiguous elements of it hate and speech even separately. In the digital domain, the term "hate speech" refers to speech that is more widely used in particular environment by particular speakers. Speech can be categorized and evaluated to indicate whether or not the speaker/author fits into specific sociopolitical groups. It is also possible to encourage and promote moral and political agendas by challenging others' approaches by using terms like hate speech (Udupa and Pohjonen, 2019).

Digital domain

The shift to an online environment is one major event and change in the landscape. However, the difficulties still exist (or they might have even increased in number). The Internet has not only brought out a deeper understanding and contact between people, but it has also contributed to an increase in polarization, violence, hate speech, and oftentimes anonymous attacks on groups that an individual disagrees with (Herz and Molnar, 2012). Understanding the term "digital domain" itself goes beyond surface-level awareness. As the boundaries between physical and digital realms blur, navigating this complex terrain demands nuanced awareness and a detail-oriented mindset.

Primarily, the term "Cyberspace" itself was introduced by William Gibson, and it was described as a multi-dimensional, artificial, or some kind of virtual reality. What is seen or even heard in this digital realm aren't necessarily physical objects or just representations of physical things. Instead, they contain pure information in terms of their form, characteristics, and existence (Tsesis, 2001; Burnstein, 1996; Uncapher, 1991).

We cannot ignore the actual definition and the essence of the digital domain, which is a concept that includes a number of neologisms that each express a unique aspect (Datasphere, Cybersphere, Infosphere, Cyberspace, and Metaverse are some of the terms that are widely used to describe its various dimensions). This kind of expansive notion incorporates both tangible elements (such as infrastructure and information and communication technologies (ICTs)) and intangible elements (such as stored data, programs, applications, and platforms). Obviously, the borders of this concept are always shifting and switching, and that exact thing leads to them being unclear at the edges most of the time (Taylor, 2022).

To sum this all up, when it comes to figuring out what exactly constitutes to hate speech, it turns out there is a heap of challenges on the table. Even if we attempt to break it down into two parts, hate and speech, there is no universally agreed description. It is a tricky phrase with various interpretations all over the map and a vast array of behaviors, making it almost impossible to pin down a clear-cut definition that satisfies everyone.

But, when we shift to the digital domain, things seem a tad less blurry. Defining hate speech here appears a bit more manageable, offering a somewhat more precise picture. However, the ease of defining hate speech in the digital space is quickly overshadowed by a whole new set of complications. This dual nature simple definition and complex application sheds light on the nuanced and intricate nature of navigating hate speech, especially in the ever-evolving landscape of the digital world.

Challenges and Controversies

“Isn’t that like cyber-bullying?”

No.

Prior to delving into the more nuances of this subject, it is imperative to acknowledge a prevailing misconception that frequently obscures discussions surrounding online interactions. There is a common tendency among individuals to mistake cyber hate speech for cyberbullying, which leads to a cascade of inaccurate perceptions right from the outset. It is paramount to disentangle these two distinct phenomena and establish a clear line between them. This delineation sets the stage for a comprehensive understanding of the challenges and intricacies inherent in both cyber hate speech and cyberbullying, fostering a more measured and analytical examination of these phenomena.

Cyberbullying is a broad phrase that encompasses a number of related concepts, such as online bullying, electronic bullying, and Internet harassment. There are several definitions of cyberbullying in the literature, out of which many of them draw inspiration from understanding traditional bullying. Every one of these terms essentially revolve around a “perpetrator” engaging in violent, hostile, or harmful behavior while utilizing some type of electronic equipment as the main weapon. The complex relationships of the individuals that are engaged in this whole situation, also the predetermined nature of the act, and additionally, the emphasis on repetition over time are the nuances generally (Tokunaga, 2010; Besley, 2008; Patchin & Hinduja, 2006; Smith et al., 2008). Although cyberbullying can at times be hateful, it does not necessarily attack a person because of the fact that they belong to a certain group; however, when it comes to hate speech, it’s totally different (Saleem et al., 2017).

Hate speech and cyberbullying (even though they are quite similar in substance) exhibit different patterns of targeting. Hate speech differs from cyberbullying in certain ways, for example, in the aspect that it focuses on individuals inside specified environments such as schools or workplaces, whereas hate speech focuses on specific groups due to a shared characteristic. Despite these variations, a significant connection appears in people's perceptions, according to which the border between hate speech and cyberbullying once again blurs, and it becomes a reason why there’s a complicated link between the two terms. This interaction could be a result of the aforementined changing dynamics of digital communication, where manifestations of hatred and bullying frequently overlap and generate a diverse landscape that challenges previous categorical distinctions (López C.A and López R.M, 2017).

César Arroyo López and Roberto Moreno López also carried out C.O.N.T.A.C.T. interviews in Spain, according to which it became visible that out of twenty interviewees, six of them associated hate speech (that they have encountered throughout their lives) to cyber-bullying or vice versa. Additionally, they thought that the main basis for both online hate speech and cyberbullying was that those who engage in these behaviors feel free to spread extreme opinions and insults on the internet due to its perceived anonymity (which may make these actions more challenging to carry out offline). That brings us to another challenge- anonymity.

"Anything, or any situation, that makes people feel anonymous, as though no one knows who they are or cares to know, reduces their sense of personal accountability, thereby creating the potential for evil action"- Zimbardo.

One apparent advantage of the Internet as a communication tool is the liberty it provides people to conceal certain elements of their offline identities until they so desire. It is proposed that the anonymity of the Internet fosters an atmosphere that is more open to free speech, encouraging people to express themselves without fear of negative consequences relevant to their skin tone, sexual orientation, or gender identity. This anonymity may, at the same time, create an environment which can be suited to open discussion and a variety of viewpoints (Brown, 2017; Graham, 1999: 143). ----------------------compared to offline

When many online identities (even from notably different platforms) are linked and cross-referenced, the complexity of Internet anonymity becomes obvious. Certain programs and services can link users' online identities to offline reality in exchange for personal details. Law enforcement's abilitie to locate and collect digital evidence raises doubt on the idea of complete anonymity. Even in cases when hate speech is lawful, the decision to express yourself online may be motivated by fears about the potential immediate and serious physical consequences that hate speech conveyed in person could have. Although there is little chance of an immediate physical threat when it comes to post-event identification online (Brown, 2017). ----------------------compared to offline

According to the C.O.N.T.A.C.T. project that we mentioned above, led by César Arroyo López and Roberto Moreno López in Spain (2017), the interviews with numerous young participants revealed that posting anonymous comments online provides a new way for expressing intolerance, rejecting diversity, and endorsing racism without taking the social constraints of offline communication into consideration. A lot of researches showcase that there is no singular profile for individuals engaging in hate or cyberbullying under the cloak of anonymity. While organized groups may aim to propagate hatred online, those behind hateful or discriminatory trolling messages are often not affiliated with openly intolerant ideological movements. Instead, they are users who may not fully comprehend the potential impact of their digital activity and its repercussions in the "real" (offline) world.

It is also worth mentioning that in most of the cases social groups appear to have an opportunity to promote prosocial conduct and stop hate speech by promoting affiliation with lawful/legal social identity groupings (this might just be a subtle effect. In some cases, it can even be unnoticeable, but still) (Seaman, 2008).

Hate speech versus Free speech

Freedom of speech is supposed to be inherently important and doesn't need any further explanation. Speech is essentially a flexible mode of communication that may be applied in a variety of environments, this can include public discourses, academic researche, and personal relationships. It is also better to remember that freedom of speech plays a universal and crucial role in shaping the structure of our society and intellectual discourse. Herz and Molnar (2012) give us several aspects of why free speech is fundamental to human existence: firstly, Critical Self-Awareness (The fundamental right to free speech is what allows people to think independently. Speech systematizes, imposes order, and enables critical reflection by objectifying and articulating thinking); Secondly, meaningful Human Life (Speech is a means for acknowledgment, self-disclosure, and creating shared memories. Restricting communication leads to wea,k shallow relationships that are marked by transparency and ignorance); Thirdly, essential Function in Society (In the world of politics, free speech allows people to critically examine views, create well-informed opinions, and serve as a check/verificator on the legitimacy of the government. Speech is also what guarantees an uninterrupted flow of ideas and cultivates a dynamic civic society); and Lastly, Intellectual Empowerment (To put it simply, free speech is essential for intellectual inquiry because it allows people to critically analyze ideas, challenge preconceived notions, pursue the truth, and develop critical thinking skills).

“Prima facie, freedom of expression can be seen as diametrically opposed to hate speech”, the idea of passing laws against hate speech is frequently seen as a possible violation of the right to free speech and a danger to the goals and values of democracy (Gagliardone, et al., 2014). Each culture has evolved solutions to somehow strike a balance between the rights to free speech and the principles of equality and dignity, which has led to the formation of particular alliances and rifts (Gagliardone, 2019; Hare & Weinstein, 2010; Rosenfeld, 2002). In this context, the case of ÜÇDAĞ seems relatable.

In the case of ÜÇDAĞ v. TURKEY Mr. Üçdağ was an imam at a local mosque who was also public official, he shared two posts on his Facebook account (It’s impossible to avoid Facebook and leave as is, so that will be discussed in one of the following sections). The Court determined that prosecuting the applicant for publishing content about a terrorist organization on his Facebook account did not strike a suitable balance between his freedom of expression and the use of violence, armed resistance or uprising, or hate speech.

Everyone is able to express their emotions freely and without any kind of hindrance as they have the right to free speech (Van, 2017). According to Fish (1994),there’s “no such thing as free (nonideologically constrained) speech“. Besides all these Gelber (2012) noted that regardless, everyone has to stick strictly to the rules of the law and show respect while giving their opinion. Comparatively speaking, hate speech disregards decency and civility. Every human being has something positive to offer. Regardless of one's political or economic background, respect guarantees that everyone feels valued. On the other hand, hate speech disregards respect since it uses derogatory terminology (Bhatia, 2016).

Another interesting idea about free speech and hate speech is given in the article “There’s no such thing as hate speech and it’s a good thing, too,” written by David Boromisza-Habashi (2013), where he notes that Similar to how speech isn't "naturally" hateful, it also isn't "naturally" inclined to be free. Another case that shows the interconnection between hate speech and free speech (in the aspect of what could be considered as an example of hate speech) is Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary case. The European court of Human Rights (ECtHR) acknowledged that some of the comments made in the Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary case were offensive and vulgar, but it emphasized that vulgarity is not unlawful and that "style constitutes part of the communication as the form of expression and is as such protected together with the content of the expression."

In the case of MTE & Index v. Hungary, MTE and Index were required to cover the applicant's expenses, and no non-pecuniary damages were granted. Nevertheless, the ECtHR emphasized that the absence of damages awarded was not pivotal in evaluating the repercussions for the applicants. Instead, the focus was on the potential implications for similar internet portals regarding their liability for third-party comments. Such a judgment could give rise to additional legal disputes where compensation might be granted (Sidlauskiene and Jurkevicius, 2017).

There was a poll conducted by the Rasmussen polling firm, in which respondents were asked whether hate speech should be prohibited. The results showed that 28 percent were in favor of hate speech prohibition, 53 percent were opposed to this idea and the other 19 percent were unclear. There was another question asked by Rasmussen polling firm, which was more concerned about whether it is better to allow free speech without government interference or do people prefer to let the government decide what types of hate speech should be banned. As an answer to this question, only 11% said that it is better to let the government decide, and 74% said it is preferable to allow free speech without restricting it at all. For the majority, the idea of the government deciding on which speeches would be tolerated and which would not be quite unacceptable, even for many who were unconcerned about hate speech limits (Greene, 2012)

David Boromisza-Habashi (2013) also mentions that it would be more appropriate to understand which definition of hate speech is more superior then the other, to avoid an attempt by politicians to give credibility to some certain forms of communication. This leads us to another problem of politicians using hate speech in their favor.

"Weaponizing Words: Hate Speech as a Political Tool"

“The term “hate speech” is used by different types of speakers for various political purposes” Udupa and Pohjonen (2019). Politicians might use this term to either warn the public about potential risks or they just use it to gather support for specific programs. In some of the cases, activists use the term Hate speech as a rally cry and they are trying to change public perception and bring attention to social justice concerns with this emphasis. Another reason for politicians to use this phrase is to back up their legal actions or other initiatives/attempts; for example, they are trying to highlight how important it can be to limit speeches that might be disruptive to the social harmony and balance.

The word "hate speech" is more than a neutral term, it is an effective weapon that individuals wield to sway others perceptions, promote their own narratives, and expand their political goals. Individuals from various backgrounds can strategically very wisely use it to influence more people. People must recognize that just like harsh words and strong/unyielding statements might be a threat to those in power, they can also serve to label certain comments as "hate speech," and it could potentially be used to help those already in charge to maintain their influence and the political support (Udupa and Pohjonen, 2019)

Liability issues based on ECtHR case law

SANCHEZ V. FRANCE

ECtHR had quite an interesting case regarding politicians being associated with free speech/hate speech in the case of SANCHEZ V. FRANCE.

In this case, Sanchez was a politician who got fined in the criminal proceedings due to failing to delete Islamophobic comments that were made by third parties on his publicly accessible Facebook page. Despite knowing about such comments and to make it all worse, this all was happening during the election campaign. The case raised major legal and ethical concerns regarding the responsibility of politicians who use social networking sites for political and electoral reasons. It also brought attention to the problem of hate speech during election seasons, and it highlighted the shared responsibility of all actors involved in combating such content.

The court sharpened the focus on the fact that the extent of accountability that a person can be assigned to is also based on their individual characteristics. They stated that a certain level of notoriety and influence adds weight to both words and actions of and individual. To determine the appropriate response, the court referred to proportionality analysis and based their approach on individual’s level of responsibility (focusing on the fact that a private individual with limited recognition and influence wouldn’t have similar obligations like a politician would). The implications of this judgment for the use of social media by politicians include the need for the shared liability in-between all actors that are involved, the implementation of a minimum degree of either moderation or prior filtering by the host or account holder to identify and remove hateful/unlawful remarks within a reasonable time. In the end, the European Court of Human Rights ruled that the politician's criminal sentence was proportionate.

To sum it all up, this case exemplifies the complex relationship of free expression, political relationships, and the responsibility for avoiding hate speech in the modern digital age.

DELFI AS v. ESTONIA

In the case of DELFI AS v. Estonia, the issue at hand was the company's liability for the defamatory comments posted by third parties. The case raised questions about the balance between freedom of expression and the need to protect individuals from defamatory and unlawful speech disseminated online. Regarding this, the company claimed that, unlike traditional print media, the unique structure of Internet media makes comment editing impossible when it is not even published yet, and that websites just play a critical role in fostering the free flow of ideas and information, particularly online.

The Court examined a number of aspects, such as the fact that the comments were in response to an article on the applicant's commercially arranged news portal. Initially, it criticized the company's failure to immediately remove comments that amounted to hate speech or incitement to violence, as well as to hold the writers accountable. Despite these considerations (in addition to the moderate sanction imposed on the applicant company) the Court held that the domestic court's decision to impose liability was appropriate and purely based on relevant grounds, taking into consideration the respondent State's margin of appreciation. As a result, the measure was not considered to be an excessive strain on the applicant company's right to free expression. Consequently, this case highlighted that web portals could be held liable for neglecting to moderate user-generated content, including hate speech used by its readers.

In this specific instance, Delfi AS did not mandate commentators to disclose their identities, and the Estonian courts could only identify some of the computers from which the relevant comments originated. Therefore, the questionable effectiveness of the means to identify the comment authors and Delfi AS's inadequate efforts to address the claim with the true authors of the comments were crucial factors supporting the decision of the Estonian Supreme Court. Additionally, the ECtHR highlighted that shifting the risk of damages recovery from the injured person to the media company, typically in a better financial position, does not constitute a disproportionate restriction on the company's freedom of expression. However, it's noteworthy that the mere economic advantage of the website operator should not, in itself, serve as the sole justification for its liability (Sidlauskiene and Jurkevicius, 2017).

On the other hand, in some of the cases ECtHR considers it a violation of a news websites' freedom of expression to impose a fine for publishing a reader's offensive comments, for example, Case of M.L. and W.W. v. Germany.

BEIZARAS AND LEVICKAS v. LITHUANIA

The case of BEIZARAS AND LEVICKAS v. LITHUANIA involved two individuals who posted a picture on their social media profile. This picture led to anti-homosexual comments on Facebook, and the applicants reported these comments to the police, but shortly, authorities refused to initiate a pre-trial investigation.

The applicants argued that because of this refusal regarding the hostile remarks that were made on the Facebook page, they were subjected to discrimination on the basis of their sexual orientation. They also argued that the freedom of expression of the commenters and their right to privacy were not sufficiently balanced by the domestic authorities.

Eventually, the Court came to the conclusion that the negative remarks were the result of bias toward the applicants' community. The same discriminating mentality was also found to be a major contributing element in the authorities' inability to carry out their positive obligation of thoroughly examining whether the remarks constituted incitement to violence and hatred. This further confirmed that the authorities were, at the very least, allowing such remarks by downplaying their seriousness. Hence, in the end, the European Court of Human Rights found that Lithuania violated articles 14,8 and 13 of ECHR, which refers to prohibition of discrimination, right to respect private and family life, and right to effective remedy.

Initially, prior case law addressing the effective investigation of hate crimes focused on physical violence or direct verbal assaults that constituted individual threats of physical harm. However, the case of Beizaras and Levickas introduced a shift as the comments involved were differed in nature. In this instance, the perpetrators did not explicitly issue direct threats to the applicant (Kundrak, 2020).

Nevertheless, the internet and cyber hate have the potential to cause harm and silence individuals. Hate speech serves as a precursor to bias-motivated violence and hate crimes. The case of Beizaras and Levickas exemplifies how verbal hate crimes, including online threats and incitement to violence directed at two young men, demonstrate the harmful impact of words (Kundrak, 2020).

Diverse Approaches and Regulatory Spectrum

The United States

Most democracies have laws against hate speech; however, the United States Supreme Court has failed to create a constitutional regulation and definition of hate speech so far. Sharpening the focus on the fact that hate speech has not been explicitly acknowledged as a First Amendment exception. According to the Supreme Court's judgment in Brandenburg v. Ohio in 1969, freedom of expression can be hindered only when it is meant to directly encourage illegal action, and there is substantial likelihood that such activity would have equivalent consequences.

There are several arguments that are widely used to oppose hate speech regulation in the United States. Above all, some consider that if they recognize "incitement to violence" or "incitement to illegal acts" as legal phrases, then it would result in a broad definition of hate speech. Additionally, broad hate speech regulations could easily jeopardize the future of political discourse and they would pose a threat to democracy. Furthermore, the objectives do not justify the methods in every case, even if there is a link between the implementation of hate speech laws and a decrease in hate speech content, it could lead to silencing minority groups (as well as people with different points of view) and these processes once again would result in the collapse of democracy.

Another really paradoxical character of the US approach is that it does not explicitly prohibit Internet intermediaries from regulating the comments or remarks on their platforms all by themselves autonomously (O’Regan, 2018; Heins, 2013-2014).

The United Kingdom

With the Defamation Act of 2013, new defenses for online publishers were introduced, offering protections for honest, true, and public interest publications. Moreover, it provides security to service providers who host material created by users. In the context of this act, a publisher is defined broadly to include anybody who distributes a defamatory remark through primary and secondary means. Secondary publishing can also include organizations like ISPs, such as Google Inc., (when the publisher does not actively exercise editorial control but nonetheless makes defamatory information available to other parties). The 2012 Christopher Anthony Mcgrath v. Professor Richard Dawkins case brought hyperlinks and their possible use in identifying an individual as a secondary publisher of linked content into question. In this case, the court accepted that this is a possibility but highlighted that it depends on the specific circumstances of the case (Ali, 2015).

Referring to the Law Commission of England and Wales (more specifically, report to the House of Commons Select Committee on Home Affairs, Hate Crime: abuse, hate and extremism online), O’Regan (2018; paras 2.4 and 8). highlighted that, though there is not enough data to reach definitive inferences, there are signs that hate speech is becoming more common online in the UK. Under Part 1 of the Malicious Communications Act prosecutions are said to have increased tenfold between 2004 and 2014, according to testimony given in 2016 to the UK House of Commons Select Committee for Home Affairs.

Based on recent developments, according to National Cyber Strategy 2022 the main objective of the new law is to impose higher standards for internet content monitoring. The law attempts to require platforms to protect users from illegal content and to take proactive steps to report and delete harmful content, even if it is legal. Companies now address such information only after user complains, but the proposed regulation requires them to be proactive. Furthermore, this legislation categorizes companies depending on their size and risks connected with their internet presence. Notably, big platforms such as Facebook and Google might be required to remove anything that is deemed harmful, even if it is legal.

The European Union

Article 10 of the European Convention on Human Rights protects freedom of speech, but the degree of protection guaranteed by this article depends on the context. Another factor that should be taken into consideration is legitimate purpose. For example, political debate is often filled with harsh language that can at sometimes qualify as hate speech, but it could still benefit from free speech protections, just because it serves to strengthen democratic processes.

The bloc's twenty-eight countries approach hate speech legislation on social media in different ways, but they all reflect fundamental values. Unlike the United States, which exclusively addresses speech that directly incites violence, the EU considers speech that incites hatred or denies or minimizes genocide and crimes against humanity (Laub, 2019).

Catherine O’Regan(2018) also gives us a detailed list of the regulatory framework of the European Union trying to combat hate speech, listed below.

The e-Commerce Directive is the fundamental legal structure that protects Internet intermediaries from responsibility for the content that was spread across their platforms. But, there are requirements that must be met if we want this immunity to work and apply. Starting from the first one, unlawful content must be removed, or the access should be disabled as soon as it is discovered; and the second one, neutrality must be guaranteed by either passive attitude or abstaining from sponsored or developed content.

In June 2016, major platforms (such as Facebook, YouTube, Twitter, and Microsoft) started their journey with the EU Code of Conduct which aimed at Countering Illegal Content Online. This code was their way of showing their commitment to reviewing and removing information that violates community standards or seems to be illegal within 24 hours. After some time, some other platforms (Google +, Instagram, Snapchat, and Daily Motion) have also joined this Code of Conduct, and they are trying to show their effort by presenting periodic reports which includes detailed data regarding their actions.

The European Commission's Code of Conduct became the main scale for monitoring and evaluating how much social media platforms were complying to the regulations by the end of 2016. During the six-week period, six hundred suspected cases of illegal hate speech on the internet were reported by twelve civil society organizations from nine Member States. In the records that these organizations have kept, it was shown that out of the six hundred notifications, 179 situations resulted in the removal of content, out of which Facebook was responsible for 28.3% of those (Alkiviadou, 2018).

Lastly, in the case of Eva Glawischnig-Piesczek v Facebook Ireland Limited, Eva Glavising-Piszczek (who was at that time the head of the Austrian Green Party) sued Facebook and stated that the social media platform failed to take down abusive and defamatory messages about her. The Court of Justice of the European Union reviewed this and supported her claim in the end. This case is likely to influence the decisions of national courts in EU Member States. Notably, it extends the obligation to delete content to instances where the material is shared on the social network, even globally, with slight modifications. This expansion aims to enhance the protection of users' personal rights by compelling social networks to proactively identify and remove similar instances of illegal content, reducing reliance on individual claims. Moreover, the decision signals an acknowledgment by the Court of the limitations and unfairness associated with social networks restricting access to offensive content based solely on IP addresses from a specific country (Knol Radoja, 2020)

France

The LICRA v. Yahoo Case should also get its well-deserved recognition. Above all, this case gained attention due to its association to Nazi incitement and hate speech. Yahoo was sued for allowing users to share the content with Nazi emblems in France. They did try to claim that French legislation could not be applied due to the fact that the company was based in the United States. However, the French Supreme Court rejected this claim and pointed out that the company's website was what made these Nazi emblems accessible worldwide and the court ordered the Yahoo to stop hosting such websites in order to prevent French citizens from accessing them.

When it comes to the regulations, the French National Assembly passed legislation targeting hate speech in May 2020, including sanctions for internet providers that fail to delete certain data rapidly. Platforms may face fines proportional to a percentage of their annual income if they do not delete information within an hour, according to the bill. There are also predefined punishments for failing to delete "manifestly unlawful" hateful content that incite racism or anti-Semitism. Despite government support, France's National Consultative Commission on Human Rights voted against the decision, citing concerns about the impractical timeframe for platform investigations, the potential removal of legal content to avoid fines, and the belief that platforms should not bear sole responsibility for content analysis without judicial involvement.

Germany

In Germany, social network operators are required to set up a procedure for dealing with complaints about "manifestly unlawful" information under the NetzDG, or Network Enforcement Act, which was implemented on January 1, 2018. This material is about particular charges defined in the German Criminal Code, such as incitement to hatred, insults, and defamation. Furthermore, the law requires semi-annual reporting and requires social network firms operating outside of Germany to appoint a designated representative within the country as well. NetzDG needs greater attention thus the next part is entirely dedicated to it.

NetzDG as a “shield” from hate speech

Governments are progressively addressing Extremism and hateful ideology issues through what is referred to as 'new school speech regulation' (NSSR) (Griffin, 2021; Balkin, 2014)**.**

Van Loo(2020) suggests that the appeal of regulating tech companies lies in their ability to control mass user speech through technological means like filtering. This trend is not exclusive to speech regulation; it signifies a broader pattern of involving 'gatekeeper' companies as regulators. All this brings us to NetzDG.

The Network Enforcement Act (NetzDG) in Germany, informally known as a "hate speech law," is a significant attempt to hold social media companies accountable for addressing online conduct that is unlawful under domestic law. According to Kohl (2022) it is “the first law that formalizes the process for platform takedown obligations”. There are some misconceptions going around regarding NetzDG and it’s important to deny them from the very beginning. First of all, NetzDG does not fall under criminal law. If we do not understand this, we would probably be ignoring or changing the whole idea behind NetzDG (Rather than criminalizing actions, it establishes regulatory measures that should be applied to online platforms and these measures are either content moderation obligations or penalties) (Griffin, 2021). Furthermore, NetzDG does not create some new categories of illegal content, it only applies 22 already-existing prohibitions from the German criminal code in the internet environment (Tworek & Leerssen, 2019).

More specific list Regarding what can be considered illegal is given by O’Regan(2018), who refers back to German criminal code (S 130, 131, 166): 1. Incitement to hatred against a national, racial, religious, or ethnic group, or specific population segments (This includes calls for violent or arbitrary measures against those groups and assaults on human dignity through either insulting, malicious maligning, or defaming segments of the population in a manner that it affects and disturbs public peace) 2. Dissemination of depictions portraying cruel/inhuman violent acts against human beings in a way that it affirms or trivializes actions or violates human dignity 3. Defamation of beliefs or religious or ideological organizations that pose a serious threat to public peace.

The main focus here is directed towards popular social media sites in Germany that have more than two million users. These platforms are required to set up a mechanism that will grant the users a lever to be able to report content that is illegal. Obviously, it would make no sense, when they get a complaint/report to not check whether this content is illegal or not, therefore it is also crucial that overseeing is mandatory. The platform has no more than 24 hours to remove anything that is found to be "manifestly unlawful," and in total seven days to remove any other illegal content. In case those rules are not followed, there are Fines of up to €50 million that will be incurred. NetzDG also adds transparency requirements, requiring platforms that receive more than 100 complaints a year to provide semi-annual reports outlining their policies for content moderation. (Tworek & Leerssen, 2019).

Neverending criticism

Griffin(2021) supports the opinions of Schulz (2018) and Wischmeyer (2020; Liesching, 2020) that the concept of country-of-origin is violated (It states that platforms should be subject to regulations in the EU member state in which they are headquartered) and due to that harmonization is also compromised. Critics also raise other procedural concerns, which includes the speed at which NetzDG was implemented, the fact that a non-independent ministerial is the agency that oversees it all, and the possibility that it might go beyond the jurisdiction of the federal government. But still, the main critics stay the same and they are the following:

Scope

NetzDG has a targeted application, specifically addressing social networks that host a user base of at least two million individuals within the German territory. Contrary to potential misconceptions, the scope of the Act is more nuanced and tailored. This deliberate focus emphasizes the legislation's precision and the intention to intervene selectively in cases where the impact and potential harm are deemed most significant (Zurth, 2021).

Given that NetzDG compliance requires large expenditures on manpower and resources, the main idea for this kind of compliance is to avoid huge expenses that could put smaller businesses at a disadvantage in a market that is already very competitive. Some argue that rather than the existing abrupt division based on a two-million-user threshold, a structured framework that imposes less demanding duties on smaller platforms would be more suitable (Griffin, 2021; Gillespie and Aufderheide, 2020).

The main problem arises while trying to measure the number of users that fall into this category, this is due to the fact that some of them might be obscuring their location by using virtual private networks (so called “VPNs”). This complicates the whole situation and makes it impossible to accurately measure the exact number, but obviously, this is not the one and only complexity, another point of conflict is the non-existent timeframe in which platforms must meet this pre-mentioned user threshold. Time frame once again blurs the already flawed picture, and as a result, all this makes the Act's execution more difficult and leaves room for interpretation and even more potential challenges (Zurth, 2021; Liesching, 2018).

Unlawful Content

Based on NetzDG "Unlawful content shall be content within the meaning of §1(1) (content disseminated on social networks) which fulfills the requirements of the offenses described in §§86, 86a, 89a, 91, 100a, 111, 126, 129 to 129b, 130, 131, 140, 166, 184b, 185 to 187, 201a, 241 or 269 of the Criminal Code and is not justified." But this definition is not enough.

According to Claussen (2018), NetzDG doesn't provide a distinct definition of what could constitute to obvious unlawful content nor does it state what are the obligations arising for the social networks.

As mentioned above, NetzDG does not establish a new liability framework, nor does it criminalize previously lawful expression. Instead of this, it focuses on managing complaints and institutes a compliance mechanism for that. This method pressures the websites into removing unlawful content and aims to hold them responsible for not doing so (Zurth, 2021).

Transparency

When it comes to effective implementation of the regulations, transparency obligations seem to play a vital role by compelling platform compliance. However, some critics argue that transparency (which is often times perceived as a mild regulatory approach) brings about minimal tangible changes. Transparency is exactly the side of the approach that sheds light to platforms' content moderation practices, whether it is effective and it is addressing the broader concerns related to hate speech properly or illegal content online remains quite questionable. At the same time, online regulations remain complex, and they require a complete framework that imposes substantial changes in order to reduce the amount of hateful information that is available (Griffin, 2021; Ananny and Crawford, 2018; Gorwa and Garton Ash, 2020).

According to Tworek and Leerssen(2019), NetzDG's transparency provisions are playing a crucial role, and they are valuable foundation for fostering more in-depth researche. As a result of exploring the nuances of companies' procedural frameworks, transparency reports provide informative elements that go far beyond simple complaint or removal numbers. These reports reveal NetzDG as primarily a "community guidelines enforcement law", that’s why by offering insights into moderation staff training, appeals mechanisms, and quality control processes, reports could actually help enhance understanding. However, critics highlight that standardized reporting formats and more detailed data are still needed to be able to get direct comparisons and a detailed evaluation of platforms' compliance with the regulations across different contents. Nevertheless, prior studies make it obvious that these reports are mostly insufficient to evaluate the frequency and visibility of hate speech, or what was the impact of NetzDG on it; they are, in a certain way, lacking of specificity, standardization between businesses, and impartial supervision.

Over-Removal, a new threat to free speech

There were heated debates and a lot of concerns at the beginning regarding the freedom of expression, when NetzDG started appearing. A major concern that was revolving around NetzDG was that it could potentially incentivize the removal of lawful content (which later became a phenomenon known as "over-removal"). Taking into considerations all the associated costs, heavy penalties and the strict deadline that NetzDG stated, it became clear that platforms would have a strong incentive to strictly abide by the majority of complaints. This dynamic would eventually result in over-removal and when the content is removed excessively to avoid potential penalties, it starts to affect freedom of expression. Critics argued that due to the lack of expertise and time for detailed assessment, online platforms might ignore and close their eyes on caution. Another criticism sharpens the focus on the fact that to be able to evaluate legality, it is essential to have high level of proficiency in German language and jurisprudence. (Tworek and Leerssen, 2019).

Going back to the freedom of speech, but this time from the NSSR viewpoint, Griffin (2021) highlights Balkin's main concerns, according to which Balkin points out few worrisome features when he tries to address the state's cooption of private power. Those are for example, the concept of collateral censorship, where intermediaries don’t have that much interest in user freedom, due to the fact that they don’t want to be held liable for user speech (a common NSSR strategy). This leads for their incentive to lean towards over-censorship and minimize potential liability risks. Balkin originally categorizes social media as intermediaries, however his subsequent publications (Balkin, 2016, 2018a, 2018c, 2020) explore the unique regulatory concerns surrounding social media. In the recent publications he advocates the state-driven speech restriction, regardless his main focus is directed towards preventing abuses of private authority. Lastly, Balkin does not claim that hate speech should never be tolerated on social media or that NSSR is never needed/useful.

Another layer of disagreements regarding Germany's speech prohibitions became noticeable, when some critics started to describe it as overly broad or fundamentally flawed system. A free speech advocacy group also argued that part of the content "should not be criminal offenses in the first place," this cited concerns about blasphemy, expansive definitions of "hate speech," and criminal defamation and insult. More general worries included the possible effects on German democracy, as well as the possibility that the vast amount of deletions may encourage anti-government sentiment or unintentionally spread the access to removed content (this is a phenomenon known as the Streisand effect). Numerous politicians in West Germany, including conservatives and Social Democrats, supported the idea of "militant democracy," according to which speech restrictions may be implemented in order to protect democratic standards (Tworek and Leerssen, 2019).

Handling Complaints

The procedures showed that the legislation has so far failed to recognize more creative and effective alternative strategies for addressing hateful content spread online. These approaches include features that encourage complainants to initiate a dialogue with the author, explain to them why the content could be harmful and promoting so called voluntary-removal. Additionally, end-user controls allow individuals to either block or mute content that they find distressing (in these case scenarios it still might not guarantee the removal). Requiring procedures that are only focused on blocking or takedowns and are supported by administrative fines leaves little opportunity for creative dispute resolution techniques and might inhibit innovation in this field. It is essential to support these alternatives because they can better safeguard users' fundamental rights, give them more individualized safety measures, and give an environment for successfully opposing and changing harmful/hateful or illegal online content (Germany: The Act to Improve Enforcement of the Law in Social Networks, 2017).

Based on NetzDG if the content is “manifestly unlawful,” platforms must remove it within 24 hours, other illegal content must be taken down within 7 days, platforms that fail to comply risk fines of up to €50 million. When it comes to processing the complaints, platforms initially have to assess them against their own guidelines. If they identify the violation, the content is supposed to be removed globally and if not, a second check should be conducted (zurth, 2021). This is the two-step process that allows platforms to enforce their internal standards universally.

Another aspect that Zurith (2021) notes regarding NetzDG, is that the procedure should be the one to guarantee swift resolution of complaints with thorough checks. An authorized personnel who has the authority to either block or delete the content must take note of the report. After this management becomes responsible for overseeing how the complaint is handled and proceeded. Next step would be informing them about the development of events (meaning that both them, including the complainant and the content poster, should receive notifications regarding the decisions that were made step by step). However, one problem still remains unsolved, the complaint procedure is accessible to all registered users and it’s unconventional to require individuals to create an account just to report the issue that they have come across.

Next step in the process would be involving the Federal Office of Justice must first present its case to the Administrative Court to hold platforms accountable for failing to remove or block the content. This process requires the pre-mentioned administrative body to submit a statement from the Social Network. One last issue is that it excludes the option for any oral hearings and specifies that the decision of the administrative court "shall not be contestable."

Private companies should not act as legal arbiters when it comes to the enjoyment of basic rights. According to the NetzDG's Sections 3(2)(2) and 3(2)(3), private actors must evaluate the legality of the contents on their own, based on notifications from parties that are not qualified to make decisions on content legality. International regulations only support accountability in cases when an intermediary disobeys a directive from a neutral, independent overseeing body, such as a court. Given the complexities of the German Criminal Code and the need to assess both the freedom of speech and the psychological state of the person who is posting it, private firms can be considered inadequate to make these kinds of complicated legal and factual assessments. Furthermore, there are no procedures under Sections 3(2) and 4 of the NetzDG that allow users whose legitimate content is taken down to exercise their legal rights. (Germany: The Act to Improve Enforcement of the Law in Social Networks, 2017).

Effectiveness of NetzDG

Utilizing a sophisticated difference-in-differences methodology, a study conducted by Duran et al. (2023) delves into the influence of the NetzDG on hate speech and hate crimes. The findings reveal a significant 4% decrease in the “toxicity of tweets” as a result of the implementation of the NetzDG. Moreover, the study identifies a tangible reduction in hate crimes in municipalities with a higher prevalence of social media users. Estimations suggest that for each standard deviation increase in social media usage, the NetzDG contributed to a 0.9 percentage point decrease in such incidents. These outcomes are further attested by a synthetic control estimate, highlighting the potential that the NetzDG has to mitigate both of the following parts starting from online hate speech and ending with its offline consequences in the real world.

In the period before the second half of 2020, YouTube received a considerable number of reported content, more specifically this number amounted to 1,757,303. Among these reports that were received in total, 23.92% (420,307 items) were successfully removed, in these cases a particular focus was on content that was falling under categories such as "defamation or insults" and "hate speech or political extremism" (Zurth, 2021).

In the latter half of 2020, YouTube showcased efficiency by processing 88.15% of the total 73,477 complaints within a twenty-four-hour timeframe. Notably, only 1,865 posts were blocked under NetzDG, constituting a mere 2.6% of the total blocks, as the majority were enacted based on YouTube's Community Guidelines (Zurth, 2021).

Shifting focus to Facebook, NetzDG-related complaints from 2018 to the first half of 2020 reached 14,114 items, resulting in the deletion of 4,431 items at a rate of 31.39%. This data underscores the platforms' commitment to addressing user-reported content, while concurrently highlighting the relatively low impact of NetzDG-specific interventions.

In the case of Facebook, NetzDG-related complaints from 2018 to the first half of 2020 amounted to 14,114 items, resulting in the deletion of 4,431 items at a rate of 31.39%. This data underscores the platforms' commitment to addressing user-reported content while illustrating the relatively low impact of NetzDG-specific interventions compared to internal content moderation policies (Zurth, 2021).

Recommendations

Griffin(2021) states that NetzDG adjusts only those systems that didn’t address online hate speech properly and it doesn’t demand new approaches at all. Besides this the only platform governance mechanism is moderation and ignores all the other aspects. Law enforcement, social work and education programs can never be replaced by on-platform solutions, that’s why instead of only paying attention to moderation, it’s important to consider these approaches. For example: Griffin considers that regulation must me more systematic only banning individual posts shouldn’t be the only solution, it should also focus on preventive measures and discouraging people from posting/viewing hate speech content.

Most of the critics belong to existing time constraint that are given for procedural requirement, as they might be affecting the accuracy of the assessments. Gorwa (2021) and Claussen(2018) both agree that it should be erased but still it doesn’t mean that they want to get rid of it totally, they both state that the evaluation should be done “expeditiously”. Furthermore, Claussen suggest that the removal of content in social networks shouldn’t lead to a retreat to smaller websites out of the scope of the NetzDG and making legislation applicable to all service providers regardless of their number of users would also solve another issue.

When it comes to transparency, O’regan offers to store the data into a publicly accessible archive or database that would show the reasoning behind the decisions, in this case we will be able to achieve even decisions and avoid different decisions from different platforms.

In case a user surpasses the defined threshold for hate speech, they can be held criminally liable for that and the legal proceedings will be initiated by the authorities, however it would be a further improvement for NetzDG to address the issues of users not being prosecuted by the authorities. (Claussen, 2018)

Griffin also highlight the fact that NetzDG doesn’t look at this topic from the victims’ or the civil society’s viewpoint, giving them the chance for more input could help make NetzDG better by promoting participatory, democratic and non-commercial platform governance structures.

Lastly, it would be better if there were incentives for independent legal experts to take part in this process or to redirect all the reported content to state authority, which would equal to quasi-judicial evaluation (Claussen, 2018)

Keeping it Civil: Facebook's Stance on Hate Speech

First question that I want to address is why Facebook? What makes it so special? And is it even worth discussing?

As of the second quarter of 2023, Facebook has around three billion monthly active users which makes it the most popular online social network globally. In the second quarter of 2017, the platform has already crossed the two billion active user mark and it took little more than 13 years to accomplish this goal. By comparison, it took 11.2 years for Instagram (As of 2021, Meta's Instagram had 1.21 billion monthly active users and It is predicted that by 2025 Instagram will have 1.44 billion monthly users, so it's nowhere close to 3 billion), and little more over 14 years for Google's YouTube to reach this milestone (Statista, 2023).

The fact that Facebook has massive user base might suggest that its policies against hate speech have some kind of potential to impact and influence millions of people worldwide. If we compare Facebook's community standards regarding hate speech with the European Union's 15th Recommendation to combat racism and other forms of discrimination, it becomes clear that Facebook is more capable to explain to its users what constitutes to hate speech (in their context and horizon) better than the European legal document does.

Additionally, if we take the Research led by Plan International (2020) into consideration, it is noticeable that Facebook was the platform with the greatest number of hate incidents (Otero, 2022).

So, Yes! It matters how Facebook establishes and upholds the regulations (Otero, 2022; Gillespie, 2018).

In this section I will be discussing the intricate landscape of Facebook’s regulations against hate speech, what mechanisms they’ve decided to use to combat it and whether it can manage to create safe and respectful online environment free from hate speech. I will also shed light to all the possible ways that Facebook offers to strike a balance between freedom of expression and hateful comments, including: flagging system, automatic detection, handling borderline content, human reviewers and the last hope - oversight board.

Facebook's Anti-Hate Speech Initiatives

Above all, the act of publishing content/commenting online has an obvious advantage (but in reality it might even be a disadvantage) in comparison to communicating in person. Just by using a smartphone/Computer hateful speaker is able to reach a larger audience immediately and doesn't even have to be present physically and spread their hateful views/opinions with no interruptions and hardships. Compared to offline/face-to-face hate speech, Online hate speech can have a unique dynamism due to its anonymity, lack of physical presence, cost-effectiveness, simplicity of usage, and instant transmission (Brown, 2017- What is so special about online (as compared to offline) hate speech?)

Over time it became clear that without safeguarding customer's rights, the platforms would end up nowhere. That's why certain internet companies started to adopt more and more regulations and policies. Fortunately, already at this stage of the digital era, online platforms and websites started to realize that combatting and stopping hate speech from spreading should be one of their goals. What is the driving force behind this is still interesting question, since it can either be their genuine commitment to protect users from harm, or just a mere commercial interest. But as a result, internet platforms started to adopt broader definitions of Hate Speech that exceed legalistic interpretations. (Brown, 2017)

Facebook defines hate speech as "direct and serious attacks on any protected category of people based on attributes like race, ethnicity, national origin, religion, sex, gender, sexual orientation, disability, or disease." We've already discussed that there's no universal definition of Hate Speech, so it could be considered that with this strategy, Facebook is trying to maintain the principles of freedom of self-expression on which the site was built around. Moreover, Facebook has been a witness of a fact that sharing hateful or cruel information frequently receives more criticism than support on the network. (Brown, 2017)

Brown also mentions that we come across consumerism every single step we take and the fact that the consumers have the option to switch to other platforms if they are uncomfortable with certain community standards around hate speech makes it more challenging for the companies. Unfortunately, users may be quite hesitant to leave well-known sites behind considering the benefits and accessibility, but having such a variety of platforms makes it easier for the users to make this decision. (Brown, 2017)

So what does Facebook actually do to combat hate speech? even if it's not the goal for them, they still have to put effort into that to maintain the users.

Facebook has quite detailed content standards to precisely define what content is not allowed on the website (Facebook Content Standards, 2021). Additionally, strict control procedures are implemented when it comes to incitements to violence or dehumanizing speech directed towards individuals on the basis of their race, ethnicity, nationality, gender, or other protected traits. This is ideologically on the same level as Matsuda's (1993) support for hate speech regulation, which emphasizes the need to identify and restrict phrases that perpetuate discrimination (Otero,2022).

There's one thing, Facebook's original definition doesn't mention a purpose to inflict harm, but it is an essential part of the definition of hate groups which may be banned. Facebook additionally indicates clearly that publishing content containing hate speech of others for the purpose of raising awareness or educating others about that hate speech is an example of a "innocent"/justifiable use of hate speech (Sellars, 2016)

Facebook has put a number of enforcement mechanisms in place to guarantee safety on the site, such as flagging system, automatic detection, human reviewers and oversight board. The main focus of all of these approaches revolves around content that is currently being shared on the network. Cummiskey (2019) highlighted that, operating all the pre-mentioned mechanisms together creates a network and composition that is aimed at continuously monitoring and evaluating whether content should be removed or kept accessible. This is an integrated approach that shows Facebook's commitments to uphold a safe and secure online environment (Otero, 2022; Cummiskey, 2019).

It's better to discuss all these mechanisms separately.

Flaging system

Gillespie(2018) describes users as omnipresent, and it's not a lie, they're the ones playing the key role for Facebook's Flagging/Reporting system. There are several alternatives and mechanisms that help users affect the content, such as reporting, hiding comments and also temporarily blocking content. But some factors, such as Facebook's unwillingness to prioritize content deletion can be seen even in the order of these alternatives, which prioritizes user-driven decisions. This approach also applies to the reporting system, which is most useful for comments rather than the profiles themselves, which are classified as violent and illegal activities as well as potentially harmful people or groups. (Otero,2022)

There are two distinct groups when it comes to the Facebook's flagging system, those are regular flaggers (this refers to the users) and trusted partners (Trusted Partner Program includes Facebook's partner groups, who provide in-depth analyses of particular abuses on the platfroms). As evaluators of Facebook's content standards, Trusted Partners can become trusted flaggers with specialized communication channels, although they can also use the standard flagging reporting system. (Otero, 2022; Cummiskey, 2019).

According to Udupa and Pohjonen (2019) Numerous posts on Facebook that combine images, text, and memes somehow mimic the effects of an actual graffiti on the walls. Several layers of meaning are added by using humor, which extends beyond morphological and lexical complexities and despite their seemingly harmless nature as "just funny stuff," these posts promote upholding social norms. However, there are significant consequences when it comes to racializing comedy on social media. The daily lives of already vulnerable citizens of marginalized groups are adversely affected by further more exclusion.

Udupa and Pohjonen (2019) presents the idea of "fun" as a lens which helps us examine the many aspects of online behaviors that consequently lead to everyday exclusion. According to the author, there are four reasons for this fun stuff. Those are using it strategically to gain attention online, embracing this type of language in political discussions, feeling accomplished with little to no assistance, and engaging in a collective (despite the fact that it's anonymous) celebration of violence and aggression.

Facebook's "Abuse Standards" operational manual from 2012 advises content moderators to recognize nine different types of "hate content," starting from so called "versus photos" that compare two people (or even an animal and a person who resembles that animal), also spreading to photoshopped images that negatively represent the subject. Despite all these, Trusted Partners are guided to follow the uncommon attempt of definitional depth, according to which "humor overrules hate speech unless slur words are present or the humor is not evident." (Sellars, 2016).

The development of the process for this reads as follows: In case users (or the pre-mentioned trusted partners) come across content that can be included in the hate speech definition (that is given in the Facebook policy they can report it by the flagging system) they report it accordingly. Upon receiving a report Facebook starts the review process to determine whether this content violates hate speech guidelines or not. Obviously, this review and the assessment is dependent on the reviewer's personal/subjective judgement (Which is yet another problem) (Alkiviadou, 2018).

Artificial intelligence/ automatic detection

Due to the growing popularity of AI and machine learning in content moderation many platforms have already implemented algorithmic techniques to automatically identify and remove inappropriate contents from online discourse. These innovative techniques train machine learning algorithms using massive data sets that have been manually marked for hate speech, offense, or abuse. This approach demonstrates a proactive commitment to use technology to improve the quality and standard of online conversation by quickly identifying and removing inappropriate content and fostering/encouraging an atmosphere that is more polite and respectful (Binns et al., 2017).

Above mentioned approaches involve the development, acquisition, and implementation of algorithms that were designed to detect and eliminate hateful content, but this time the advantage is frequently preventing it from even reaching users' News Feeds. (Cummiskey, 2019).

Facebook made its content moderation algorithms public in mid-2017 and they provided insight into the standards by shedding light to which information classified as "hate speech" is identified and removed. Apparently, the standards took an unwavering stance against insults and incitements to violence, especially when it came to targeting the whole "protected categories." But unfortunately there was an important weakness regarding this layer of protection, it did not apply to groups that were subcategories of these categories. Meaning that while material that was labeled as "hate speech" directed towards Black people would be taken down, identical speech directed at particular subgroups, like Black drivers or Black kids, would not be held to the same standard of attention. That's why there were concerns raised over the impartiality and consistency of Facebook's content moderation procedures. The revelation once again underlined the difficulties platforms have in creating and enforcing content guidelines that take the complexity of hate speech across a range of demographics into account (Gelber, 2019; Angwin, 2017; Klonick, 2017).

In 2017 Mark Zuckerberg made a statement that emphasized Facebook's commitment to developing automated techniques for identifying hateful content. According to Transparency Reports (Facebook Transparency Report, 2020), the platform showcases its removal of 98% of the content that was believed to breach the community standards through automatic detection. Nevertheless, as Chris Guilliard (2020) accurately points out, these statistics might also be hiding the reality that 98% of the data on the network is actually what Facebook itself categorizes as a hateful speech and this is based on kind of an internal consideration, meaning that potentially harmful stuff may remain unnoticed if Facebook doesn’t acknowledge it as hate speech. (Otero, 2022)

Generally, the use of automated methods does not provide a universally applicable answer to all issues, despite the fact that algorithms are acknowledged as highly sophisticated and effective instruments for content control. Algorithms even with their advanced capabilities are always "designed and tested by people, enacted and maintained by people, and deployed or overridden by people" (Gillespie, 2018). This demonstrates the importance of humans in the creation, execution, and maintenance of algorithmic moderation systems, stressing the necessity of continual human supervision and engagement.

Another challenge in hate speech detection lies in the susceptibility of many existing algorithms to manipulation. Gröndal et al. (2018) demonstrated that introducing spaces and misspellings can deceive several prevailing hate speech detection algorithms. Their findings suggest that achieving an algorithm capable of flawlessly detecting all forms of hate speech remains an elusive goal.

Once again, we come to a conclusion that, it is important to recognize that not all automatic detection is designed to remove content entirely. Certain initiatives aim to minimize exposure and restrict dissemination, leading to what Facebook refers to as "borderline content," which exists but has limited accessibility. (Otero, 2022; Cummiskey, 2019).

Borderline content

Based on Heldt (2020) The conversation around "borderline" content on social networking sites focuses on the categorization of speech that is considered unfit/inappropriate for public discussion but at the same time does not violate the principles of free speech. When an opinion statement is classified as "borderline," it suggests that it is on the verge of being illegal without really going over that line.

When it comes to borderline content Facebook's suggested solution is a "penalization tool" that reduces the amount of content which crosses the line and limits engagement of the society without completely removing the content itself (Zuckerberg, 15 November 2018). This kind of intervention targets content that is near to the policy boundaries, but this all happens with the intention of limiting visibility and potential escalation within Facebook's standards. Hence, Facebook's adoption of the concept of borderline material enables it to control the spread of information and manage the visibility of content while still upholding the right to free speech.

Human reviewers

When it comes to the content management process Facebook uses human reviewers to evaluate posts that have been marked as potentially harmful. Platform strives to provide various viewpoints when evaluating content against universal standards by assigning reviewers in accordance with different language and culture. These groups of teams can be separated into external and internal groups, or workers who are outsourced. Additionally, Facebook made a major investment in safety by hiring 30,000 people by the year 2019. Out of these 30,000 people, half of them-15,000 were assigned to content review. This significant investment is once again indication of Facebook's commitment to improving security protocols and content filtering, while human reviewers are playing a fundamental role to prevent and combat hate speech and maintain platform security (Otero,2022).

Reviewers classify already flagged content as “confirmed” for deletion, “un-confirmed” for preservation, “confirmed difficulty” for limited visibility, or "escalate" for additional internal review using the Facebook system. They only have restricted access to user data, and only in extreme circumstances will they be given more access to the information. In order to avoid conflicts of interest and minimize the chances of partiality, Facebook uses alert and separation systems that identify shared connections between users and reviewers (Otero,2022).

In 2020 Facebook decided to suspend the majority of its human content moderation worldwide in favor of automated methods and this caused serious concerns regarding the dynamics of free expression. This change once again highlighted concerns about the fine line that exists in-between excessive and minimal censorship. Additionally, there's still a threat of unknown long-term effects that could endanger user rights and the safety of everyone. Under these circumstances, the Oversight Board was the one to become an essential procedural component with the intention to protect free speech worldwide from the constantly shifting methods of content management while also encouraging user involvement in governance (Klonick, 2020).

Oversight board

To challenge or dispute a moderation decision made by Facebook's content moderation team, users can use the tools that allows them to either express their disapproval and challenge a moderation decision that was made by Facebook's human reviewers. Mark Zuckerberg decided to present the Oversight Board as an external remedy. As a result, the Oversight Board was introduced on November 15, 2019, as part of "A Blueprint for Content Governance and Enforcement." Main advantage of this was oversight board serving as an external enforcement tool, especially when content owners decide to challenge the removal of their work. But as a whole, there are three key concerns addressed by the Oversight Board, including prevention of the gathering of the whole power inside Facebook teams, ensuring accountability and transparency of the process, and making decisions that serve the Facebook community rather than being motivated by profitability (Zuckerberg, November 15, 2018).

The “Supreme Court” consists of 22 members. It gathers people with different professions and experience. In the list we can see: former Prime Minister of Denmark Helle Thorning-Schmidt, Nobel peace prize laureate Tavakkol Karman, Queensland University law professor Nicholas Suzor, Stanford University Professor and Director of the Constitutional Law Center Michael McConnell. These and all the other members of the Oversight Board are highly qualified and reputable and they are the ones who should play the role of "judges", which is a serious challenge, both for them and for Facebook (Oversight board, 2023).

Despite the fact that the Oversight Board cooperates with experts on individual issues, it would be more fair and result-oriented to staff the Council on a regional basis, so that a local perspective is taken into account when solving problematic issues.

The Oversight Board's formation and goals mark a novel improvement in internet governance. Some people worry that the Oversight Board would potentially encourage more restrictions on user speech, especially when it comes down to "borderline" content, and it’s not baseless. In addition, as mentioned in publications like Klonick’s (2020) citation of Ghosh (2019), there are concerns that it could act as a demonstration of self-governance, which could hinder government efforts to control Facebook.

Publicly, Facebook claims to remove over 90% of hate speech on its platform. However, internal documents reveal a different reality, with the company admitting that only 3 to 5% of hate speech is addressed. The discrepancy between public assertions and private acknowledgments raises concerns about transparency. As a result, Facebook's emphasis on high percentages in transparency reports contrasts with internal documents suggesting a much lower efficacy in hate speech moderation, underscoring the importance of assessing the proportion of hate-speech takedowns relative to the total instances of hate speech on the platform. This discrepancy highlights the need for a more accurate representation of Facebook's efforts in tackling hate speech for a comprehensive understanding of the issue (Giansiracusa, 2021).

To put it briefly, even though Facebook offers different mechanisms to combat hate speech, it still faces difficulties with content moderation, including inconsistent application of rules to hate speech across a range of demographic groups and subjective judgements in the review process. But the fact that they are trying to make it somehow transparent, that they're trying to switch to automated techniques instead of human workforce (minimizing the bias and being able to do that as swiftly and quickly as possible), also implementing all the above mentioned systems and most importantly, having defined hate speech with its essence indicates that Facebook is trying to take a one step forward in achieving "safe and secure" environment for the platform's users.

Conclusions

In the contemporary society that we live in, social media companies hold a crucial role in combating hate speech and generally safeguarding the right to be free from discrimination in the digital domain. However, these companies have quite unwavering responsibility, as these platforms are the ones to bear the task of overseeing published content and determining whether to limit the access or remove it in contentious situations. Decision-making within social networks is guided by internally established norms, effectively serving as the digital world's equivalent of laws, gradually supplementing traditional legal frameworks with normative documents.

One significant issue that has been widely discussed is the mislabeling of hate speech. Many individuals lack a clear understanding of the distinction between cyber-bullying and cyber-hate, often using these terms interchangeably without recognizing the impact on societal responses. Frequently, people retract their efforts when they perceive that an incident is not repetitive, assuming that only cyber-bullying, a term more familiar to them, is worth addressing. This misconception hinders effective action against what is commonly referred to as "normal" hate speech, resulting in unintended and undesirable consequences.

Another problematic aspect of this topic is balancing hate speech against free speech, where is the border between them and what society thinks is more superior. Apparently, based on the Rasmussen polling firm’s research only 28 percent of the respondents prefer to ban hate speech. Moreover, 74% of the respondents prefer to allow free speech without restrictions rather than letting government decide which speeches would be tolerated. This once again highlights that they don’t take hate speech seriously and people need more information and raising awareness.

Case law that was adjudicated by the European Court of Human Rights offers one way to address hate speech. These cases serve as examples, establishing who may be held accountable in instances where a citizen's rights are infringed upon on a social network.

Leading European nations have enacted regulatory laws for social media, aimed at enhancing the accountability of these platforms in safeguarding users’ rights to be free from hate speech. For instance: The German NetzDG law represents a significant step taken by Germany to heighten the responsibility of social media platforms, albeit with specific limitations and some shortcomings. Despite these constraints, (such as 1. terms of the threshold for user numbers and the difficulty in measuring and enforcing compliance for smaller platforms; 2. Definition of Unlawful Content: There is a need for a clearer and more distinct definition of unlawful content under NetzDG; 3. Transparency; and 4.existing time constraints affecting the accuracy of assessments and resulting in Over-Removal) the implementation of such measures appears to contribute to a greater level of protection for social media users against hate speech.

The detection and the removal of hate speech on social platforms present a multitude of challenges. Even though there's a lack of a universal definition for hate speech, which significantly complicates the detection process, Facebook explicitly prohibits hate speech in its public standards and defines it in its own way. Social networks encounter difficulties in accurately pinpointing instances of hate speech due to certain limitations in available tools and mechanisms (including flagging systems, use of detection tools and AI, addressing borderline content, having human reviewers and referring to Oversight board as a supreme court). This complexity often leads to the unintentional misidentification and also mislabeling of expressions that may contain hate speech, disrupting the nuanced harmony of freedom of expression. Striking the right balance between mitigating hate speech and preserving the principles of free expression remains a complex and ongoing endeavor for social platforms like Facebook.

In addition to external regulation of social networks, the significance of self-regulation cannot be overlooked. An illustrative example is Facebook's Oversight Board and similar entities in other platforms. These structures not only signify a form of internal governance but also present novel ways to address the multifaceted challenges encountered by social media companies in upholding fundamental human rights and tackling hate speech.

Summary

In the contemporary digital landscape, social media companies bear the complex responsibility of addressing hate speech and safeguarding individuals from discrimination. Serving as custodians of online content, they navigate intricate decisions on content removal or restriction based on internally established norms, akin to digital legal frameworks.

Mislabeling hate speech, often conflated with cyber-bullying, poses a pervasive challenge, hindering effective countermeasures against what is erroneously perceived as accepting hate speech. Hence, striking a delicate balance between mitigating hate speech and preserving the principles of free expression becomes more intricate.

Legal adjudication of hate speech cases by the European Court of Human Rights sets a foundational precedent, establishing accountability when citizens' rights are violated.

Leading European nations respond with regulatory laws like the German NetzDG, aiming to amplify the accountability of social media platforms against hate speech and obliges them to tackle hate speech in their own way.

However, detecting and removing hate speech remains challenging, compounded by the absence of a universal definition and limitations in available tools. Social networks grapple with nuanced identification, often leading to misidentifications and mislabeling, disrupting the delicate balance of freedom of expression.

The interplay between external regulatory measures and the significance of self-regulation becomes more apparent. Entities like Facebook's Oversight Board symbolize not only internal governance but also offer innovative solutions to address multifaceted challenges. As platforms strive to protect human rights and combat hate speech, the evolving dynamics underscore the intricate interplay between technology, regulation, and societal expectations, defining the path forward in this digital era.

References:

1. Alexander Brown (2017). "What is Hate Speech? Part 2: Family Resemblances."

https://www.jstor.org/stable/44980889

2. Alexander Brown (2017). "What is So Special About Online (as Compared to Offline) Hate Speech?"

https://journals.sagepub.com/doi/10.1177/1468796817709846

3. Alexander Brown (2015). "Hate Speech Law: A Philosophical Examination."

https://www.routledge.com/Hate-Speech-Law-A-Philosophical-Examination/Brown/p/book/9781138062740

4. Amélie Heldt (2020). "Borderline Speech: Caught in a Free Speech Limbo?"

https://policyreview.info/articles/news/borderline-speech-caught-free-speech-limbo/1510

5. Andrew F. Sellars (2016). "Defining Hate Speech."

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882244

6. Anthony Lewis (2007). "Freedom for the Thought That We Hate."

https://www.hachettebookgroup.com/titles/anthony-lewis/freedom-for-the-thought-that-we-hate/9780465012930/?lens=basic-books

7. Catherine O’Regan (Year not provided). "Hate Speech Online: An (Intractable) Contemporary Challenge?"

https://academic.oup.com/clp/article-abstract/71/1/403/5258889

8. César Arroyo López and Roberto Moreno López (2017). "Online Hate Speech in the European Union: A Discourse-Analytic Perspective."

https://link.springer.com/book/10.1007/978-3-319-72604-5

9. Claussen, Victor (2018). "Fighting Hate Speech and Fake News: The Network Enforcement Act (NetzDG) in Germany in the Context of European Legislation."

https://www.medialaws.eu/wp-content/uploads/2019/05/6.-Claussen.pdf

10. David Boromisza-Habashi (2013). "Speaking Hatefully: Culture, Communication, and Political Action in Hungary."

https://www.jstor.org/stable/10.5325/j.ctt32bb7q

11. Sahana Udupa and Matti Pohjonen (2019). "Sahana Udupa and Matti Pohjonen. "

https://www.researchgate.net/publication/334587804_Extreme_Speech_and_Global_Digital_Cultures_Introduction

12. Iginio Gagliardone; Alisha Patel;, Matti Pohjonen (2014). "Mapping and Analysing Hate Speech Online."

https://www.researchgate.net/publication/314552833_Mapping_and_Analysing_Hate_Speech_Online

13. Jamal Greene (2012). "Hate Speech and the Demos."

https://assets.cambridge.org/97805211/91098/frontmatter/9780521191098_frontmatter.pdf

14. Tommi Gröndahl; Luca Pajola; Mika Juuti; Mauro Conti; (2018). "All You Need is 'Love': Evading Hate-speech Detection."

https://arxiv.org/abs/1808.09115

15. Heidi Tworek and Paddy Leerssen (2019). "An Analysis of Germany’s NetzDG Law."

https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf

16. Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths (2017). "A Web of Hate: Tackling Hateful Speech in Online Social Spaces."

https://www.researchgate.net/publication/320163517_A_Web_of_Hate_Tackling_Hateful_Speech_in_Online_Social_Spaces

17. Jayson Seaman (2008). "Experience, Reflect, Critique: The End of the 'Learning Cycles' Era."

https://doi.org/10.1177/105382590803100103

18. Jeremy Waldron (2012). "The Harm in Hate Speech."

https://danielwharris.com/teaching/394/Waldron.pdf

19. Jeremy Waldron (2010). "Dignity and Defamation: The Visibility of Hate."

https://www.jstor.org/stable/40648494

20. Kate Kionick (2020). "The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression."

https://heinonline.org/HOL/LandingPage?handle=hein.journals/ylr129&div=48&id=&page

21. Katharine Gelber (2019). "Differentiating Hate Speech: A Systemic Discrimination Approach."

https://doi.org/10.1080/13698230.2019.1576006

22. Katharine Gelber (2019). "Critical Review of International Social and Political Philosophy."

https://philpapers.org/rec/GELDHS-2

23. Laub, Z. (2019). "Hate Speech on Social Media: Global Comparisons."

https://indianstrategicknowledgeonline.com/web/Hate%20Speech%20on%20Social%20Media_%20Global%20Comparisons%20_%20Council%20on%20Foreign%20Relations.pdf

24. Michael Herz and Peter Molnar (2012). "The Content and Context of Hate Speech."

https://assets.cambridge.org/97805211/91098/frontmatter/9780521191098_frontmatter.pdf

25. Natalie Alkiviadou (2018). "Hate Speech on Social Media Networks: Towards a Regulatory Framework?"

https://doi.org/10.1080/13600834.2018.1494417

26. Paloma Viejo Otero (2022). "Governing Hate: Facebook and Hate Speech."

https://doras.dcu.ie/26615/1/Paloma_Viejo_Otero_Final_%20Thesis.pdf

27. Patrick Zurth (2021). "The German NetzDG as Role Model or Cautionary Tale? Implications for the Debate on Social Media Liability."

https://ir.lawnet.fordham.edu/iplj/vol31/iss4/4

28. Rachel Griffin (2022). "New School Speech Regulation and Online Hate Speech: A Case Study of Germany’s NetzDG."

https://sciencespo.hal.science/hal-03586791/document

29. Rafael Jiménez Durán, Karsten Müller, Carlo Schwarz (2023). "The Effect of Content Moderation on Online and Offline Hate: Evidence from Germany's NetzDG."

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4230296

30. Robert S. Tokunaga (2010). "Following You Home from School: A Critical Review and Synthesis of Research on Cyberbullying Victimization."

https://www.researchgate.net/publication/222416503_Following_you_Home_from_School_A_Critical_review_and_Synthesis_of_Research_on_Cyberbullying_Victimization

31. Sahana Udupa and Matti Pohjonen (2019). "Extreme Speech and Global Digital Cultures."

https://www.researchgate.net/publication/334587804_Extreme_Speech_and_Global_Digital_Cultures_Introduction

32. Stanley Fish (1994). "There Is No Such Thing as Free Speech and It's a Good Thing, Too."

https://www.jstor.org/stable/43592917

33. Van Loo, R. (2020). "The New Gatekeepers: Private Firms as Public Enforcers."

https://www.virginialawreview.org/wp-content/uploads/2020/12/VanLoo_Book.pdf

34. Victor Claussen (2018). "Fighting Hate Speech and Fake News: The Network Enforcement Act (NetzDG) in Germany in the Context of European Legislation."

https://www.medialaws.eu/wp-content/uploads/2019/05/6.-Claussen.pdf

35. Zurth, P. (2021). "The German NetzDG as Role Model or Cautionary Tale? Implications for the Debate on Social Media Liability."

https://ir.lawnet.fordham.edu/iplj/vol31/iss4/4

36. Kundrak V. (2020). “Beizaras and Levickas v. Lithuania Recognizing Individual Harm Caused by Cyber Hate?”

https://heinonline.org/HOL/Page?handle=hein.journals/etepnybk2020&div=4&g_sent=1&casa_token=w6mDQsAIW8gAAAAA:HFbRuQsJaW8T2Jt9LUZx2ZzNi1RtK1_FVzIuVD802EBu3vGfHg-FUFFq52zMErNUj2mFspthwA&collection=journals

37. Sidlauskiene J. and Jurkevicius v. (2017) “WEBSITE OPERATORS' LIABILITY FOR OFFENSIVE COMMENTS: A COMPARATIVE ANALYSIS OF DELFI AS v. ESTONIA AND MTE & INDEX v. HUNGARY”

https://heinonline.org/HOL/Page?public=true&handle=hein.journals/bjlp10&div=15&start_page=46&collection=journals&set_as_cursor=1&men_tab=srchresults

38. KNOL RADOJA K. (2020) “FREEDOM OF EXPRESSION ON THE INTERNET - CASE 18/18 EVA GLA WISCHNIG-PIESCZEK V FACEBOOK IRELAND LIMITED:

https://heinonline.org/HOL/Page?public=true&handle=hein.journals/bssr15&div=4&start_page=7&collection=journals&set_as_cursor=0&men_tab=srchresults

39. Cohen-Almagor R. (2011) “Fighting Hate and Bigotry on the Internet”

https://www.researchgate.net/publication/215660527_Fighting_Hate_and_Bigotry_on_the_Internet

40. Delgado R. (1982) “WORDS THAT WOUND: A TORT ACTION FOR RACIAL INSULTS, EPITHETS, AND NAME-CALLING” https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2000918

41. Matsuda M. (1989) 22. “PUBLIC RESPONSE TO RACIST SPEECH: CONSIDERING THE VICTIM'S STORY”

https://www.jstor.org/stable/1289306

Case law:

Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary - 22947/13

Sanchez v. France [GC] - 45581/15

CASE OF DELFI AS v. ESTONIA (Application no. 64569/09)

CASE OF M.L. AND W.W. v. GERMANY (Applications nos. 60798/10 and 65599/10)

CASE OF BEIZARAS AND LEVICKAS v. LITHUANIA (Application no. 41288/15)

Brandenburg v. Ohio, 395 U.S. 444 (1969)

Yahoo!, Inc. v. LICRA

Eva Glawischnig-Piesczek v Facebook Ireland Limited C-18/18

Christopher Anthony McGrath and Another v Professor Richard Dawkins and Others Claim No HQ11D 01227

Others:

Statista, 2023 https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Oversight board, 2023 https://www.oversightboard.com/meet-the-board/

Giansiracusa (2021) https://www.wired.com/story/facebooks-deceptive-math-when-it-comes-to-hate-speech/

Germany: The Act to Improve Enforcement of the Law in Social Networks (2017) https://www.article19.org/wp-content/uploads/2017/09/170901-Legal-Analysis-German-NetzDG-Act.pdf

National Cyber Strategy (2022)

https://assets.publishing.service.gov.uk/media/64e60e4b1ff6f3000d70ae7c/14.283_CO_National_Cyber_Strategy_Progress_Report_Web_v3.pdf