More moderation, more problems?

A qualitative study on the perceptions of Bulgarian users surrounding the debate of content moderation and increasing regulation on social media.

Student Name: Daiana Danova
Student Number: 607477

Supervisor: René König

Master Media Studies - Media & Business
Erasmus School of History, Culture and Communication
Erasmus University Rotterdam

Master Thesis
June 2022

More moderation, more problems?

A qualitative study on the perceptions of Bulgarian users surrounding the debate of content moderation and increasing regulation on social media.

ABSTRACT

Content moderation has for years been a contentious topic of discussion within academia and the public policy realm. The spread of information, misinformation, disinformation or the so-called 'fake news' especially during the Covid pandemic has been facilitated by the means of social media platforms, and this has brought the issue of moderation even further into the spotlight. Misleading information regarding public health could lead to devastating consequences, making the issue of moderation an essential element for its curbing. From discussions about more regulation, to increasing transparency and accountability of platforms, or giving more power and responsibility to the users, there have been numerous ideas and proposals to address the issue, all receiving their fair share of criticism. This whole debate on how to move forward, should however, take into account the opinions and perceptions of the 'ordinary users' of social media platforms, and research on those, especially outside of the United States is still scarce. Cultural norms, political contexts, historical backgrounds, or different legislations all have an impact on how content moderation practices should be addressed and how users would perceive them, so further research is necessary into how those conditions should be accounted for. Therefore, this research aims to answer the question: *How do users in Bulgaria perceive content moderation on social media in light of widespread misinformation about Covid vaccines?*

The results mostly go in line with what has been suggested by previous studies, in terms of the importance of increasing transparency in content moderation practices, reluctance towards the motivations of platforms for doing content moderation and skepticism towards more regulation from the government. Furthermore, this study shows that users understand the complexity of the issue and admit that they should also be placed with some responsibility regarding online content.

Keywords: Social Media, Content Moderation, Regulation, Online Expression, Online Interactions, User perspective

Table of Contents

1. Introduction

Social media platforms have become an immense part of our everyday lives, and as such have also played a pivotal part in shaping public discourse. They are often praised for being the podium of free speech by allowing people to voice their opinions, by creating a space for active public debate and easy access to all kinds of information (Leerssen, 2015). This, however, has created an ecosystem of information, that is populated by all sorts of opinions and false or simply misleading information. As our age is marked by the abundance of freely available content online, there is the issue of the accuracy of the information and of people forming harmful or wrong beliefs, due to misleading information (Stewart, 2021). The Covid-19 pandemic has brought the discussion on this issue even further, by underlining how disinformation shared on social media can represent a risk for public health (Common and Nielsen, 2021). The increasing mistrust in traditional media has created a suitable environment for the spread of ambiguous information or deliberate disinformation regarding public health on social networks, leading to widespread pandemic misinformation (Gollust et al., 2020). Platforms have been externally pressured to take the practice into their own hands and have established content moderation practices. This model, however, establishes them as the new governors and the judges of freedom of expression and content that can be shared online (Klonick, 2018). Following this, there has been a push for further regulation on social media platforms, which in itself has sparked a critique that further government regulation would lead to limitation of public expression and could serve as cover for governments to monitor their constituents (Satariano, 2019). Thus, as platforms have grown in size and importance, so has the problem of moderating them, and the consequences of information shared on these platforms extend way beyond the platforms themselves (Gillespie, 2020). Increasing moderation has also brought up accusations of censorship (Stewart, 2021) and the debate about freedom of speech and what can be said online. Moreover, another problem surrounding content moderation is how the decision is made as to what content should be flagged as 'false', and if it is possible to make the decision in an unbiased manner (Stewart, 2021). Also, while much of the discourse around platform regulation revolves around political actors and platforms themselves, there is little attention to how people perceive further regulation of and by platforms (Riedl et al., 2021). This paper will therefore aim to discover the perspectives of Bulgarian users regarding

content moderation, in order to uncover existing patterns of thought and give insight for future discussions on the topic of content moderation.

1.1. Social and scientific relevance

The study of content moderation is essential for policy formulation and regulation, and scholars have the unique position to contribute to the debate (Gillespie et al., 2020). The focus on content moderation has brought out many aspects and perceived issues - the possible biases of social media platforms (Allen & VandeHei, 2019), and the role they have played in the information sharing regarding important events, such as elections (Allcott & Gentzkow, 2017), and more recently, the Covid pandemic. This has in turn demonstrated the complicated relationship between platforms and society (Riedl et al., 2021). Ongoing debates and discussions on the topic are numerous, however they often fail to take into account what users of those platforms think (Riedl et al., 2021). Despite the vast scholarly research on the topic, there is still the need to expand the scope and range of research in order to receive more valuable insights (Gillespie et al., 2020). Also, understanding how different social dynamics, political backgrounds and values might influence the desire or skepticism towards certain policies and perceptions of content moderation, is essential for making informed decisions (Wihbey et al., 2022). Therefore, this study will contribute to existing research by showcasing the user perspective in a country with a very specific political background, social context and linguistic barrier. The results could then serve as the basis for future research in different societies, in order to assess the impact of those underlying circumstances.

Furthermore, there is existing research on how misinformation on social media impacts people in the political context (Allcott & Gentzkow, 2017), but understanding the scope of misinformation and people's perceptions during a public health crisis is crucial, as people need to have access to reliable information in order make well-informed decisions (Kim & Tandoc, 2022). Therefore, this research will contribute, not only to the general debate regarding content moderation, but also more precisely with regards to the Covid-19 pandemic, and shed a light on how people perceived the information they found online and whether they found that it influenced overall outcomes.

1.2. Research focus

Given the aspects mentioned above, this study will focus on uncovering perceptions, understandings, and opinions of users towards content moderation on social media. The aim of the research will be to contribute to the discourse on content moderation, by adding the aspect of the user perspective. This will be done by using semi-structured in-depth interviews with active social media users, with a focus on Facebook users. The focus will be on users from Bulgaria and the discussion of content moderation will be in light of information shared online during the Covid-19 pandemic and regarding Covid vaccines. The topic guide for the interview will be focused on a particular issue, as this could uncover participants opinions regarding the topic by using a real situation, instead of hypothetical scenarios, and would help them to better understand the aspects of content moderation, in case they are not very aware of the general practice. The debate about information regarding Covid vaccines shared on social media and the focus on public perceptions in Bulgaria has been chosen for two main reasons. First, Bulgaria has the lowest vaccination rates in the EU (ECDC, 2022), which is largely attributed to misdisinformation spread on social media, and lack of trust in government and even medical staff (Kantchev, 2021). Furthermore, as vaccination has become a highly polarizing topic, tensions in Bulgaria have increased, leading to anti-vaccine protestors trying to storm the Bulgarian parliament building, urged by the nationalist party "Vazrazhdane", which currently holds 13 seats in parliament (Tsolova, 2022). Second, Bulgaria is rated with the highest perceived corruption levels among all European Union countries (Transparency International, 2021). Because of this, trust in the government is very low, with only around 30% of the population expressing trust in the national government (Eurofound, 2018). Given the significance of these aspects, Bulgaria would be a suitable example, as it would allow to explore the perceptions of users in a place where information shared on social media played a significant role in shaping public opinion regarding a worldwide phenomenon. By doing this and using this particular scenario, we can understand how users take in the information they read on social media, how they feel about content moderation, whether they think curbing the spread of misinformation shared on social media is the responsibility of the platforms, and how they perceive their freedom of expression in the online world. Since users themselves are the ones ultimately impacted by social media content, how it is moderated and by further regulation on platforms, it is important to gain their perspectives, as this can discover new perspectives and help guide future

discussions on the topic of content moderation. The research question is consequently formulated as follows: *How do users in Bulgaria perceive content moderation on social media in light of widespread misinformation about Covid vaccines?*

In order to address this research question, the questions asked in the interviews will revolve around different aspects of content moderation, how users would feel by increased government regulation and who in their opinion should be responsible for content shared online. Furthermore, their own experiences with content removal or restrictions will be discussed in light of the idea of freedom of expression online. Lastly, participants will be asked about their personal stance regarding Covid vaccines in order to assess whether people with fundamentally different views have different sentiments on the topic and to assess whether people with different opinions had different experiences with the content they viewed and shared.

The next chapter will serve as a theoretical framework and will summarize and analyze previous literature and research on the topic, as well as outline and try to explain the main concepts and phenomena surrounding content moderation, that are relevant for the research at hand. Following that, there would be a chapter describing in detail the methodology of this study, the sampling criteria and the analysis of the results. The following chapter will present and discuss the results derived from the in-depth interviews with active social media users and the paper will close up with concluding remarks, reflections and limitations of the research process and the implication of its findings for future research.

2. Theoretical framework

This chapter of the paper will outline in detail all the related concepts, by analyzing and discussing the existing relevant literature on the topic. It will further present a framework for the theoretical aspects of this study, emphasizing the fundamental topics in the sphere of content moderation on social media platforms. It will explore the nature of content moderation and will try to define the scope of its activities. It will also present the different 'labels' for misleading information, define and interpret the ambiguity of terms such as misinformation, disinformation and fake news. The different understandings and schools of thought when it comes to freedom of speech and freedom of expression will be explored and previous research on user perceptions of content moderation will be presented. Lastly, it will outline the situation in Bulgaria regarding vaccine hesitancy.

2.1. Content moderation

Social media platforms have contributed tremendously to the widespread sharing of information in a much faster and easier manner than ever before, leading to broad connectivity among people and new social dynamics. However, researchers have also brought up the risks and complexities associated with the characteristics of these new platforms, such as the balancing of rights, interests, privacy and honor (Hartmann, 2020). This research will look at social media platforms through the definition by Flew et al. (2019), where they are regarded as 'platforms as integrated software systems, providing the infrastructure, business models and cultural conditions for social networking and publishing, by organizing economic actors to create scalable, re-configurable, multi-sided information and communications markets.'

As outlined in the introduction, the relationship between platforms and society has become highly complicated as is shown by the debate around platform regulation, the possible biases of social networks and their role in major social events such as elections, and more recently - in the supply of information around the Covid pandemic (Riedl et al., 2021). Nowadays, social media platforms have millions of users and given that rapid increase in numbers, platforms have had to discover ways in which to go through, process, examine and curate the large-scale content

produced by their users (Jhaver et al., 2019). Currently, platforms cannot be regarded as a mere mediator of public discourse because they also constitute it (Baym & Boyd, 2012). Those factors have consequently led to the recognition that social media platforms provide a podium for public expression, and knowledge and information sharing, and this should include what behavior is prohibited and how this prohibition would be enforced (Balkin, 2017). This in turn has led to the development of complex and multi-layered content moderation systems that incorporate computational mechanisms, interfaces, and communication practices and most importantly, rules and guidance mechanisms (Myers West, 2018). Despite this, the pressure by governments, regulators, and users on digital platforms to regulate the information shared on their platform has been increasing. Many, however, are critical of the risk associated with the growing power of governments and platforms in being the deciding force of what information is misleading, false or harmful (Zeng & Kaye, 2022). Some of the concerns related to the goals of the European Commission and governments for more regulation on content could fall somewhere in between "proxy governance" and the coercion of private companies by states for the limitation of content, which the government itself cannot lawfully 'censor' (Wehba, 2019), which has happened in a few authoritarian countries. Due to the complexity of the issue, it is still unclear on how authorities can increase the oversight of those platforms, and 'perfect enforcement' is a very ambiguous goal (Renard, 2020).

Content moderation is a very complex phenomenon, entailing social, legal and regulatory aspects, and is defined as the process through which platforms decide and filter the appropriateness of the content shared and the exchanged information according to cultural understandings, legal norms and policy requirements (Witt et al., 2019). Also, content moderation is how platforms shape user participation into a deliverable experience (Gillespie, 2018). Despite its relatively simple definition, content moderation is still a very complex process, raising lots of questions. It is an expensive and intensive approach, requiring the delicate and often flawed decision between what is acceptable and what is not, consideration of different cultural norms, languages, and interpretations, and this makes the standards for content moderation very unclear (Gillepsie, 2018). Consequently, academia has produced an extensive amount of research related to the convoluted systems of content regulation and all the social, regulatory, and political matters that surround it (Bucher, 2012; Gillespie et al., 2020; Helberger, 2020; Suzor et al., 2019). As nowadays the

quantity of online content is enormous, the consequences of harmful or misleading content shared online extends way beyond the platform it appears on and criticism of the failure of how it is being handled has grown, fueled by Myanmar, revenge porn, the 2016 US Presidential elections, Christchurch (Gillespie, 2020), and more recently, by the Covid-19 pandemic. While the screening of the content can occur prior to it being posted on the platform's page, most commonly content moderation occurs after the post has made its way on the platform and is flagged for review by other users who find that it could be harmful and offensive, or that it contradicts platform guidelines (Roberts, 2016). And while some of the filtering could be done through machine learning and automation, the ultimate decision is a vastly complex process, which is beyond the capabilities of softwares alone and requires the human 'factor'. Thus, content moderation still requires a degree of human intervention (Roberts, 2016). Despite the dreams and promises of the CEOs of these platforms for fully automated and AI led content screening that could alleviate the work of human workers and could solve the complexity of the issue and remove harmful content before it ever being seen, in March and April 2020, when many of the platforms shifted to almost entirely automated operations due to the offset of the pandemic, there was an uptick in errors made during the screening of contents (Magalhaes & Katzenbach, 2020). A statement by Twitter pointed out the issue, referring to it as the lack of context that a human brings that results in mistakes (Gadde & Derella, 2020). As already mentioned, there are a number of challenges with content moderation that also translate to automated content moderation. Different ways of expression could lead to very high ambiguity regarding a piece of content, the consideration of context could change the meaning within a post, or the potential bias of the moderator could impact what is considered as appropriate. AI curated content conforms to the same issues, along with a few additional technological challenges (Llansò et al., 2020). Language also presents another barrier to the practices of content moderation and more precisely to automated content moderation. First of all, most of the currently approved language processing tools are only effective for texts in English or other 'high-resource' languages, such as French, Spanish, German or Chinese (Duarte et al., 2017). While the majority of social media users are non-English speakers, machine learning language tools tend to have low accuracy for languages that are not 'popular' on the internet, leading to fewer examples of these languages for the models to learn from, which could in turn lead to bias and disproportionate outcomes (Mina, 2015), and algorithms might amplify existing biases within languages, when excluding the human factor (Duarte et al., 2017). Furthermore,

machine learning cannot make sense of context, doesn't have empathy and cannot grasp humor, sarcasm or irony, and can only mirror decisions that have already been made, by only adapting gradually (Ruckenstein & Turunen, 2019). Human moderators on the other hand can adapt quickly, can react to changes in context and linguistic expressions, and learn (Ruckenstein & Turunen, 2019). Some critics even argue that in most instances content moderation should not be an automated task as it depends on the circumstances of the context and rule of law, and automated systems decrease the transparency and accountability in the decision-making process (Renard, 2020). These factors impede even further the progress in content moderation and present additional challenges for the improvement of content moderation practices.

For all the reasons stated above, the discussion of content moderation is a cross-disciplinary work, since it involves legal, political, economic, social and communication issues (Gillespie et al., 2020). Additional research on the user's knowledge of content moderation practices of platforms shows that knowledge is often unauthoritative and is a result of 'gossip' between online members and communities (Bishop, 2019). For example, the practice of 'shadow banning', which is when a platform makes a users' content less visible, without notifying them, is nearly always denied by the companies. However, users and creators often dispute that, and claim otherwise, which is a proof of the asymmetry in information between platforms and users (Cotter, 2021). In turn, Bucher (2016) emphasizes the importance of understanding how users perceive these platforms, as that would help understand what expectations they might have towards decision making algorithms, which can in turn shape the algorithms themselves. Riedl et al. (2021) add that while much of the discourse around platform regulation revolves around political actors and platforms themselves, there is little attention to how people perceive further regulation of and by platforms. Given this, in the next sections of this study a light will be shed into the discourse about people's perceptions of content moderation and give insights for future research on the topic.

2.2 Information, misinformation, disinformation, fake news

In order to understand the severity of and address properly the threat posed by the quality of information, it is important to develop techniques that help identify misleading information and that help its spread. In order to do that, however, a good understanding of the scope of

disinformation is needed, along with distinguishing it from other types of wrongful information (Fallis, 2015). Harmful content and information include all different sorts of speech - misinformation, disinformation, malinformation, fake news, hate speech, illegal activities and so on. For the purposes of this paper, only the terms related to misleading information will be looked at, as they are the most relevant to this study and the situation with the pandemic.

Furthermore, in this paper misinformation will be used as an umbrella term for all sorts of misleading or false information. However, for a more detailed and precise definition, it is important to distinguish between what information, and misinformation, disinformation and fake news are. The spread of wrongful information is not anything new, however new technologies have facilitated the creation and dissemination of incorrect and misleading information (Hancock, 2007). There isn't one straightforward way to define what constitutes misleading information versus the notion of information itself. What is inherently true can vary depending on what theory lies underneath and within philosophy there are debates as to whether information actually requires truth to be information (Søe, 2019).

Floridi can be regarded as the philosopher of information in the world, and he is the one that established it as a subdiscipline of philosophy, with the main task of defining what information actually is (Fallis, 2011). Floridi (2011) defines information as "well-formed, meaningful and truthful data". This definition, and more precisely the element of 'truth, has caused debates in the academic community and among philosophers'. Some philosophers agree with the notion that 'false' cannot be an element of information and go even further to claim that misinformation or disinformation should not be looked at as varieties of information (Dretske, 1983; Grice, 1989), while others believe that any meaningful data counts as information (Scarantino and Piccinini, 2010; Fetzer, 2004a), since at the moment of its reception from an information source, an individual has solely received information, without yet knowing if it is true or false (Fallis, 2011). Therefore, it can be argued that truth and falsity are not the distinguishing factors between information and mis- or disinformation, but it is rather misleadingness and non-misleadingness, and intentionality (Søe, 2019).

Based on these philosophical ideas, previous research has proposed frameworks in which differentiation is made on the basis of the truthfulness of the content shared and whether the person who shared it had malicious intentions with it or not, and if they were even aware of it being misleading (Wardle & Dershkan, 2017). Furthermore, whether something is malicious is often a matter of perspective, which makes it difficult to determine and challenging to assess. These criteria are the main differentiating factor between misinformation, disinformation, malinformation or fake news.

Misinformation should be looked at as "information that is false, but not created deliberately with the intention of causing harm" (Wardle & Dershkan, 2017; Floridi, 2005). This entails that those who share misinformation are usually not aware of the untrue essence of the information and are usually driven to share it rather because of social incentives, or because it corresponds with and reinforces their current beliefs (Wardle, 2019).

Disinformation, on the other hand, is information or semantic content that is intentionally misleading and is deliberately created to harm a person or a certain group (Wardle & Dershkan, 2017; Fetzer, 2004b; Jackson & Jamieson, 2007; Floridi, 2005). In other words, the main difference between misinformation and disinformation, is that with disinformation, the source is aware of the falsity of the information. Some researchers have argued, however, that a more narrow definition and further distinguishing factors for disinformation are needed, since even if some content would lead people to inaccurate beliefs, it doesn't necessarily mean that it is dangerous (Fallis, 2011).

Fake news, on the other hand, presents the biggest challenge and ambiguity when it comes to finding a definition. While the application and explanation of the term is problematic, it is often used to frame the larger issues in the social media ecosystem (Albright, 2017). The term 'fake news' gained popularity since Britain's referendum on EU membership and gained substantial traction with the 2016 US Presidential election and the Donald Trump presidency (Tarran, 2017). Fake news can be defined as ‘made-up stuff, masterfully manipulated to look like credible journalistic reports, that are easily spread online to large audiences willing to believe the fictions, and willing to spread the word (Holan, 2016). However, there is no common understanding or

definition of the meaning of fake news within academia, and researchers have diverging opinions regarding how and if the term should be used. Burger et al. (2019) for example, refrain from using the term 'fake news' at all due to the imprecise definitions and connotations related to it, such as deceitful, slanted or false, and resort to the term 'junk news' instead. Nielsen & Graves (2019) on the other hand conducted a study on audience perspectives on 'fake news' and found that when asked about 'fake news', people identify poor journalism, propaganda or various advertising, instead of false information masqueraded as the truth or as a news report. Furthermore, people view 'fake news' as a politicized buzzword, that politicians use to criticize news media and social media platforms, and the distinction between 'fake news' and 'real news' is seen as a matter of degree, rather than a definite distinction (Nielsen & Graves, 2019). In line with that, Talisse (2018) argues that 'fake news' is a highly politically charged term and as such people from different political ideologies will disagree on what sources count as 'fake news'. Also, Benkler et al. (2017) point out that 'fake news' could constitute true or partly true information, relied in a way that creates an ultimately misleading message. Fallis and Mathiesen (2019) analyze various definitions of the term and suggest that the most precise definition of 'fake news' would be 'counterfeit news', which is a false story that is presented as genuine news, with the intention to deceive and mislead. They describe genuine news as those that have gone through the standard modern journalistic process with reporters, editors and rigorous fact checking. Some academics have gone as far as suggesting the discontinuation of the use of the concept of fake news (Habgood-Coote, 2019). Egelhofer and Lecheler (2019) however state that abandoning the term is just not feasible and suggest classifying it as a two-dimensional phenomenon. The first dimension would constitute the deliberate creation of pseudo-journalistic disinformation, while the second dimension entails the instrumentalization of the term with the aim to delegitimize traditional media (Egelhofer and Lecheler, 2019). These diverse definitions within academia, and the politicized meaning across public discourse, demonstrate the ambiguity of the concept. As such, there is no single way to define fake news and distinguishing it from other types of misleading information is not a clear process.

Again, to avoid confusion and misunderstandings, in this paper 'misinformation' will be used as an umbrella term for all types of misleading information. In any case, no matter if deception results from a mistake, negligence, or is an intentional act, the spread of false information could be harmful to people (Fallis, 2015). Therefore, apart from the importance of distinguishing

between the different types of false information, researchers, politicians and policy advisors are also looking into ways to curb its spread, now more than ever, with the fast-spreading means created by the Internet and social media platforms.

2.3. Freedom of expression in the online world

Another topic that emerges from the debate on content moderation and how to curb the spread of misinformation, is the one of freedom of expression. It is important to first note that social media platforms are owned by private tech companies and as such they are guided primarily by economic considerations such as profits, managerial interests, good company image and generating engagement. Private interests are usually commercial and not necessarily aligned with government interests such as the protection of freedom of speech and the curbing of hate speech or wrongful information (Irving, 2019). Despite that, however, social media platforms have led to tremendous transformations not only in the way people communicate, but also on how businesses go about their marketing initiatives, how goods and services are sold or on the political activism spectrum as well, since social media platforms have become a tool for political campaigns, organization of protests and expression of political affiliation (Jackson, 2014). Even though social media companies are privately owned, a lot of politicians have started engaging with their constituents through these platforms, which redefines the participation principle of representative democracy (Nunziato, 2018). Given this, social media platforms have begun to resemble public forums, essentially providing space for free expression and the discussion of important social and political events (Jackson, 2014). This status of social media platforms has not only been emphasized by academics, politicians and policy activists, but also by prominent figures. In March 2022, Elon Musk, who is now also the owner of Twitter, posted a tweet in which he describes Twitter as the de facto public town square, and claims that by failing to adhere to free speech principles, it fundamentally undermines democracy (Musk, 2022). Similar to this, many questions have arisen as to how the freedom of speech principles apply to these 'private forums' and how should they be followed in order to not interfere with one of our basic human rights (Leetaru, 2017). Furthermore, the debate about content moderation has in its core the question of how to create content moderation practices that do respect freedom of expression and do not end up censoring unpopular views and opinions, since if social media companies are to be considered

public forums, then the right of freedom of speech applies to their practices (Nunziato, 2018). In any case, even though in many instances social media platforms could be regarded as public spaces, there should be further clarification if that should apply to all of their communication practices. (Bittencourt, 2020). This leads to another discussion, concerning the fact that if social media platforms have to indeed protect freedom of speech and expression, then what is the proper limit of hate speech, and if any hate speech should even be allowed on the platforms (Bittencourt, 2020). Also, who has the responsibility to regulate and determine what actually is harmful content? Last but not least, this discussion has to presume the fact that different countries have different laws and different 'levels' of what speech and expression is protected under law, and what speech would be illegal.

Freedom of speech is a fundamental human right, which has gained even more significance with its entanglement with democracy (Nieuwenhuis, 2000). Article 19 of the Universal Declaration of Human rights gives a universal definition of freedom of expression, which is formulated as follows '*Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers'.* However, despite this general definition, there are different approaches towards what it encompasses, depending on the type of democracy. For example, in many European countries there is extensive legislation against racist utterances, such as inciting hatred, insulting particular groups or denying the Holocaust, and these legislations are not according to the law at odds with the fundamental right of freedom of expression (Nieuwenhuis, 2000). These provisions, however, would not stand the test of the First Amendment in the United States, where the Supreme Court has posed very strict limits on what speech should be restricted (Nieuwenhuis, 2000). These different approaches to human rights have given rise to problems for human rights groups that operate internationally and have posed challenges ever since the creation of the Internet, regarding the different types of content that are deemed acceptable in Europe and in the United States (Nieuwenhuis, 2000). This problem has been extended also to social media platforms and poses a lack of clarity as to what should be deemed as harmful content and how that translates across borders. Following this, some researchers have also stated that it could be problematic to delegate complex decisions regarding freedom of expression to private companies, because of the lack of clarity and guidance of how the law should be interpreted

(Keller, 2018). Social media platforms have already generated different user experiences depending on the laws of the specific countries they operate in (Silverman & Singer-Vine, 2018), however this is often costly, difficult to manage and leading to discontent among users (Keller, 2018).

Social media platforms have been pressured externally to censor content in a few authoritarian and restrictive regimes (Jackson, 2014). This, however, has not happened solely in authoritarian regimes and although not so common in democratic societies, platforms like Facebook have provided user information and restricted content under pressure by the government in countries like the United States and India (Jackson, 2014). Also, oftentimes, platform-established content guidelines are stricter than the laws themselves when it comes to what speech is deemed tolerable, but the principles under which this is established lack transparency, reason giving or review (Wehba, 2019).

Even if the debate about whether social media platforms should follow free speech protections is very unclear, many of the big platforms have committed to transparency in their content moderation practices, and to adhere to the principles of freedom of speech (Meta, 2022; Twitter, 2022). However, despite this, academics have pointed out that after all those platforms are private companies, who pursue their own financial interests, and they often face external and internal pressures to censor or block particular content or particular users (Jackson, 2014). Following this, some researchers have even pointed out that current content moderation practices could result in over-removal of content that is not false or misleading, which some argue could amount to censorship and the under-removal of actual misinformation (Keller, 2018). With growing suspicion about the over-removal of content, even if current practices prove to be very effective, the overblocking, censorship and possibly biased decision making regarding what is deemed acceptable can represent a risk to human rights (Oliva, 2020).

2.4. The user perspective of content moderation

As much of the discussion surrounding content moderation has been on the policy level, between policy makers, academics, journalists, and legislators, it is essential to get the user's

perspective on the topic. Research in that aspect is still scarce and there is still a need for further understanding of how users in different countries perceive content moderation and regulation of content moderation in order to bridge the gap between what the user experiences in reality and what ideas researchers have. Cook et al. (2021) conducted a large-scale survey of social media users on six different social media platforms, where they evaluated users' social media habits and their understanding of current content moderation practices, and what they believe the best practices would be when it comes to content moderation. They find that users tend to be more knowledgeable of content moderation practices in user-moderated platforms, rather than in the ones with employee moderators. Also, they find that users thing none of the strategies that users can employ to moderate content themselves (such as reporting and blocking) are very effective, and they emphasize that in employee moderated platforms, the platform should take more responsibility for moderation, while in user moderated platforms, users should be better educated and responsible for the content they share. In any case, they find that despite the emphasized importance of transparency, users prefer to leave the responsibility of content moderation to the platforms.

The Cato Institute also conducted a Speech and Social Media National Survey on over 2,000 American citizens, in which they find that users do not think that platforms are generally unbiased and fair with their content moderation practices (Kemp & Ekins, 2021). They find that users generally prefer that social media companies give them more control over what they see, and want to be part of the decision-making process, rather than platforms solely restricting their access to harmful content and misinformation. Also, they find that conservative users tend to have their content restricted more often than liberal users do.

Wihbey et al. (2022) conducted research on public opinion in four different democracies - the United States, the United Kingdom, Mexico and South Korea, gathering the perspectives of people on issues such as social media regulation, online expression and censorship. The results from their study show differences across countries in the extent to which people are supportive of freedom of expression or stricter measures for content moderation. These results show the importance of underlying global dynamics in the discussion of content moderation, such as trust in government and societal values in norms. Furthermore, they demonstrate that different

democracies might have different needs when it comes to moderating online content and underscore the nuances in which rules have to be applied.

As the study by Wihbey et al. (2022) demonstrates, there needs to be further research into how people from different backgrounds perceive content moderation and what the implications of those sentiments are when discussing and drafting further regulations and policies for social media platforms. Current studies of the user perspective have focused mainly on the United States, and therefore, there is yet to be a more complete account of the viewpoints of users coming from different countries. It is also important to study how the underlying societal characteristics might have an influence on people's opinions. This study will build on existing literature, by examining the user perspective in a country with some interesting social dynamics - low trust in government and institutions, high collective consciousness, and a unique language, which can be a challenge for content moderation.

2.5. Covid misinformation, the anti-vaccine movement, and the case of Bulgaria

Public health has been a topic giving rise to misinformation and conspiracy theories for long before social media. However, social media platforms have expedited the speed with which they are shared, and this can be seen by the presence of countless conspiracy theories and hoaxes regarding vaccination or cancer treatments for example (Oyeyemi et al., 2014; Drezde et al., 2014). Misleading health-related information can have devastating consequences and provides obstacles to health officials to properly give health education and services (Rodgers & Massac, 2020). Covid-19 was no different, as the focus became not only on containing the spread of the virus, but also on the spread of misinformation around it (Kim & Tandoc, 2022). This led the WHO to even declare an infodemic, as misinformation regarding the virus, remedies and prevention spread online (Thomas, 2020). That was also accompanied by the discourse on rising vaccine hesitancy. Previous research has demonstrated that vaccine hesitancy has been following a slow evolution and has existed long before the Covid pandemic (Jones & McDermott, 2022). The infamous 1998 publication by Andrew Wakefield, where he argued that the MMR (measles, mumps, and rubella) vaccine could cause autism, could be considered the start of the anti-vaccination movement as we know it today and the rise of the internet has played a role in spreading the word around (Hussain

et al., 2018). There is no clear definition about what vaccine hesitancy or being anti-vaccine means, but generally researchers have defined vaccine hesitancy as a delay in acceptance, reluctance, or refusal to vaccinate, despite availability of vaccines (MacDonald et al. 2015).

As of June 2022, only around 30% of Bulgaria's population has been vaccinated, the lowest in the European Union (ECDC, 2022). There is no single reason to explain that, but it is often attributed to a combination of local particularities and global concerns, all revolving around trust (Boytchev, 2021). In Bulgaria trust in authorities has been low, as the government was already for a long time involved in corruption scandals (Boytchev, 2021). As the pandemic begun, cases in Bulgaria were very low compared to the rest of Europe, but the preventive measures taken were so strict, including roadblocks, closing of natural parks and permits to travel between provinces, that people were generally unconvinced about the danger of the pandemic (Boytchev, 2021). Furthermore, many have blamed widespread misinformation on social media as having a direct impact on distrust and vaccination outcomes (Kantchev, 2021). It can be argued that misinformation, especially regarding health, spreads more easily since people are more likely to trust and accept advice from their family and friends or people from their community that they trust (Rodgers & Massac, 2020). The Bulgarian society is marked by a very high collective consciousness, which combined with low trust in authorities, leads people to be way more trusting of information coming from their peers and people inside their communities (Pehlivanova, 2009). On top of that, countries in Eastern Europe, given their Socialist past, have been found to be more susceptible to the lure of conspiracy theories (Marinov & Popova, 2022). All these factors, amid unclear guidance and contradictory advice from authorities, plus general distrust, could be considered a contributing factor to vaccine hesitancy in Bulgaria. The challenge of addressing this issue is emphasized even further, as vaccination can be a polarizing topic and vaccine hesitant people can often form tight communities on social media where any countering information would be deemed untrustworthy (Schmidt et al., 2018), and this presents an issue to content moderation. Therefore, it is important to study how people with different perceptions regarding vaccines perceive the issue of content moderation, in order to make a more informed decision about how this topic can be approached. For all the reasons stated above, Bulgarian users with different opinions regarding Covid vaccines are a suitable subject of analysis.

2.6. Summarizing remarks

To sum up, the debate surrounding content moderation and how it should be executed is a mix of several different factors that lack clarity and are inherently controversial. Platforms are private actors and as such can make their own decisions and restrict speech and content differently than democratic governments otherwise would (Keller, 2019). However, with the way they are shaping communication and public discourse, governments have pushed for further regulation on how and on what grounds content should be removed on these platforms and for more guidance as to what type of expression is acceptable. Addressing content moderation requires in its core a clear definition and understanding of what constitutes misinformation and what constitutes subjective opinions, what type of speech is not harmful or misleading to others and whether those types of expression go in line with the fundamental rights of freedom of expression. All these factors are heavily analyzed by academics, researchers, policy advisors and politicians but there is still more progress and work to be done. In addition to the opinions of professionals, however, it is also important to understand how users perceive and what their sentiments are towards the way content is monitored. For this reason, this paper will aim to give some insight into this aspect, in order to provide a basis for future research on users' perceptions.

3. Methodology

In order to answer the research question at hand "*How do users in Bulgaria perceive content moderation on social media in light of widespread misinformation about Covid vaccines?"*, a deeper understanding of people's perceptions, sentiments and opinions towards content moderation is needed. The theoretical framework has already outlined the existing debates and problems related to content moderation and discussed and analyzed in detail existing literature on the topic, including the diverging concerns and arguments arising from it. As mentioned in the introduction and theoretical framework of this paper, previous literature has extensively covered the legislative debates, difficulties with navigating different cultures, political regimes and understandings of freedom of expression. However, there has been little research into how users of these platforms actually perceive these issues and how aware and well-acquainted they are with content moderation itself, and how and why it is performed. Furthermore, as the future of content moderation, at least in the European Union, is a big topic of discussion in the European Commission, it would be important to gain deeper understanding of how the issue is understood in one of the smaller and newer member states - Bulgaria, and whether more education regarding content shared online and how its legitimacy can be verified could be needed, in order to make users more knowledgeable and understanding of what content moderation is and why it is necessary. Therefore, this study will contribute to existing literature through a series of in-depth interviews that aim to discover recurring patterns in what Bulgarian users' opinions and perceptions of content moderation are, if they understand what it is and how it is done, and what their sentiments are towards further regulation on the matter. Again, Bulgaria is a suitable example, since, as it is a smaller member state, it is often not at the center of the debates surrounding content moderation and can present a challenging case given the different language and often particular attitudes towards sensitive topics and 'being told what can be said or not'. Furthermore, as the Covid pandemic presented new challenges regarding the wave of misinformation on social media, and Bulgaria proved to be an interesting case with that, asking questions particularly regarding Covid and vaccine information that was shared online will allow this research to discover understandings of how people perceive the topic of moderating content shared online, by using a recent and existing example instead of subjective scenarios. The aim of this chapter is to provide

a comprehensive overview and justification of the methods chosen for this research and to elaborate on the sampling process, data collection process and data analysis process.

3.1. Research methods and design

As mentioned previously, this study will employ qualitative methods, namely in-depth interviews. Qualitative methods are used when researchers need information about peoples' experiences, attitudes, or opinions about the issue at hand (Hammarberg et al., 2016). As the data needed conceptualizes the opinions, perceptions, or emotions of an individual or a group, qualitative methods would be of higher relevance than numerical quantitative research (Hammarberg et al., 2016). Also, as this study focuses on relevant social affairs and the focus is put on gaining in-depth insight of perceptions and understandings of the topic, a quantitative method might be too constraining (Brennen, 2017). Going further than that, in-depth interviews are typically employed when the researcher seeks more profound and detailed information about a person's opinions, understandings or thoughts and strives to uncover new issues and problems (Boyce and Neale, 2006). In-depth interviews require interviewing a small number of participants, which in this case will be ten, in order to uncover underlying patterns (Boyce and Neale, 2006). Moreover, interviews are a suitable method for this research because as the study aims to measure the perceptions and sentiments of Bulgarians towards content moderation and further legislation and regulation by the government of content related to Covid shared on social media, interviews will provide more profound knowledge than for example, surveys. Additionally, interviews will be used instead of focus groups, as overall sentiments towards Covid vaccines can be a polarizing and sensitive topic, and participants might not feel comfortable or be truthful in discussing a sensitive topic in front of others (Boyce and Neale, 2006).

To fully make use of the advantages that in-depth interviews offer, in terms of flexibility and gaining profound understanding, using semi-structured interviews was chosen for this research. An interview guide was prepared and used in all the interviews, however it was solely used as a basis and as a guide for the most relevant concepts and questions that should be covered with each participant. The questions were not asked in any exact order and were chosen depending on the directionality of the interview. Furthermore, some questions from the topic guide might

have been skipped and other ones asked instead, depending on the course of the discussion. This allowed for more flexibility and for uncovering new perceptions or topics that were not considered by the researcher prior to starting the interview process. Moreover, allowing for flexibility in the topic guide ensured the removal of any potential bias and subjective opinions instilled by the researcher when constructing the questionnaire. The semi-structured interview method proved to provide a far-reaching insight for answering the research question in this study.

3.2. Sampling criteria

When conducting in-depth interviews, the first step is to identify what information would be required and who is a good source of that information, in order to establish the sampling criteria for the participants that would be recruited for the research (Boyce & Neale, 2006). The proposed sample for this study is people aged between 18-50, who are active users of Facebook. The research will focus on Facebook, as this is the most used social media platform in Bulgaria (Statista, 2021). The age group has been chosen as people between 18 -50 years old make up the highest percentage of Facebook users in Bulgaria (Statista, 2021). The goal of this study is to reach theoretical saturation, which Glaser and Straus (1999) defined as the point when researchers see similar instances over and over again and find no additional data for a given category, which allows them to develop the characteristics of that category. Therefore, this paper aims to uncover patterns of behavior and perceptions, by using a small sample, that can be divided into relevant categories. Participants were picked using non-random, purposive sampling. In qualitative research, purposive sampling is typically used to identify and select participants that are information rich and very knowledgeable about a topic, in order to obtain comprehensive results with limited resources (Patton, 2002). In this research, purposive sampling was found to be the most suitable method, as it allowed the researcher to handpick participants of all backgrounds, and more precisely both people who are pro and people who are against the Covid vaccine. In turn, a total of ten participants were recruited, from various Facebook groups, with the criteria that they are active users of the platform (this means that they post or comment frequently on other people's posts in the group they were found in) and have diverging opinions towards Covid vaccines. This is initially established through their public posts and comments on the group they were found on, without going through their personal profiles, in order not to invade their privacy and raise ethical concerns. A person will be described as vaccine hesitant or against the vaccine using the definition outlined

by MacDonald (2015), which goes as follows: *Vaccine hesitancy is understood as a delay in acceptance, reluctance, or refusal to vaccinate, despite availability of vaccines*. Covid vaccines in Bulgaria are widely available, therefore a participant will be regarded as being against the Covid vaccine, if they answer that they are unvaccinated, and the reason for that is hesitancy or distrust in the vaccine. Pro-vaccine participants will be defined as those who have taken the Covid vaccine and believe in its effectiveness and trustworthiness. Pro-vaccine people are usually expected to believe that anti-vaccine people are generally misinformed, because they derive their information from unscientific sources, while anti-vaccine people are expected to express more skepticism and believe that pro-vaccine people are deriving their information from biased or manipulated sources (Maciuszek et al., 2021). Furthermore, anti-vaccine groups are often associated with orthodox religiousness, moral purity concerns, conspiratorial thinking, and hierarchical worldview (Hornsey et al., 2018). Also, vaccine hesitancy and rejection has played an important role in the Covid pandemic, and it is important to understand the potential underlying role of social media in these dynamics. Therefore, it could be assumed that those two groups would have diverging sources of information and different opinions towards information found online, and that makes them suitable for analysis. Choosing participants with different sentiments towards Covid vaccines is important, because it will ensure the inclusion in the sample of people with different opinions, which can in turn uncover contrasting attitudes towards content moderation and diverging exposure towards misleading information on the platform. The objective of qualitative research and in particular in-depth interviews is not necessarily to be highly representative and involve a high number of participants, but that the sample allows to uncover recurring structures that can then be generalized to the rest of the society. Thus, picking people with diverging opinions would allow us to uncover more and different pre-existing structures.

The use of the case of both pro-vax people and anti-vax people in Bulgaria, allows to comprehend a larger population and shed light on an issue, using smaller units and a specific case (Seawright & Gerring, 2008). In this research this is relevant as it would help generate an in-depth understanding of a very complex issue in a real-life setting (Crowe et al., 2011). Furthermore, particular cases are often used to explain causal links that have resulted from new policy initiatives, and it can help shed light as to what course of action should be implemented instead of another (Yin, 2009). According to Siggelkow (2007), case studies are useful to demonstrate the occurrence

of a phenomena and to identify research gaps or previously unidentified scenarios that can guide future research. In this study, the case of pro-vax and anti-vax people in Bulgaria will identify possible gaps in the consideration of the perceptions of people with different opinions, on a particularly polarizing topic, when implementing regulations on social media on the European Union level. The drawback of using specific cases, as would be in this scenario, is that they are very specific, and generalization might be misguided. However, it could be used to generate an initial hypothesis and uncover a new side of the problem, which can be further tested in future research (Abercrombie et al., 1984).

3.3. Operationalization and data collection

As the researcher lives in the Netherlands and the participants are Bulgarian, most interviews were conducted online, using the Zoom service, while three of the interviews were conducted in person, since the researcher found participants from the target sample who live in the Netherlands. Typically, in-person interviews are considered to be the standard, as they provide a comfortable environment, but it has been suggested that they are only marginally superior to video interviews (Krouwel et al., 2019). The limitation that was caused by conducting the interviews online, was the decreased ability to pick up on non-verbal cues and the lack of personal contact and exact face-to-face interaction, which prolongs the building of rapport. As the topics discussed are indeed sensitive, ensuring a controlled and safe environment for each participant was of utmost importance (Babbie, 2014), so all participants were interviewed in a comfortable environment for them, ensuring their privacy and anonymity. Given this aspect, conducting the interviews online proved to be a beneficial option, as nearly all participants were doing the interviews from the comfort of their own homes, which could be considered the 'safest' environment. Also, by not conducting the interviews face-to-face, and by being able to be at home, the participants felt more in control of the situation. All participants were presented with and asked to sign an informed consent form, including the consent to record the entire interview. Their consent for recording was asked once again at the start of the interview. In addition to that, the researcher made sure to make the participants feel comfortable and at ease and refrained from expressing her personal opinions on the topics discussed, so the participants would feel comfortable to speak freely.

The interviews started with ice breaker questions in order to establish rapport and make the participants at ease. Then, there was a short conversation in which the goals of the research were explained, participants were reminded of the informed consent they signed, were again ensured about the privacy and anonymity of their responses and that they would only be used for the research at hand and were again asked for their consent to record the interview. All participants were also given the option to receive the conclusions and results from the study, once it was completed.

The ten interviews lasted anywhere between 40 and 60 minutes, not counting the time the researcher and participant spent talking before and after the interview itself. Some questions from the topic guide were used in all the interviews. For example, the questions regarding how much the participant knew about content moderation, how they thought it was currently executed, their opinion about more government regulation given the amount of misleading information that was shared about Covid on social media platforms and of course their stance towards Covid vaccines, were essential for the research question at hand and thus were included in each interview. Following the semi-structured approach, the rest of the interviews were adjusted depending on the course of the discussion and depending on how knowledgeable and opinionated the participant was on the topic of content moderation. For instance, people who had more profound familiarity on how and why content moderation is executed and on the whole debate surrounding this issue, were asked more detailed questions about different aspects of content moderation. Participants who appeared to have less awareness of the issue were given some explanation as to what the debate is and were asked more general questions, such as their opinions on who should be responsible for it and what would be their preferred way of it being approached.

The recordings of the interviews were initially transcribed using a transcription software called Otter.ai, for the interviews conducted in person, or the Zoom transcription feature, for the interviews conducted online and were then checked by the researcher for any gaps and inconsistencies, which were then corrected. The interviews were all conducted in English, to ensure that there were no inconsistencies or misunderstandings due to the need for translation and as all of the participants indicated they were comfortable with doing the interviews in English. For this reason, there was no need for translation of the interviews, prior to the transcription.

3.4. Analysis of results

The data collected was consequently analyzed using an inductive approach, which establishes a framework and categories based on evidence in the collected data (Thomas, 2006). Inductive analysis creates categories from the material itself, and not from an established theoretical framework, which is suitable for interviews, as not all the material in an interview was regarded for analysis (such as off-topic or ice-breaker conversations) (Mayring, 2014). Results and conclusions were drawn based on the categories and frequencies of occurrence of each category. In particular the data was analyzed using the grounded theory approach, in which understanding is drawn based on the data collected and not on existing theoretical models to explain the results (Thornberg & Charmaz, 2014). The thematic analysis was done following the six steps of thematic analysis described by Braun & Clark (2006).

In order to do this, all the transcripts were put into ATLAS.ti, which is a qualitative data analysis tool that helps to analyze unstructured data and provides tools to code findings in the initial data material (Silver & Lewins, 2014). The initial coding was done by selecting all the relevant parts of the data (excluding general, icebreaker questions). After that all the codes were assigned to the potential main categories and themes that were formed based on the theoretical framework and on the overall topics that emerged from the interviews. Those themes were subsequently re-checked with the contents of the interviews and finally, all the categories and themes were assigned with relevant names. The results chapter will outline the main themes that emerged from the analysis and will explain in detail what the perceptions of Bulgarian users regarding content moderation on social media were.

4. Results

As explained in the methodology chapter, this study employed semi-structured interviews with users of social media platforms in order to identify their perceptions of content moderation. This chapter will outline and analyze in detail the results from this research, by grounding them in the concepts and academic literature discussed in the theoretical section. After the coding and thematic analysis, there are 4 main categories that emerge that describe the perceptions of users and can be used to answer the research question "*How do users in Bulgaria perceive content moderation on social media in light of widespread misinformation about Covid vaccines?".* Those key categories and their related themes will be thoroughly discussed in this chapter using the data produced from the interviews, grounded in the concepts discussed in the theoretical framework.

4.1. Research findings

The following table outlines the main categories and themes identified after the open, selective and axial coding of the thematic analysis, which was done as described by Braun and Clarke (2006). The first category revolves around users' perceptions of who and to what extent should be responsible for content moderation and how much should different actors be involved in the process. The second category relates to how the participants perceive freedom of expression online, regarding not only their own, but also other people's opinions, and what type of content in their point of view should be moderated on social media platforms. Lastly, the third category relates to the social implications that content moderation has. This category involves both the impact of different types of content, but also the reasons and benefits different actors might have for and from content moderation. This part also touches upon the topic of increasing transparency in content moderation practices.

Table 1: Main categories and themes

Category	Main themes
Content moderation in practice	The role of government Platform liability The role of the user Transparency in content moderation
Opinions and online expression	Presence of opinions and discussions Sentiments towards online expression Opinions vs (mis)information
Social implications of content moderation	The bias in moderation Platform vs user interest Freedom of expression online Impact of online information

4.2. Content moderation in practice

The first category in the results from this research uncovers who users believe should take the responsibility for content moderation. Interestingly, all of the interviewees had similar perceptions of who should be responsible for content moderation. All of them mention the need for government guidelines and regulation on online content but emphasize their skepticism towards too much government involvement in the process. Similarly, they show distrust towards letting the platforms make the decision about what can be said and what cannot, but at the same time point out the complexity of the issue and indicate that social media platforms should not be judged too hard on how they manage this task. Given this, they also place some responsibility on the users to be more mindful of the things they see online.

All users interviewed indicated that responsibility lies somewhere in the middle between government, platforms, users, and a potential independent party. The prevailing sentiment towards government involvement and further regulation of social media was that of skepticism, especially because of the threat of censorship and the danger it could pose to the free flow of opinions.

> *I do feel like every time the government gets too involved, that could lead to censorship, it could lead to loss of freedom of speech, loss of democracy. And I mean, the internet and social media in particular, they've always been branded as these democratic places where everyone can get their opinion. So no, I don't think that ultimately the government should be involved to that extent, I don't know to what extent I want the government to be involved. (Natalia, May 2022)*

Natalia was not alone in this sentiment as nearly all participants expressed similar concerns, as to the further involvement of the government in content moderation practices. Most participants indicated that in their opinion the political realm and social media companies are already very intertwined, which makes platforms more leaning in one political direction rather than the other. By increasing the regulation of platforms, participants fear that information online will be skewed towards what the government is promoting or pushing for and would limit the freedom of expression and free flow of information. This goes in line with what some critics, both academic and political actors, have pointed out, in terms of increased government insight on moderation practices (see Hicks, 2021; Wehba, 2018). This has already happened in some countries with authoritarian regimes, such as China and Iran, where governments place very tight restrictions or even block some platforms from use in the countries (Jackson, 2018). With increasing external pressures for regulation, however, users have expressed their concerns with reverting to a society whose expression is limited and guided by the government.

> *During Soviet times, during the Iron Curtain the government controlled everything we said and what we did, and I don't believe they should be messing up with my time and what I say. (Eduard, May 2022)*

In a country like Bulgaria, these sentiments are probably strengthened by memories of an authoritarian past and further emphasized by lack of trust in the government. This was demonstrated by some statements such as:

> *…I have what I would describe as a healthy level of distrust that any government. (Kaloyan, May 2022)*

Thus, as discussed in existing literature and as has been demonstrated by the data from the interviews, increasing government involvement in content moderation definitely comes with a concern about how much the government will be involved in the process and about the potential censorship of opinions that challenge the status quo. This effect could potentially differ across countries, since different socio-political circumstances can influence sentiments towards online content (Wihbey et al., 2022).

Despite this skepticism, participants also highlight the need for overall guidance and a general regulatory framework established by governments. All interviewees pointed out that they believe that the government should help with establishing overall rules for platforms as to what content should be removed, and ensure the follow through of the process, but they should not go beyond that and monitor content themselves or have a say in individual cases.

> *Probably some government regulation could be in some cases, good, but only in terms of the structure of how things should run. So, if the government can impose some sort of regulation, saying: you must have this third-party moderator and you must, you know, leave, let them assess completely your content and your information provided on the platform, but not in terms of the government actually having any say in individual cases. Because at the end of the day, media is the one way to challenge the government as an individual. (Lora, May 2022)*

What Lora has indicated is that the government should solely provide the structure for how things run, and someone else needs to be responsible for the actual assessment. Similarly, another interviewee, Martin, mentioned that the government should establish the minimum criteria for

what cannot be said online, but as long as that is met, people should be free from interference in expressing their opinions online, and that the actual moderation should be done by algorithms. These results underline the complexity of the situation and go in line with the growing literature on how platforms exist side by side with the state and how intertwined the two are with policy decisions on user speech and security (Benvenisti, 2018; Keller, 2019). By operating in a given country, it is obvious that the platforms have to follow the laws of the country. When it comes to content moderation, drawing the line is still challenging, and the results from these interviews demonstrate that users themselves have the same skepticism that many critics have pointed out, but still believe in the importance of governmental guidelines and principles.

When it comes to the role of the platforms in content moderation practices, all interviewees acknowledge that currently the process is mainly the responsibility of the platforms. When asked whether platforms were doing a good job with their practices, the sentiments were mixed. Some participants thought that platforms were doing a very good job with how content is being moderated.

> *I think, right now, it's quite good in the sense of how content is moderated. (Hristo, May 2022)*

Those participants highlighted the complexity of the task and identified that their feeds have been 'free' from harmful or misleading content for the most part and that they have been exposed to all types of opinions and discussions, which they find appropriate for a social media platform. On the other hand, a part of the interviewees identified that in their opinion platforms had a long way to go before doing a good job in managing the content that is on their platform. Some noted that they still see plenty of obviously false or misleading content and it is sometimes hard to distinguish it from 'truthful' content. Interestingly, however, some of those participants also indicated that the reason that platforms are not doing a good job is because they often tend to remove too much, instead of too little. For example, Natalia commented:

> *I feel like the problem is not so much that they're moderating very little, it's more that they're moderating too much. When it comes, for example, to platforms like*

Twitter, I would say they're even overachieving with their content moderation to the point where they're like censoring people, and I don't think that's fine. (Natalia, May 2022)

Going in the same direction Martin also pointed out that platforms might be removing too much, which could be unfair to some people who have their content removed for sometimes not very obvious reasons. These sentiments go in line with what has been argued by Keller (2018) that current content moderation practices often result in censorship and over-removal of information that should not be removed, and under removal of actually harmful content. This indicates the need for better understanding of the basis of the content moderation practice and could mean that the discussion should focus on improving the practices as they are now, rather than establishing more and stricter rules or regulations. The core of this is ensuring that people maintain their freedom of expression and that removal is based solely on harmful or misleading content, excluding personal opinions. This will be discussed further in the next parts of this chapter.

Despite disagreeing on whether platforms are doing a good job, interviewees agreed that platforms shouldn't be judged too harshly on this matter. They note that content moderation is an extremely difficult task, with many ambiguities, and that it would be very hard, if not impossible to reach a point where it is done well according to different people's standards. Therefore, they suggest that despite the responsibility social media companies have, we shouldn't expect from them to solve this issue right away.

I think that there is a continuous belief that social media platforms are doing something wrong. In both cases, if they do moderate the content, there are going to be people who say they shouldn't. If there are not, there are going to be people that would say, sure, they should. And I think they are faced with a particularly hard philosophical question, which they surely didn't think they would face once creating the social media platform itself. (Nikolay, May 2022)

In light of this, the interviewees also made observations about the responsibility of the user and how users can also take part in the decision-making process of what and how content should

be moderated. On the one hand, participants suggested that adult people should put some thought into and be careful with what they share. They put emphasis on the fact that platforms do not have the responsibility to educate people on what is right and what is wrong. Furthermore, people talk outside of social media, so even if platforms limit the sharing of misleading and hurtful information, there is a high chance people would still hear it 'on the streets'. Thus, when discussing content moderation, and evaluating the work of the platforms in that area, we should also look at the users, and put some of the responsibility on them.

> *Of course, I believe it's the responsibility of the person to limit themselves on what they're saying…. let's say we live in a world where you can have the best social media, the most censored one, but you go outside, and you'd hear everything. So, if we consider that there are issues right now with social media, I mean, us people need to be able to filter for ourselves, what makes sense and what does not…. (Hristo, May 2022).*

Also, some of the interviewees noted that the spread of misleading or false information can also be blamed on the users. For example, Martin noted that it is part of human nature to be captivated by controversial content and that people find it interesting and tend to share and reshare, even though they do not find it believable or useful. In that sense, previous research confirms that controversy tends to spark conversation, and that controversial information tends to be discussed more often (Chen & Berger, 2013). On the other hand, in terms of user responsibility, many of the interviewees indicated the importance of including the users and having an open discussion about the rules and principles of content moderation. They emphasized that in order for users to understand and support content moderation, they should be included in the process and be given the opportunity to voice their opinions.

> *I think it should be an open discussion for all the people in the world to discuss whether content moderation is needed to see both perspectives to hear both sides of the argument, because I'm sure that me as a regular person is missing out a lot of important points and maybe continue moderation is super important, but I just don't realize it. (Stefan, May 2022)*

This goes in line with the whole idea of this research and what Riedl et al. (2021) highlight in terms of the current discussion focusing on the governments and platforms, without taking into account the user perspective and opinions.

In line with the inclusion of users in the discussion on content moderation goes the next theme in this category that emerged from the analysis of the interviews, which relates to transparency in the content moderation practice. All participants underlined the fact that they are not aware of the rules or on what grounds content is removed and thus they don't think the process right now is transparent. They advocated for more transparency, which in their opinion would make users of the platforms more trusting and less judgmental when it comes to content moderation.

> *On the other hand, I think that they could be more transparent with their users, as I said, regarding what should be shared with them, and maybe the user should have the power to decide what to see and what not to see (Nikolay, May 2022)*

According to them, transparency would mean being very open about the rules and grounds on which a piece of content is removed, give reasoning as to why a certain piece of content is flagged and provide reliable sources instead for each post that is flagged as potentially misleading information. Furthermore, Lora, one of the interviewees, noted that if we are trying to achieve more trustworthiness of information shared online, a big part of achieving that would be by being honest and transparent about the way the reliability of information is being decided. This goes in line with previous research that implies that transparency in content moderation is essential to ensuring users are aware of what they can say and how they can behave online (West, 2018; Jhaver et al., 2019; Suzor et al., 2019). Platforms like Facebook have previously committed to transparency, but that dedication has been often put into question, due to their tight control of information on the platform (Felella, 2021). In any case, there have been frequent calls in demanding more transparency in content moderation, as this would allow for a more informed public debate (Suzor et al., 2019). Given that similar sentiments were expressed by users, it can

definitely be argued that increasing transparency would be a step in the right direction for informed debate.

4.3 Opinions and online expression

Content moderation and online expression are deeply intertwined. As the discussion about more regulation and rules on content moderation is growing, there have been various criticisms as to what that would mean about freedom of speech. Some academics and legislators consider that increased regulation on social media, would lead to lower protections of freedom of expression, while others argue that rules and laws would be appropriate and not deteriorating to democracy, if they make the internet a safer space (Satariano, 2019). As social media platforms are often praised for being the podium of free speech and giving access to alternative opinions (Leerssen, 2015), it is essential to see what influence content moderation has on that aspect from the user perspective. When inquired about how they feel about their ability to express themselves freely online, users shared some different experiences. Generally, people like Eduard, who were more skeptical towards Covid vaccines and online Covid information and could thus be classified as holding more alternative opinions, felt more restricted in what and how they can express themselves online. However, all participants agreed that social media platforms should allow the free flow of information and opinions as that would enhance the dialogue about different topics. All of the people interviewed expressed their support for allowing diverse opinions on social media platforms, as this was the only place where those could be presented and as that would allow people to get exposed to alternative thoughts and perceptions, and that would allow them to make more informed decisions

> *I think social media is one of the, if not the only medium of sharing alternative thoughts, because the news, the official news on the TV, or the official news that share on social media would never share information that they're not at least 80% sure about. (Nikolay, May 2022)*

When asked about the diversity of opinions they could find on the platform, in particular regarding Covid and Covid vaccines, interviewees had seemingly different experiences. Some of

them identified that they could see different opinions and discussions happening on the topic, while others indicated that they saw predominantly polarizing opinions, confrontation and no healthy discussions. Most of them indicated that this would probably have to do with the algorithm, rather than with how content moderation is being done. Social media algorithms are designed in such a way as to promote content similar to what the user interacts with and what the user is assumed to be interested in, and form groups of like-minded individuals, which has been blamed to result in filter bubbles and echo chambers (Cinelli et al., 2021). In that sense, there are multiple factors that impact information spreading. One of them would be online polarization, which can aid in the spread of misinformation (Vicario et al., 2019). The other one would be the fact that individuals tend to favor information that confirms their own beliefs and interact with people and content that reinforced their beliefs, which causes and reinforces the existence of echo chambers (Cota et al., 2019). In relation to that, all of the interviewees for this research indicated that they wouldn't mind and would even like to see comments or posts with diverse opinions, and most importantly different than their own. Lora and Nikolay indicated that this could lead to healthy discussions in society, where people learn from each other and see different aspects of a given topic, which they otherwise wouldn't. Stefan mentioned that currently there are no discussions, because alternative opinions are often removed, but that he hopes that with the improvement of content moderation, platforms would be a safe space for online discussions, rather than just reading what someone has decided is good for the users. Natalia on the other hand expressed that she does not believe discussions on Facebook or other social media platforms are healthy, because people tend to enter them with the idea to start an argument or a confrontation, but still, she thinks people should freely share their opinions on the platform. Along those lines Kaloyan said that even though seeing different opinions would sometimes 'piss him off', he would still rather see that, than only posts and articles that go in line with his opinions.

In addition to that, some interviewees pointed out that a big reason for them to use social media as a source of information is to get the 'unfiltered' information, coming straight from people and how they experience a given situation. For example, Eduard highlighted that he used Facebook during the pandemic, in order to get 'the real news', which he described as what is happening in his community and what people's experiences are, rather than what the journals and mass media are saying. Nikolay added more to that by saying:

Social media is like a secondary news, because you would hear the official news from the official outlets, but then you could engage with the users and hear what they what actual people think. Or if you're based in one country, and there is a big event happening, like it was in Italy, with the spread of the virus, you could actually go on social media and see pictures and comments from people actually living there compared to just hearing the story from a news company in your country. (Nikolay, May 2022)

Therefore, this shows that users might be using social media platforms for getting that extra layer of information, on top of official sources and reports. This goes in line with the fact that the Bulgarian society has very high collective consciousness and low trust in institutions, which leads them to seeking and believing people's experiences, rather than what officials are saying (Pehlivanova, 2009). In that aspect, however, importance should be placed on differentiating between opinions and 'facts' or information, as the harm could come from people being misled from taking a certain opinion as a given fact (Stewart, 2021). In that sense, it was pointed out that maybe opinion sharing could be limited in certain scenarios. Some interviewees expressed that there are certain instances in which the presence of different or alternative opinions might be harmful. For example, Natalia mentioned:

I feel like a lot of people on Facebook tend to present their opinion as a fact. And I think it could be said that, when you say okay, that's my opinion, then it's fine. Well, it really also depends on the topic actually. I think if it's some serious topic, then I would say it could be harmful to maybe present our opinion as a fact, because maybe someone would read that and think: "Oh, well, that must be the truth. You know? And it's actually not." So that truly depends on the topic. (Natalia, May, 2022)

After this, she further emphasized that she would often see articles from accredited scientific journals, where in the comments some people would argue against scientific facts, which in her opinion would do more harm than good. In this case, the presence of different opinions and

discussions might be harmful, as it could mislead people and lead them to challenge information that has been proven scientifically. This, however, leads to the challenge of distinguishing between what is an opinion and what is a fact, and also making the decision about what is indeed a fact. Along those lines, Martin said:

> *I think this is just a general problem in the world, like you don't know whether something is true or not until sometimes after the fact. (Martin, May 2022)*

This points to what can be described as one of the prevailing issues in the world of content moderation. For instance, Stewart, 2021, specifies that there are two disagreements related to the detection of misleading information online. The first one relates to distinguishing between what is an opinion, and what is information, thus what should be subject to moderation, and the second one questions who should decide that, and if it could be done in an unbiased manner (Stewart, 2021). The interviewees expressed their skepticism towards the involvement of the companies in the latter part. The results from this study show the importance for users in tackling those issues, as they believe in the importance of opinion sharing online, but also feel skeptical about who should make the decisions and if in the current state of affairs, freedom of speech online even exists. Those latter sentiments will be further described and analyzed in the next part of this chapter.

4.4. Social implications of content moderation

As already described in the theoretical framework, previous research has shown that users tend to be skeptical when it comes to the fairness of the way social media companies are doing content moderation (Kemp & Ekins, 2021). The interviews conducted in this research demonstrate similar reluctance, as users show distrust towards the reasoning platforms have behind conducting content moderation and hesitation about their power of deciding what is right or wrong. For instance:

> *….They obviously want to make money. Or that's how companies work, making profit to the stockholders so obviously, they'd be afraid to get rid of certain stuff, if*

that affects the shareholders…. the truth is just what makes that company shareholders the most money. (Kaloyan, May 2022)

In the same line of thought, Lora mentioned that content moderators are employees of a company who has investors that need to be pleased, and therefore it is impossible to moderate content without taking that aspect into account. Sandra also said that she doesn't know where platform interest lays, but she knows it is not in the public interest. She also added that misleading articles or so-called fake news, often have flashy titles and generate more engagement in terms of clicks and comments, which she thinks could be a discouraging factor for social media companies to remove these contents. In that sense she was doubtful if those platforms might be making profits from harmful content. Stefan, who said that he thinks governments are ultimately responsible for content moderation as they are the ones that put the obligation on platforms to do it, added another layer to this argument by saying:

I do think that the platform itself doesn't really care that much [about content moderation]. What they only care about is that the governments are allowing them to exist, in order for them to be able to make profit and business with advertisers. And governments are the body that allows a certain social media platform to operate in their country or not. (Stefan, May 2022)

This goes in line with the argument made by some scholars that social media platforms generally regulate their content in order to keep their users and advertisers, which are the main sources of income for these companies (Keller, 2018). This was demonstrated by the situation in which YouTube severely tightened its content moderation policies almost overnight, after major advertisers pulled their money out of the platform, as their brand advertisements appeared next to content that was deemed to be extremist and harmful (Solon, 2017). In addition to that, Roberts (2018) argues that content moderation is a matter of brand protection and social media companies undertake content moderation practices for the sole purpose of protecting themselves from liability, keeping investors happy, maintaining advertising revenue, and providing a decent user experience. Roberts (2018) blames this perception on the lack of transparency that social media companies have in terms of content moderation, by pointing out their unwillingness to disclose too

much information about how they approach the moderation practice. This goes back to the point that interviewees made about the need for more transparency in content moderation and in that sense, it demonstrates even further how increased transparency about the rules and practices of content moderation could lead to higher trust and less judgment on the process. In turn, interviewees also made some observations as to how this situation could be improved in their opinions, and how content moderation can be done in a way that is more transparent and in line with the user's interest. Stefan suggested that in order to escape the 'trap' he pointed out, there needed to be a more inclusive discussion on the topic:

> *I think it should be an open discussion for all the people in the world to discuss whether content moderation is needed to see both perspectives to hear both sides of the argument, because I'm sure that me as a regular person is missing out a lot of important points…. And then as a society, we should be able to make a decision whether this is something good or something bad, but it shouldn't be imposed on us by the government. (Stefan, May 2022)*

This emphasized the points made by this study and by Riedl et al. (2021) about the importance of not excluding users from the debate on content moderation. Nikolay also shared a similar opinion. He said that platforms should be the one making the decisions about content moderation, but they should be very open with users about the decisions they make, get opinions from users and have open dialogues, get feedback and reiterate. On the other hand, some participants suggested a different approach, where responsibility for content moderation is taken outside of the platform itself.

> *I would say there needs to be a third party. Something like an ombudsman, like someone who is.... I don't know who would be sponsoring them. But if it could be not attached to either government or company, that would be the best way of like, the sponsorship could come, but if it was not entailed with any other affiliations politically, or company, then something like that, so that the person is not influenced by anything political or financial aspects. (Lora, May 2022)*

Nikolay also shared that he thinks there should be a well-established third party who ultimately gives the thumbs up and thumbs down for certain pieces of content, while Sandra mentioned that she thinks content moderation should be a collaboration between the platform and an established third party. Those results support the findings of a report by Business Insider Intelligence, in which they find that 70% of users studied preferred a stakeholder outside of the company making the final decision regarding content moderation, with 26% of respondents stating that it should be a non-governmental independent agency (Schomer & McCarthy, 2019). Independent fact checking has existed in the United States since the early 2000s and has gained traction in Europe as well over the last decade with over 60% of those being fully independent civil organizations, and some being affiliated with media organizations (Graves & Cherubini, 2016). Despite their rise, their presence is still scarce, and their work hasn't gone without criticism, mainly due to the complexity and disagreements from users regarding the decisions made (Graves & Cherubini, 2016).

Independent fact checking could serve as a solution to concerns regarding the interests of platforms and how those influence content moderation practices, and would mitigate the skepticism of further government regulation, however there is another aspect of the issue that it cannot address yet - the human bias. Content moderation practices have often been blamed to be biased, with the personal bias of human moderators often weighing in (Diakopoulos & Naaman, 2011). The backlash often comes from people with more conservative ideologies, who claim that their views are being censored and that platforms are biased against them (Usher, 2018). Fact-checking organizations have been accused of the same bias, leading to conservatives pledging for 'fact-checking of the fact-checkers' (Richardson, 2018). Those statements were also confirmed by the sentiments expressed from the participants in this study. Eduard, who was the most doubtful participant of Covid information and of the Covid vaccine, was also the one who felt his opinion was 'censored' and content he shared was restricted the most.

> *They only allow you supportive posts, they never allow you to be critical of the EU and are always pro NATO, pro vaccines, pro everything. and they talk about free speech and free discussion, but they never let you say your opinions. (Eduard, May 2022).*

Eduard further shared that he often had his posts taken down during the pandemic, with no feedback or explanation from Facebook itself, and that social media was too influenced by the government and 'the West' given the content they are allowing on the platform. Other participants confirmed that even though they haven't felt to be subject to the 'bias', they also think platforms are often leaning in a certain direction. Natalia also said:

> *I do think that social media are very left leaning politically. I think that it's, you know, I don't think it's possible to talk about social media nowadays and not also mention politics with it, because it's so intertwined. (Natalia, May 2022)*

Martin also acknowledged that platforms tend to be more left leaning as San Francisco where most big social media companies are based generally has different views than other parts of the US. What he emphasized, however, is that as long as the rules are transparent, it is acceptable for some platforms to be more liberal, and for others to be more conservative, as different traditional media outlets are also usually leaning towards a particular political direction. Christina, on the other hand, mentioned that people are usually very biased, so the information you get on social media, no matter how moderated, will always be more biased than a traditional media source. As mentioned, human bias has been emphasized by academia, as a problem for content moderation, and it is one that is particularly hard to curb. Martin made a suggestion as to mitigating the bias of human based content moderation by automating the process:

> *I think it should be algorithm based so basically no one person should be scrolling through Facebook and saying: "Oh, this is allowed, this not allowed", because every person is biased in their own way……So I feel like there should be sort of like Community guidelines and I feel like it should be algorithms deciding based on rules that are clear. (Martin, May 2022)*

However, researchers have pointed out that algorithms often carry a certain amount of bias, since they are created and trained by humans, who inherently have their own bias. In that sense, many algorithms have been blamed in causing discrimination and being biased and creating an

algorithm that is as fair as possible has been a huge topic of discussion (Binns et al., 2017). Bias is a part of human nature and therefore it is a factor that is nearly impossible to eradicate. Therefore, possibly the goal of content moderation should not be to remove human bias, but to include as much as possible different perspectives and opinions, as to make the decision-making process as fair as possible. In that scenario the previous points about transparency, open discussions and reiteration would be essential to achieving this goal.

Another point that emerged in the discussion of content moderation was the impact of information and opinions shared online. The whole reason for the existence of content moderation in the first place is to remove harmful content, and information that can have a negative impact on the people that see it. When asked about the impact of information shared online, participants expressed that it depends on the topic and the circumstances. In regard to the Covid pandemic, interviewees felt that information shared online definitely had an impact on the outcomes and vaccine hesitancy in Bulgaria. Some participants even shared those online discussions and information initially had impact on their sentiments towards measures and Covid vaccines:

> *I got maybe a little bit influenced by the whole discussions and comments and opinions, because it was a very hot topic, also, on social media platforms, Facebook, as well. And a lot of people commenting, a lot of people without any medical background, commenting…. For me, it was confusing, I was confused (Sandra, May 2022)*

Christina also shared that this effect might have been more amplified than usual as everyone had to stay home and thus spent more time on social media than usual. However, participants shared that social media isn't the only one to blame for that. First of all, participants emphasized the importance of people seeking information outside of platforms, and that the platforms should not bear the responsibility for educating people on how to be careful with the information they receive. On that note Natalia mentioned:

> *Because at the end of the day, like social media can't give you all the information, to teach you, or Facebook or any other type of social media. It's not their*

responsibility to educate, but you know, they're not educational platforms, if you want to educate yourself, go somewhere else. (Natalia, May 2022)

In that sense, she said that what social media platforms could do to diversify the information a person receives, is to adjust the algorithms in a way that could expose people to more diverse information and opinions, instead of only the ones they interact with. Sandra also added that after her confusion, she realized that she needed to seek information outside of social media in order to inform herself based on research. Interviewees also noted other factors that could have had a more significant impact in terms of sentiments regarding the pandemic. They all pointed to mixed statements and guidance from the government and the WHO, as well as the overall uncertainty in the situation. They also placed some blame on media outlets, who used flashy titles in order to generate clicks and engagement, with the actual content of the article being different from the title. Therefore, interviewees agreed that the impact of information online was only a small part of the overall cycle. An interesting point that was made by a few of the participants was regarding the underlying social dynamics and how that could have also influenced how people perceive online information and the impact of the pandemic:

So paradoxical, but I think that overall, people in Bulgaria are, in the first place, very susceptible to conspiracy theories. And one small thought that is a conspiracy, sparks a big idea in them that it is a conspiracy, so it doesn't take a lot for them to believe things that are wrong or misinformation. (Nikolay, May 2022).

Researchers have previously linked the post-Communist transition in Eastern Europe with the rising power of conspiracy theories, mostly attributed to political dynamics (Marinov & Popova, 2022), and therefore can be considered an important social dynamic in Bulgaria and probably other Eastern European countries. This highlights the importance of the underlying social dynamics in a society, when discussing content moderation, which was pointed out by Wihbey et al. (2022) and reviewed earlier in this paper.

Last but not least, participants felt that the impact of online information was dependent on the topic and on the source of the information. Natalia for example emphasized that with topics

that can be considered important to the overall well-being of people, such as public health in the case of Covid, or politics and elections, then information shared online could have a bigger and more detrimental outcome. On the other hand, when it comes to entertainment, celebrity news or events, sharing opinions or information that could be misleading is not particularly harmful and would not impact people in a negative way. Similarly, Martin and Nikolay both emphasized that people often seek information about or talk about their interests and hobbies on social media, where incorrect information would not necessarily mislead people to a significant extent, and therefore the free flow should be allowed. Nikolay added another layer to that by saying that online information has a diverging impact depending on the topic - if the matter discussed is scientifically proven or factually backed, then information or opinion that question or oppose that should be removed from the platform as they would impact users in a negative way. However, if a topic carries a high degree of uncertainty- such as the topic of Covid, he felt that there should be a free flow of information and opinions, as that would generate an open discussion. Thus, the topic of discussion, the underlying social dynamics and the complexity of the situation should be taken into account when discussing the impact of online information.

5. Conclusion

Content moderation has proven over time to be a highly controversial and challenging topic. As social media has transformed the way in which we communicate and expedited the sharing of information, the issues related to misinformation, conspiracy theories and 'fake news' have gained more prominence than ever. Those dynamics were further changed by Covid, but also had a significant impact on the course of the whole pandemic (Magalhães & Katzenbach, 2020). For all those reasons, content moderation has been at the center of policy debates, legislative action and academic discourse. This dialogue, however, needs to take into account the user perspective, in order to account for underlying social dynamics, geographical and political factors, and diverging values (Gillespie et al., 2020; Riedl et al., 2021).

This research has explored the perceptions of users in Bulgaria regarding content moderation, in light of the widespread misinformation regarding the Covid pandemic. The findings of this study aim to uncover existing patterns within a society with a particular political background, social values and language barrier, which can be used as a basis of analysis of other countries with similar characteristics in order to reach a point of theoretical saturation. The insights gained from this study have provided the vehicle for answering the research question that lays at the core of this research: *How do users in Bulgaria perceive content moderation on social media in light of widespread misinformation about Covid vaccines?*

This study employed in-depth interviews as a method for the research and a total of ten participants, who are all active users of social media, were interviewed. Following the analysis, a total of three overarching topics were identified. Firstly, participants expressed their opinions towards content moderation in practice and shared their sentiments towards government regulation, platform liability, the role of the user and the importance of transparency. Overall, interviewees felt skeptical towards further regulation on content moderation, but also feared leaving the task solely to the platform. Contrary to what has been emphasized in previous research, they also highlighted the role of the user in being more careful with the information they see and share and how platforms shouldn't be judged too harshly, due to the complexity of the issue. Most importantly, they reiterated the importance of increasing transparency, in order to mitigate all of

those concerns. The second topic was regarding users' sentiments towards expression of opinions online, the importance of different opinions and discussions, and the separation between mere opinions and misinformation. Participants supported the notion of social media platforms being a public forum and underscored the importance of allowing the free flow of information and alternative opinions on the platforms, in order to spark productive discussions. They agreed that as in day-to-day life, freedom of expression online should not be absolute, but should also be protected. Lastly, participants discussed the social implications of content moderation, in terms of the commercial interests of platforms, the possible bias in moderation and the impact of information shared online. Participants showed their skepticism towards the motivations of social media platforms to moderate content and confirmed their sentiment of the existence of bias within those platforms. They also shared that the impact of information shared online shouldn't be attributed to just social media and their content moderation practices, and that the underlying political and social dynamics are equally as important.

5.1 Social and theoretical implications

This study has a few social and theoretical implications. In the first place it is important to note that the results of the research conducted are similar to what previous studies on the user perspective have found, and what researchers have pointed out and discussed in the past. Therefore, from a theoretical perspective, the results from study can be considered as a validation of understanding in current academic discourse (Sousa, 2014), but it contributes to it by adding the nuance of the user perception. As has been stated in this paper and by other researchers, there is a lack of consideration of the user perspective in current content moderation debates. The user perspective is essential, as it would allow us to understand how the different social dynamics, political backgrounds, values and context influence sentiments towards content moderation (Gillespie et al., 2020). This research gives insight into how some particular social dynamics in Bulgaria could influence the way users there feel about their expression online and about content moderation practices. Participants showed concern regarding increasing government regulation on content moderation and the government having too much power as to what can be said online, due to a socialist past, where speech was very restricted. In that sense, this study bridged the gap in current research and policy discourse as to how users in a country with an authoritarian past could

feel towards increasing regulation on social media and aims to pave the way for future research of user sentiments in other regions of the world. On the other hand, participants also expressed concerns regarding the motivations with which platforms themselves execute content moderations, and how as private entities, their interests are mostly commercial, and do not necessarily lie with the public interest. However, interviewees expressed that they believe platforms should not be judged too harshly about their content moderation practices, as it is a very controversial topic, and users shouldn't be stripped of any responsibility regarding the information they consume and share. This finding contributes to the current debate, which focuses almost entirely on the platform and government responsibility.

From a societal perspective, this study adds to the understanding of the relationship between users and platforms and the role social media platforms play in today's society. As previous research on the topic of content moderation has emphasized the importance of not restricting freedom of expression and censoring alternative opinions. This research has contributed to previous findings that conservative users, and users with more alternative opinions experience more restrictions and content removal on social media (Usher, 2018). This implies that there is the need for further discussion as to how to make the process as fair and unbiased as possible, in order not to marginalize or restrict certain groups. In this line of thought, this research also uncovered that users would like to be involved in the decision making process regarding content moderation, and that they insist on the overall process and rules being as transparent as possible. As previous research has also confirmed, transparency would increase trust in the process and would help avoid discrimination and silencing in terms of online content (Suzor et al., 2019). Lastly, this research contributes to the still scarce research about the user perspective of content moderation and aims to pave the way for future research into user perspectives, and for the inclusion of users in the debate regarding content moderation.

5.2. Limitations and implications for future research

Despite the measures undertaken to suppress them, there are several limitations to this research. First, it is important to note that the sample of ten people is a relatively small sample compared to the millions of users of social media platforms. Also, even though the focus on

Bulgaria and in the particular case of Covid allowed us to uncover underlying patterns, cultural characteristics and diverging circumstances might impede the generalizability of results and their translation to other cases. Furthermore, as the researcher used grounded theory and an inductive approach, and despite grounding them in existing literature, the established categories and themes that constituted the theoretical findings were subject to the interpretation of the researcher. This was mitigated by critical reflection, careful analysis of the data and double-checking of all the categories with the data in the interviews. Moreover, the analysis reached theoretical saturation, as at a particular point, the analysis of further interviews only confirmed existing observations and did not generate any new themes (Thornberg & Charmaz, 2014). Therefore, the results of this study have high reliability, and could be used as a useful insight to guide future research and discourse regarding content moderation but should not be generalized due to the size of the sample (Sousa, 2014). Another drawback that could result from interviews, especially as in this case the topic of Covid vaccines can be a rather polarizing topic, is that some participants might have not felt comfortable to share all of their opinions and experiences as to not feel judged.

This study has provided relevant insights which can be investigated further in future research. As the sample aimed to include people of different opinions regarding Covid vaccines, in order to get the perspectives of people with fundamentally different sentiments and consequently different online experiences, the results combined diverging perspectives and presented an extensive understanding of how users perceive content moderation and the content they see online. However, future research can focus on wider samples, in a more international context, in order to eliminate country specific characteristics. Bulgaria was a suitable example in order to discover patterns in a society with a different language than the predominantly used ones on social media and represents smaller Eastern European countries that are not usually the subject of thorough research in the context of content moderation. However, the legacy of socialism in Bulgaria has left a very strong collective consciousness related to communities and groups of people, leading people to have very high trust in micro-level personal relationships in contacts, than in government and institutions (Pehlivanova, 2009). Also, Bulgarian people generally have very low trust in government and governmental institutions (Eurofound, 2018). This cultural characteristic possibly has an effect on how Bulgarian users perceive their interactions online, which might differ from for example countries in Western Europe. Therefore, it would be useful to conduct further research

on the user perspective in different countries, or on a more international context, and compare the similarities and differences in the results and the context.

Moreover, future research could focus on different social and political phenomena than Covid, in order to see whether the perceptions of online content of users differ according to the topic discussed and according to their different opinions on various topics. This would help in gaining understanding about whether specific issues require more strict moderation according to users, but it would also shed a light into whether users with diverging opinions from what is 'the popular belief' are usually the ones who are subject to 'more moderation' than others.

Additional research could further explore some of the topics brought up in this research, such as the importance of transparency in content moderation and the possible inclusion of the user community in the decision-making process regarding content moderation. Current literature has focused mainly on 'protecting' the user from harmful content online but hasn't extensively discussed whether and how the user can become part of making the decision as to what they see online.

Last but not least, future research should take into consideration that the actors and challenges surrounding online content and misinformation is constantly shifting, and the actors and dilemmas are continuously changing. It is thus important to analyze the actions that need to be taken over time and the impact they would have on society as a whole.

References

Abercrombie, N., Hill, S., Turner, B.S. (1984). The Dictionary of Sociology. Penguin Books

Albright, J. (2017). Welcome to the era of fake news. Media & Communication, 5(2), 87–89. Doi: 10.17645/mac.v5i2.977

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Allen, M., & VandeHei, J. (2019). Trump allies plot new war on social media. Axios. https://www.axios.com/trump-2020-campaign-social-media-bias-41bbed1e-0bd3-445d-bf6f195b5c3e65a6.html

Babbie, E. (2014). The Basics of Social Research (6th ed.). Wadsworth: Cengage Learning

Balkin,J.M. (2017). Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation. Yale Law School, Public Law Research Paper No. 615. http://dx.doi.org/10.2139/ssrn.3038939

Baym, N.K., Boyd, D. (2012). Socially Mediated Publicness: An Introduction, Journal of Broadcasting & Electronic Media, 56:3, 320-329 https://doi.org/10.1080/08838151.2012.705200

Benkler, Y., Faris,R., Roberts, H., Zuckerman, E. (2017). Study: Breitbart led Right-wing Media Ecosystem Altered Broader Media Agenda. Columbia Journalism Review. http://www.cjr.org/analysis/breitbart-media-trump-harvard-study.php.

Benvenisti, E. (2018). Upholding Democracy Amid the Challenges of New Technology: What Role for the Law of Global Governance?, European Journal of International Law, Volume 29, Issue 1, Pages 9–82 https://doi.org/10.1093/ejil/chy013

Bishop, S. (2019). Managing visibility on YouTube through algorithmic gossip. New Media & Society, 21(11–12), 2589–2606.

Bittencourt, C.M. (2020). Social Media Companies and Harmful Online Content: responsibility, free speech and a possible regulatory framework. Tilburg Law School http://arno.uvt.nl/show.cgi?fid=153001

Boyce, C., Neale, P. (2006). Conducting in-depth interviews: A Guide for Designing and Conducting In-Depth Interviews for Evaluation Input. Monitoring and Evaluation - 2. Pathfinder International. http://www2.pathfinder.org/site/DocServer/m_e_tool_series_indepth_interviews.pdf?docID=6301

Boytchev, H. (2021). Covid-19: Why the Balkans' vaccine rollout lags behind most of Europe. BMJ 2021;375:n2412 http://dx.doi.org/10.1136/bmj.n2412

Braun, V., Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology. 3. 77-101. DOI: 10.1191/1478088706qp063oa

Brennen, B.S. (2017). Qualitative Research Methods for Media Studies (2nd ed.). Routledge. https://doi.org/10.4324/9781315435978

Bucher, T. (2012). Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. New Media & Society, 14(7), 1164–1180. https://doi.org/10.1177/1461444812440159

Burger, P., Kanhai, S., Pleijter, A., & Verberne, S. (2019). The reach of commercially motivated junk news on Facebook. PLOS ONE, 14(8). https://doi.org/10.1371/journal.pone.0220446

Chen, Z., & Berger, J. A. (2013). When, Why, and How Controversy Causes Conversation. Journal of Consumer Research, 40 (3), 580-593. http://dx.doi.org/10.1086/67146

Cinelli, M., Morales, G.D.F., Galeazzi, A., Quattrociocchi, W., Starini, M. (2021). The echo chamber effect on social media. PNAS, 118 (9) e2023301118. https://doi.org/10.1073/pnas.2023301118

Common, M.F., & Nielsen, R.K. (2021). Submission to UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression report on disinformation. Reuters Institute for the Study of Journalism, University of Oxford https://reutersinstitute.politics.ox.ac.uk/news/how-respond-disinformation-while-protecting-free-speech

Cook, C.L., Patel, A., Wohn, D.Y. (2021). Commercial Versus Volunteer: Comparing User Perceptions of Toxicity and Transparency in Content Moderation Across Social Media Platforms. Frontiers in Human Dynamics, Vol. 3, ISSN 2673-2726 https://doi.org/10.3389/fhumd.2021.626409

Cota, W., Ferreira, S.C., Pastor-Satorras, R., Starnini, M. (2019). Quantifying echo chamber effects in information spreading over political communication networks. EPJ Data Sci. 8, 35. https://doi.org/10.1140/epjds/s13688-019-0213-9

Cotter, K. (2021).'Shadowbanning is not a thing': Black box gaslighting and the power to independently know and credibly critique algorithms.Information, Communication & Society, Online First, 1–18. https://doi.org/10.1080/1369118X.2021.1994624

Crowe, S., Cresswell, K., Robertson, A. et al. (2011). The case study approach. BMC Med Res Methodol 11, 100. https://doi.org/10.1186/1471-2288-11-100

Diakopoulos, N., Naaman, M. (2011). Towards quality discourse in online news comments. In Proc. of CSCW. https://doi.org/10.1145/1958824.1958844

Dretske, F.I. (1983). Précis of Knowledge and the Flow of Information. *Behavioral and Brain Sciences*, 6:55-90 https://doi.org/10.1017/S0140525X00014631

Dredze, M., Broniatowski, D. A., & Hilyard, K. M. (2016). Zika vaccine misconceptions: A social media analysis. Vaccine, 34(30), 3441–3442. https://doi.org/10.1016/j.vaccine.2016.05.008

Duarte, N., Llanso, E., Loup, A. (2017). Mixed Messages? The Limits of Automated Social Media Content Analysis. Center for Democracy & Technology. https://cdt.org/wp-content/uploads/2017/11/Mixed-Messages-Paper.pdf

European Centre for Disease Prevention and Control (2022). Covid-19 vaccine tracker. https://vaccinetracker.ecdc.europa.eu/public/extensions/covid-19/vaccine-tracker.html#uptake-tab

Egelhofer, J., & Lecheler, S. (2019). Fake news as a two-dimensional phenomenon: A framework and research agenda. Annals of the International Communication Association, 43(2), 97–116. https://doi.org/10.1080/23808985.2019.1602782

Eurofound (2018). Societal change and trust in institutions https://www.eurofound.europa.eu/sites/default/files/ef_publication/field_ef_document/ef18036en.pdf

Felella, L.H. (2021). A Call for Legislated Transparency of Facebook's Content Moderation. Brennan Center for Justice. https://www.brennancenter.org/our-work/analysis-opinion/call-legislated-transparency-facebooks-content-moderation

Fetzer, J. H. (2004a). Information: Does It Have to Be True?. Minds and Machines, 14:223-29. https://doi.org/10.1023/B:MIND.0000021682.61365.56

Fetzer, J.H. (2004b). Disinformation: The Use of False Information. Minds and Machines, 14:231-40 https://doi.org/10.1023/B:MIND.0000021682.61365.56

Fallis, D. (2011). Floridi on disinformation. Etica & Politica, 13(2), 201–214 https://www.openstarts.units.it/bitstream/10077/5802/1/Fallis_E&P_XIII_2011_2.pdf

Fallis, D. (2015). What Is Disinformation? Library Trends 63(3), 401-426. doi:10.1353/lib.2015.0014.

Floridi, L. (2005). Semantic Conceptions of Information. The Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/entries/informationsemantic/

Floridi, L. (2011). The philosophy of information. Oxford Scholarship Online. DOI:10.1093/acprof:oso/9780199232383.001.0001

Gadde, V., Derella, M. (2020, March 16). "An update on our continuity strategy during COVID-19. *Twitter Blog.* https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuitystrategy-during-COVID-19.html

Gillespie, T. (2018). Custodians of the internet: Platforms, content moderation, and the hidden decisions that social media make. Georgetown Law Technology Review,22, 198–216.

Gillespie, T. (2020). Content moderation, AI, and the question of scale.Big Data & Society, 5, 371. https://doi.org/10.1177/2053951720943234

Gillespie, T., Aufderheide, P., Carmi, E., Gerrard, Y., Gorwa, R., Matamoros-Fernández, A., Roberts, S. T.,Sinnreich, A., & West, S. M. (2020). Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates.Internet Policy Review,9(4) https://doi.org/10.14763/2020.4.1512

Glaser, B.G., & Strauss, A.L. (1999). The Discovery of Grounded Theory: Strategies for Qualitative Research (1st ed.). Routledge. https://doi.org/10.4324/9780203793206

Gollust, S.E., Nagler, R.H., Fowler, E.F. (2020). The Emergence of COVID-19 in the US: A Public Health and Political Communication Crisis. J Health Polit Policy Law 45 (6): 967–981. https://doi.org/10.1215/03616878-8641506

Graves, L., & Cherubini, F. (2016). The Rise of Fact-Checking Sites in Europe. In Digital News Project Report (Reuters Institute Digital News Report). Reuters Institute for the Study of Journalism. https://ora.ox.ac.uk/objects/uuid:d55ef650-e351-4526-b942-6c9e00129ad7

Grice, P. (1989). Studies in the Way of Words. Cambridge: Harvard.

Habgood-Coote, J. (2019). Stop talking about fake news! Inquiry, 62(9–10), 1033–1065. doi:10.1080/0020174X.2018.1508363

Hammarberg, K., Kirkman, M., Lacey, S. (2016). Qualitative research methods: when to use them and how to judge them. Human Reproduction, Volume 31, Issue 3. https://doi.org/10.1093/humrep/dev334

Hartmann, I. A. (2020). A new framework for online content moderation. Computer Law & Security Review, 36, 105376. doi:10.1016/j.clsr.2019.105376

Helberger, N. (2020). The political power of platforms: How current attempts to regulate misinformation amplify opinion power. *Digital Journalism*, 86, 842–854. https://doi.org/10.1080/21670811.2020.1773888

Hicks, P. (2021). Press briefing: Online content moderation and internet shutdowns. UN Human Rights Office https://www.ohchr.org/sites/default/files/Documents/Press/Press_briefing_140721.pdf

Holan, A.D. (2016). 2016 Lie of the Year: Fake news. Politifact. The Poynter Institute. https://www.politifact.com/article/2016/dec/13/2016-lie-year-fake-news/

Hornsey, M. J., Harris, E. A., & Fielding, K. S. (2018). The psychological roots of anti-vaccination attitudes: A 24-nation investigation. Health Psychology, 37(4), 307–315. https://doi.org/10.1037/hea0000586

Hussain A., Ali S., Ahmed M., et al. (2018). The Anti-vaccination Movement: A Regression in Modern Medicine. Cureus 10(7): e2919. doi:10.7759/cureus.2919

Irving, E. (2019). Suppressing Atrocity Speech on Social Media' (2019) 113 AJIL Unbound 256. doi:10.1017/aju.2019.46

Jackson, B.F. (2014). Censorship And Freedom Of Expression In The Age Of Facebook. New Mexico Law Review https://digitalrepository.unm.edu/nmlr/vol44/iss1/6

Jackson, B., Jamieson, K.H. (2007). Unspun: Finding Facts in a World of Disinformation. *New York: Random House.* Print.

Jhaver, S., Bruckman, A., & Gilbert, E. (2019). Does transparency in moderation really matter? User behavior after content removal explanations on Reddit. Proceedings of the ACM on Human-Computer Interaction, 3CSCW,1–27. Medium. https://medium.com/acm-cscw/does-transparency-in-moderation-really-matter-b86bab9b4810

Kantchev, G. (2021). Covid-19 Vaccine Rollout Falters in Bulgaria Amid ‘Perfect Storm’ of Mistrust, Fake News. The Wall Street Journal. https://www.wsj.com/articles/covid-19-vaccine-rollout-falters-in-bulgaria-amid-perfect-storm-of-mistrust-fake-news-11632133548

Keller, D. (2018). Internet Platforms: Observations on Speech, Danger, and Money. Hoover Institution's Aegis Paper Series, No. 1807, 2018. Available at SSRN: https://ssrn.com/abstract=3262936

Keller, D. (2019). Who Do You Sue? State and Platform Hybrid Power Over Online Speech. Hoover Institution's Aegis Series Paper No. 1902, 2019 https://www.hoover.org/research/who-do-you-sue

Kemp, D., Ekins, E. (2021). Poll: 75% Don't Trust Social Media to Make Fair Content Moderation Decisions, 60% Want More Control over Posts They See. Cato Institute. https://www.cato.org/survey-reports/poll-75-dont-trust-social-media-make-fair-content-moderation-decisions-60-want-more#introduction

Klonick, K. (2018). The New Governors: The People, Rules, and Processes Governing Online Speech, Harvard Law Review, Vol. 131:1598 https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech/

Kim, H.K., Tandoc, E.C.Jr. (2022). Consequences of Online Misinformation on COVID-19: Two Potential Pathways and Disparity by eHealth Literacy. Front. Psychol. 13:783909. doi: 10.3389/fpsyg.2022.78390

Krouwel M. , Jolly K. , Greenfield S. (2019). Comparing Skype (video calling) and in-person qualitative interview modes in a study of people with irritable bowel syndrome—an exploratory comparative analysis. BMC Med Res Methodol 19, 219. https://doi.org/10.1186/s12874-019-0867-9

Leerssen, P. (2015). Cut Out By The Middle Man: The Free Speech Implications Of Social Network Blocking and Banning In The EU. Journal of Intellectual Property, Information Technology and Electronic Commerce Law, 6(2), 99-119. https://www.jipitec.eu/issues/jipitec6-2-2015/4271

Leetaru, K. (2017). Is Social Media Really A Public Space?. Forbes. https://www.forbes.com/sites/kalevleetaru/2017/08/01/is-social-media-really-a-public-space/

Llansó, E., Van Hoboken, J., Leerssen, P., Harambam, J. (2020). Artificial Intelligence, Content Moderation, and Freedom of Expression. https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/05/Artificial_Intelligence_TWG_Llanso_Feb_2020.pdf

MacDonald N.E., et al. (2015). Vaccine hesitancy: Definition, scope and determinants. Vaccine. 2015; 33 (34):4161–4164. https://doi.org/10.1016/j.vaccine.2015.04.036 PMID: 25896383

Maciuszek J., Polak M., Stasiuk K., Doliński D. (2021). Active pro-vaccine and anti-vaccine groups: Their group identities and attitudes toward science. PLoS ONE 16(12): e0261648. https://doi.org/10.1371/journal.pone.0261648

Magalhaes, J.C., Katzenbach, C. (2020). Coronavirus and the frailness of platform governance. Internet Policy Review. https://policyreview.info/articles/news/coronavirus-and-frailness-platform-governance/1458

Marinov, N., & Popova, M. (2022). Will the Real Conspiracy Please Stand Up: Sources of Post-Communist Democratic Failure. Perspectives on Politics, 20(1), 222-236. doi:10.1017/S1537592721001973

Mayring, P. (2014). Qualitative content analysis: theoretical foundation, basic procedures and software solution. Leibniz Institute for Social Sciences. URN: http://nbn-resolving.de/urn:nbn:de:0168-ssoar-395173

Meta. (2022). Our Actions. https://about.facebook.com/actions/

Mina, A.X. (2015). From Digital Divide to Language Divide: Language Inclusion for Asia's Next Billion, in The Good Life in Asia's Digital 21st Century. *Medium.* https://medium.com/meedan-labs/from-digital-divide-to-language-divide-language-inclusion-for-asia-s-next-billion-7792db117844.

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. New Media & Society, 20(11), 4366–4383. https://doi.org/10.1177/1461444818773059

Musk, E. (2022, March 26). Given that Twitter serves as the de facto public town square, failing to adhere to free speech principles fundamentally undermines democracy. [Tweet]. https://twitter.com/elonmusk/status/1507777261654605828

Nielsen, R.K., Graves, L. (2017). "News you don't believe": Audience perspectives on fake news. Reuters Institute. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2017-10/Nielsen%26Graves_factsheet_1710v3_FINAL_download.pdf

Nieuwenhuis, A. (2000). Freedom of Speech: USA vs Germany and Europe. Netherlands Quarterly of Human Rights, 18(2), 195–214. https://doi.org/10.1177/092405190001800203

Nunziato, D.C. (2018). From Town Square To Twittersphere: The Public Forum Doctrine Goes Digital. 2018-40 SSRN Electronic Journal; P. R Biju, Political Internet: State And Politics In The Age Of Social Media. Routledge 2017. http://dx.doi.org/10.2139/ssrn.3249489

Oliva, T.D. (2020). Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression. Human Rights Law Review, 2020, 20, 607–640 doi: 10.1093/hrlr/ngaa032

Oyeyemi, S. O., Gabarron, E., and Wynn, R. (2014). Ebola, twitter, and misinformation: a dangerous combination? BMJ 349:g6178. https://doi.org/10.1136/bmj.g6178

Patton, M. Q. (2002). Qualitative research and evaluation methods. Thousand Oaks, Calif: Sage Publications. https://us.sagepub.com/en-us/nam/qualitative-research-evaluation-methods/book232962

Pehlivanova, P. (2009). The Decline of Trust in Post-communist Societies: The Case of Bulgaria and Russia. https://www.ceeol.com/search/article-detail?id=99871

Renard, C. C.(2020). Algorithmic Content Moderation on Social Media in EU Law: Illusion of Perfect Enforcement. University of Illinois Journal of Law, Technology & Policy (JLTP), Forthcoming. http://dx.doi.org/10.2139/ssrn.3535107

Richardson, V. (2018). Conservative project seeks to fact-check the fact-checkers accused of liberal bias. The Washington Times. https://www.washingtontimes.com/news/2018/mar/27/conservative-project-seeks-fact-check-fact-checker/

Riedl, M.J., Whipple, K.N., & Wallace, R. (2021). Antecedents of support for social media content moderation and platform regulation: the role of presumed effects on self and others. Information, Communication & Society. https://doi.org/10.1080/1369118X.2021.1874040

Roberts, S. T. (2016). Digital refuse: Canadian garbage, commercial content moderation and the global circulation of social media's waste. Media Studies Publications.14(1), 1–12 ons. https://ir.lib.uwo.ca/commpub/14

Roberts, S. T. (2018). Digital detritus: 'Error' and the logic of opacity in social media content moderation. First Monday, 23(3). https://doi.org/10.5210/fm.v23i3.8283

Rodgers, K., and Massac, N. (2020). Misinformation: a threat to the public's health and the public health system. J. Public Health Manag. Pract. 26, 294–296. doi: 10.1097/PHH.0000000000001163

Ruckenstein, M., Turunen, L.L.M. (2019). Re-humanizing the platform: Content moderators and the logic of care. New media & Society 2020, Vol. 22(6) 1026–1042. https://doi.org/10.1177%2F1461444819875990

Satariano, A. (2019). Europe Is Reining In Tech Giants. But Some Say It's Going Too Far. The New York Times. https://www.nytimes.com/2019/05/06/technology/europe-tech-censorship.html

Seawright, J., & Gerring, J. (2008). Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options. Political Research Quarterly, 61(2), 294–308. https://doi.org/10.1177/1065912907313077

Scarantino, A., Piccinini, G. (2010). Information without Truth. *Metaphilosophy*, 41:313-30. https://doi.org/10.1111/j.1467-9973.2010.01632.x

Schmidt AL, Zollo F, Scala A, Betsch C, Quattrociocchi W. Polarization of the vaccination debate on Facebook. Vaccine. 2018; 36(25):3606–3612. PMID: 29773322. https://doi.org/10.1016/j.vaccine.2018.05.040

Schomer, A., McCarthy, D. (2019). The Content Moderation Report: Social platforms are facing a massive content crisis — here's why we think regulation is coming and what it will look like. Business Insider Intelligence. https://www.emarketer.com/content/our-data-suggests-social-platform-users-are-skeptical-of-platforms-content-moderation-abilities-heres-why-the-era-of-self-regulation-might-be-over-2019-9

Siggelkow, N. (2007). Persuasion with Case Studies. The Academy of Management Journal, 50, 20-24. https://doi.org/10.5465/amj.2007.24160882

Silver, C., & Lewins, A. (2014). Using software in qualitative research: A step-by-step guide. Sage. https://dx.doi.org/10.4135/9781473906907

Silverman, C., Singer-Vine, J. (2018). Here's Who's Been Blocked By Twitter's Country-Specific Censorship Program. Buzzfeed https://www.buzzfeednews.com/article/craigsilverman/country-withheld-twitter-accounts

Statista (2021). Distribution of Facebook users in Bulgaria as of July 2021, by age group. https://www.statista.com/statistics/805460/facebook-users-bulgaria/

Stewart E. (2021). Detecting Fake News: Two Problems for Content Moderation. Philosophy & technology, 34(4), 923–940. https://doi.org/10.1007/s13347-021-00442-x

Søe, S. O. (2021). A unified account of information, misinformation, and disinformation. Synthese: An International Journal for Epistemology, Methodology and Philosophy of Science, 198(6), 5929. https://doi.org/10.1007/s11229-019-02444-x

Solon, O. (2017). Google's bad week: YouTube loses millions as advertising row reaches US. The Guardian. https://www.theguardian.com/technology/2017/mar/25/google-youtube-advertising-extremist-content-att-verizon

Sousa, D. (2014). Validation in qualitative research: General aspects and specificities of the descriptive phenomenological method. Qualitative Research in Psychology, 11(2), 211–227. https://doi.org/10.1080/14780887.2013.853855

Suzor, N. P., West, S. M., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. International Journal of Communication,13(2019), 1526–1543. https://ijoc.org/index.php/ijoc/article/view/9736/2610

Talisse, Robert B. 2018. “There’s No Such Thing as Fake News (And That’s Bad News).” 3: AM Magazine. https://www.3ammagazine.com/3am/theres-no-such-thing-as-fakenews-and-thats-bad-news/.

Tarran, B. (2017). Why facts are not enough in the fight against fake news. Significance, 14(5), 6–7. https://doi.org/10.1111/j.1740-9713.2017.01066.x

Thomas, D. R. (2006). A General Inductive Approach for Analyzing Qualitative Evaluation Data. American Journal of Evaluation, 27(2), 237–246. https://doi.org/10.1177/1098214005283748

Thomas, Z. (2020). WHO Says Fake Coronavirus Claims Causing 'Infodemic'. BBC [Online]. Available online at: https://bbc.in/2xUCaAh

Thornberg, R., & Charmaz, K. (2014). Grounded theory and theoretical coding. In The SAGE Handbook of Qualitative Data Analysis (pp. 153–169). https://dx.doi.org/10.4135/9781446282243.n11

Transparency International (2021). Corruption Perceptions Index. https://www.transparency.org/en/cpi/2021/index/bgr

Tsolova, T. (2022). Anti-vaccine protesters try to storm Bulgaria's parliament. Reuters. https://www.reuters.com/world/europe/anti-vaccine-protesters-try-storm-bulgarias-parliament-2022-01-12/

Twitter. (2022). Twitter Transparency Report 19. https://transparency.twitter.com/

Usher, N. (2018). How republicans trick facebook and twitter with claims of bias. The Washington Post. https://www.washingtonpost.com/news/posteverything/wp/2018/08/01/how-republicans-trick-facebook-and-twitter-with-claims-of-bias/

Vicario, M.D., Quattrociocchi, W., Scala, A., Zollo, F. (2019). Polarization and fake news: Early warning of potential misinformation targets. ACM Trans. Web 13, 1–22. https://doi.org/10.1145/3316809

Wardle, C. (2019). Misinformation has created a new world disorder. Retrieved from https://www.scientificamerican.com/article/misinformation-has-created-a-new-worlddisorder/

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. https://edoc.coe.int/en/media/7495-

information-disorder-toward-an-interdisciplinary-framework-for-research-and-policymaking.html

Wehba, H.B. (2019). Global Platform Governance: Private Power in the Shadow of the State. SMU Law Review, 27. https://scholar.smu.edu/smulr/vol72/iss1/9

West, S. M. (2018). Censored, suspended, shadow banned: user interpretations of content moderation on social media platforms. New Media Soc. 20, 4366–4383. https://doi.org/10.1177/1461444818773059

Wihbey, J., Chung, M., Peacey, M., Morrow, G., Tian, Y., Vitacco, L., Reyes, D. R., Clavijo, M. (2022). Divergent Global Views on Social Media, Free Speech, and Platform Regulation: Findings from the United Kingdom, South Korea, Mexico, and the United States http://dx.doi.org/10.2139/ssrn.3999454

Witt, A., Suzor, N., & Huggins, A. (2019). The rule of law on Instagram: An evaluation of the moderation of images depicting women's bodies. The University of New South Wales Law Journal,422, 557–596. https://eprints.qut.edu.au/129978/

Yin, R.K. (2009). Case study research, design and method. Sage Publications Ltd., 4th Edition https://doi.org/10.33524/cjar.v14i1.73

Zeng, J., Kaye, B.V. (2022). From content moderation to visibility moderation: A case study of platform governance on TikTok. *Policy & Internet, 14:79–95* https://doi.org/10.1002/poi3.287

Appendix A - Interview guide

Basic introduction:

- Mention the informed consent and ask them if they agree to the interview being recorded; remind them about the anonymity and ensured privacy of their responses
- Icebreaker, questions/conversation to build rapport and make the interviewee comfortable and predisposed to talk
- Asking how they have been doing
- Tell them a little more about what the research is about and explain why I chose to speak with them - because of their active participation on Facebook and particular Facebook groups regarding Covid vaccines

Guiding questions (subject to change depending on interview directionality):

- Do you enjoy the platform / social media in general? How often do you use it? What do you use it for? How often do you post?
- Do you use Facebook / social media to obtain news or relevant information?
- Did you use the platform to obtain news or information regarding Covid measures, vaccines, remedies etc. ?
- How did you perceive that information - did you find it useful, deceiving etc.?
- What was the most interesting / surprising / captivating piece of information you read on Facebook regarding Covid and Covid vaccines? What did you think about it/
- How careful are you with what you post? Do you do a double-check of the information you share?
- Have you had a piece of content they shared removed? What was it and why do you think that is?
- How do you double check or verify information you find on Facebook?
- Do you know what the term content moderation means and how do you understand it? The researcher can give a brief unbiased definition and overview on the topic in case the participant doesn't understand completely the concept of content moderation

- Who do you think is responsible for content moderation on Facebook and social media platforms?
- Who do you think should do it?
- Should the government have a say and establish guidelines and laws for that?
- Do you think content on social media should be moderated? Should misinformation be removed or should it be up to the user to check info they see online?
- Do you think platforms can do a good job in removing misinformation?
- Do you feel like you have the right to say whatever you want on social media? Would you feel censored if an opinion you shared was deleted or reported?
- What in your opinion is freedom of speech online? Is content moderation restricting it?

Stance towards Covid/ Covid vaccines:

- What is your stance on Covid vaccines?
- Could you see information on Facebook that supported your opinion towards Covid vaccines?
- Could you see all types of opinions regarding Covid vaccines on Facebook and do you think that was fair?
- Do you think there was a lot of misinformation on Facebook during Covid?
- Did this impact vaccination rates in Bulgaria in your opinion?

Final remarks and interviewees asked if they have anything else they would like to add

Appendix B - Sample description

Name	Age	Activity on Facebook	Covid vaccine stance
1. Hristo	20's	Moderately	Pro-vaccine
2. Martin	30's	Moderately	Pro-vaccine
3. Eduard	50's	Very active on groups	Vaccine hesitant
4. Christina	20's	Moderately	Pro-vaccine (changed opinions)
5. Nikolay	20's	Very active	Vaccine hesitant
6. Lora	20's	Very active	Pro-vaccine
7. Natalia	30's	Moderately	Pro-vaccine
8. Sandra	40's	Very active	Pro-vaccine (changed opinions)
9. Kaloyan	30's	Moderately	Pro-vaccine
10. Stefan	40's	Very active	Vaccine hesitant