Master Thesis

[Ambiguity, Autonomy & Arbitrariness: The European Vision of Freedom of Expression in the Digital Sphere]

by

Julian Alexander Samol

S2639882

j.a.samol@student.utwente.nl

Submitted in partial fulfillment of the requirements for the degree of Master of Science, program [Public Administration / European Studies], University of Twente

5th of July, 2024

Supervisors:

Dr. Ringo Ossewaarde, BMS

Dr. Elifcan Karacan, BMS

Abstract

Prior to the Digital Services Act, the digital sphere in the European Union was largely unregulated. Here, the nature of the digital sphere enables a rampant dissemination of information, which on one hand provides individuals with useful information allowing them to make informed decisions in their daily-life. On the other hand, one observes abuse by nefarious actors who spread hatred, dangerous ideologies and lies, ultimately aiming to cause serious harm. A turning of the tides was initiated with the passing of the Digital Services Act, which tasked service providers with realising the European Union's vision of the digital sphere. By laying out several obligations which ought to be met, the European Union determined what content is deemed permissible in the digital sphere. With this incision on the part of the European Union, a question arises of what strategy to combating abusive expressions is provided, including if this strategy infringes on fundamental rights or if it suitably addresses the seeming lawlessness in the digital sphere. Moreover, it begs the question of how issues such as disinformation and hate speech are interpreted, as well as where the European Union draws the boundaries to the right of freedom of expression. In order to answer these questions, an interpretive content analysis was conducted. Documents published by various institutions under the umbrella of the European Union addressing dynamics related to freedom of expression in the digital sphere have been analysed. With this, it was aimed to paint a picture of how the European Union interprets freedom of expression in the digital sphere, which required the interpretation of coded passages in the selected documents. As a prerequisite for this approach, a theoretical framework was established which investigated different conceptions of freedom of expression and looked at different dynamics which scholars raise in regard to freedom of expression in the digital sphere. The analysis found that the European Union’s vision, as laid out in the selection of documents, appears incomplete. While it was expected to gain findings that allow the painting of a picture of what the European vision entails, it was found that due to a combination of ambiguity, autonomy, and arbitrariness in both what actually ought to be tackled and how it ought to be tackled, service providers are able to arbitrarily interpret what the European vision entails. Thereby, it was concluded that the actual European vision remains largely vague, bearing potential for conflict within the digital sphere and between service providers and the European Union.

Abbreviations

Artificial Intelligence	AI
Artificial Intelligence Act	AIA
Charter of Fundamental Rights of the European Union	CFR
Code of Conduct	COC
Code of Practice	COP
Digital Services Act	DSA
E-Commerce Directive	ECD
European Union	EU
Freedom of Expression	FoE
Interpretive Content Analysis	ICA
Regulation on Addressing the Dissemination of Terrorist Content Online	RADTC
Universal Declaration of Human Rights	UDHR
Very Large Online Platform	VLOP
Very Large Search Engine	VLSE

Table of Contents

1 Introduction

1.1 Background

As the digital sphere progressively swallows up an increasing amount of human day-to-day activities, social media platforms came to be regarded as a substitute for the local town square, serving as a venue for the exchange of ideas and opinions. A prominent pioneer of this idea is the richest man alive, Elon Musk. In 2022, Musk acquired the social media platform Twitter for a sum of $44 Billion (Browne, 2022). He motivated this gargantuan purchase with his high regard for free speech, calling it the "bedrock of a functioning democracy", whereas according to him, Twitter poses as "the digital town square where matters vital to the future of humanity are debated" (PR Newswire, 2022). For Musk, and like-minded "free-speech absolutists" (Milmo, 2022) promising changes were planned, however, as Twitter, which was renamed to X, is widely used by Europeans and accessible in the European Market, it must adhere to European law. In the past, the digital sphere was mostly unregulated in regard to the dissemination of content, whereas in the European Union (EU), the E-Commerce Directive (ECD) posed as the only major legislation applicable. The ECD however did not obligate the monitoring of user's content and removed any liability from service providers as long as they were not aware of any illegal activities on their platform and quickly reacted to any related notification (Heldt, 2022). Musk's takeover provoked a wide-ranging commotion, including speculation about whether the technology tycoon's intentions were merely altruistic.

Coincidentally, the take-over occurred during the formative process of the European Union's Digital Services Act (DSA), whereby the EU immediately confronted Musk and his newly acquired platform by reminding about the DSA's obligations (Datta & D'Silva, 2022). Strikingly, the European Commission followed through by opening formal infringement proceedings against the platform (European Commission, 2023). This chain of events illustrates a clash of conceptions surrounding how the digital sphere ought to be regulated. On the surface level, it seems as if the EU introduced a regulation that contradicts a vision as communicated by Musk, thereby interfering in the attempt to create a safe haven for freedom of expression (FoE) and the exchange of ideas. Simultaneously, the EU's determination to implement the DSA raises questions concerning the feasibility of Musk's vision as well as what potential challenges lurk in its shadows, especially as the EU deems interference as necessary. This conflict ultimately calls for research into how the DSA aims to transform the digital sphere, particularly with regard to how freedom of expression is interpreted by the EU.

1.2 The Digital Services Act

On the fourth of October 2022 the EU passed the DSA, which lays its focus on regulating the digital sphere with regard to fundamental rights, data privacy, protecting stakeholders as well as promoting European digital sovereignty (Turillazzi et al., 2023). The DSA specifically addresses "services that involve the transmission and storage of user-generated content" (Wilman, 2022, p.1). With the act, the EU hopes to transform the digital sphere into a "safe, predictable and trusted online environment that facilitates innovation, in which the fundamental rights enshrined in the [Charter of Fundamental Rights of the European Union] are effectively protected." (Wilman, 2022, p.2). Moreover, the policy's main goals are also summarised as reducing illegal or potentially harmful content, allocating liability to service providers that host third-party content, protecting fundamental rights in the digital sphere and bridging information asymmetries between the service providers and their users (Turillazzi et al., 2023). Based on these aims, it already becomes evident how freedom of expression is of tremendous relevance in the context of the DSA. Not only by aiming to protect fundamental rights but also by influencing what content and therefore what expressions are permissible in the digital sphere. To achieve these aims, the relatively young EU legislation provides online platforms with several obligations. It also introduces the classification of very large online platforms (VLOP) for some intermediary service providers, whereas a platform is considered a VLOP when it has a user base of (over) 45 million.

The act goes on to define intermediary service providers as service providers offering network infrastructure, whereas next to the aforementioned categories, the DSA also refers to hosting providers such as cloud services and (smaller) online platforms (Turillazzi et al., 2023). Notably, VLOPs and very large search engines (VLSE) have to abide by all DSA obligations compared to other service providers who only have to follow distinct obligations (Wilman, 2022), here, the obligations of the DSA are meant to be proportional to the size of the impact a given service has on the European market (Turillazzi et al., 2023). Prominent examples of very large online platforms include Facebook, Twitter, but also Wikipedia, whereas Bing and Google are categorised as VLSEs (Hohmann & Kelemen, 2023). The main obligations of the policy are threefold, firstly the DSA establishes a new liability system (Article 6-8), secondly a removal order (Article 9) and thirdly a notice-and-action mechanism (Article 16) (Sulmicelli, 2023). From the latter emerges that providers are required to implement signalling mechanisms allowing anybody to flag content as being potentially illegal. Here, a given provider is then required to investigate the flagged posting and only after, the provider begins to be held liable for a given posting (Turillazzi et al., 2023). This mechanism applies to both content prohibited by law and the formulated terms and conditions of a provider (Heldt, 2022). If a provider in fact suspects that a (possible) threat to the life or safety of a person or criminal offence is at hand, the provider has to notify the corresponding authority depending on the involved member state (Hohmann & Kelemen, 2023). For this and the monitoring of the enforcement and respective compliance of the DSA, two new oversight institutions are established, these being Digital Services Coordinators at the national level, and the Board for Digital Services at the EU level. For this, Article 20 and 21 foresee that whatever sanction service providers undertake, it must be governed by clear and foreseeable rules, moreover, affected users must then be notified and have their applied sanction explained, whereas an ability for an appeal must be provided to users (Heldt, 2022).

Further obligations include the disclosure of how a provider moderates content, user's rights and the publishing of transparency reports on how content is moderated, which however does not apply to small or micro enterprises. VLOPs and VLSEs also need to conduct a risk assessment for their services. Notably, those providers are obligated to create a crisis-response mechanism, which should become active in extraordinary circumstances leading to a serious threat to security or health in the EU or significant parts of it. Here, the European Commission may require service providers to act in accordance with the European Commission. Regarding Artificial Intelligence (AI), the DSA obligates a low error ratio when service providers opt to utilise AI. Moreover, whenever AI is eventually utilised, any decision ought to be subject to human oversight (Hohmann & Kelemen, 2023). Finally, the DSA requires the publication of a yearly audit concerning DSA compliance, including the creation of an authority which supervises the enforcement of the DSA. Given this overview of the DSA aims and instruments, a question arises of how this vision ultimately asserts the boundaries of freedom of expression. In an effort to find an answer to this question, it is sensible to investigate contemporary research in the field, whereby potential knowledge gaps can further be identified.

1.3 State of the Art

The previous section established a broad summary of the DSA and its instruments, what follows is an overview of the current state-of-the-art discussing the DSA in terms of its goals and potential consequences. Academic discourse surrounding the DSA provides both promising and concerning findings with regard to FoE in the digital sphere. On a more general note, Leerssen's analysis showed that the DSA is the first piece of legislation to directly address shadow-banning as well as expanding on content moderation practices by addressing demonetisation and visibility restrictions (Leerssen, 2023). Hohmann & Kelemen identified how the DSA aims to mitigate issues relating to social media platforms acting as gatekeepers. Specifically, it is stressed how in a data-driven and information-dependent society, service providers are capable of arbitrarily affecting political discourse through the deplatforming of politicians (Hohmann & Kelemen, 2023). Furthermore, research identified that legislation aimed at combating hate speech as present in China, can be used to suppress free speech, as well as motivate platforms to create a general ban on politically sensitive content when legislation creates administrative liabilities for these platforms (Chen, 2022). Here, Turilazzi and peers argue that the change to the DSA's regulatory regime runs the risk of encouraging affected service

providers to implement a "delete first, think later" approach, enabling the infringement of user's rights. It is further argued how the risk of over removal is exacerbated by the DSA's removal clause (Turillazzi et al., 2023). Another scholar in Sulmicelli, agrees and enumerates how based on Article 16's notice and action mechanism, service providers are faced with time constraints combined with a threat of financial punishment. As a result, platforms are incentivised to over-restrictively moderate content and make use of algorithms in hopes of avoiding liability, whereby AI would be utilised to preemptively identify risky content prior to being alerted (Sulmicelli, 2023). Turillazzi and peers also argue that an obligation to follow the DSA may shy away platforms from entering or remaining in the European market. More critically, however, it is argued how the DSA fails to clearly discern between harmful and illegal content, raising concerns of potential FoE infringements (Turillazzi et al., 2023).

The findings provided until now paint a picture of the DSA posing as a dire risk to freedom of expression. Conversely, contemporary researchers also provide findings which contradict this picture. Hohmann & Kelemen describe the DSA as a milestone in European digital constitutionalism, "characterised as a set of rules shielding individuals from abuse of power in the digital environment" (Hohmann & Kelemen, 2023, p.226). Research by Paige, which attempted to clarify whether the DSA is compliant with FoE, concluded:

> "that some of the structures of the DSA restrict online expression (…). However, as understood by relevant legal authorities in Europe, the freedom of expression likely remains unviolated due to ever-expansive criteria by which authorities may limit that freedom." (Paige, 2023, p.1).

From these findings, one may infer that in spite of asserting boundaries to FoE in the digital sphere, the DSA does not violate the right as laid out in the Charter of Fundamental Rights of the European Union (CFR). By having identified two opposing views in the state-of-the-art, accompanied by the concerns of over-regulation raised before, it is called into question which of the two interpretations bares more truth. Moreover, further analysis into how FoE and its boundaries are actually interpreted within the DSA is motivated. In doing so, it becomes necessary to investigate forms of expressions that ought to be subject to interference. Here, Sulmicelli explains that by attempting to regulate content moderation on online platforms, the DSA tries to balance the challenges of combating abusive expressions with safeguards for FoE (Sulmicelli, 2023). Therefore, for both service providers and users, the question of what content and therefore what types of expression are deemed to be (un)protected under FoE arises. Related to this bipolarity of expressions' protection in the digital sphere, Heldt discusses that initially, social media platforms were regarded as enabling free speech and therefore facilitating a democratisation of public discourse (Heldt, 2022), mirroring the sentiment of Musk's remarks concerning his Twitter purchase. Over time however, as a result of AI content recommendation, problems with illegal and harmful content arose, which remain to be tackled by service providers who neglect the spread of mis- and disinformation, as well as criminal and harmful speech. This ultimately led to member states beginning to experience adverse effects of online speech harms, resulting in the realisation for policymakers that the current regime is insufficient. From this diagnosis emerged a call by the President of the European Commission von der Leyen, demanding that issues such as disinformation and online hate messages ought to be addressed (Heldt, 2022).

1.4 Knowledge Gap

Based on the summary of the provided research on the DSA, a knowledge gap can be identified. The current state-of-the-art is, firstly, characterised by research explaining the function of DSA mechanisms, providing a general account of what the act aims to achieve, exemplified by Wilman's work (Wilman, 2022). Secondly, research evaluates the DSA against the backdrop of its formulated aims, including in terms of its potential impact on FoE. Here, scholars either found how, for instance, the DSA can lead to disproportionate content moderation, undermining rights of minority groups such as the LGBTQ-community (Sulmicelli, 2023), or contrarily, how the DSA does not actually undermine the right. Here, scholars such as Paige (Paige, 2023), or Hohmann & Kelemen (Hohmann & Kelemen, 2023) find the DSA to positively contribute to safeguarding FoE in the digital sphere.

With this, key issues in contemporary research revolve around simplifying, or summarising what is laid down in the DSA, as well as scrutinising its mechanisms against the backdrop of potential infringements of FoE. A particular gap is identifiable in regard to an analysis of how freedom of expression is actually interpreted in the act, as in how the DSA itself portrays the right. While a debate persists surrounding the question if the act provides mechanisms that may undermine the right, an actual analysis of how the act explains and conceptualises freedom of expression (in the digital sphere), thereby, a crucial first step to this debate, is missing. Moreover, attributable to the DSA's implementation still being in its infancy, the quantity of research on it is generally limited, whereby research with the focus presented in this thesis is entirely missing. When scholars debate whether FoE is infringed upon, each scholar arguably holds a unique perspective of what expressions ought to be protected under the right, even when consulting its definition in the CFR or even case law. The necessity of the missing findings is further underlined by the contradictory nature of the findings found in contemporary research on how the DSA affects FoE, suggesting how a prior step, that enables to establish clarity, is missing. With this in mind, an analysis of how the DSA discusses and interprets FoE and its boundaries is paramount. Ultimately, this research will attempt to bridge the knowledge gap of how broad the scope of FoE in the DSA is, therefore answering the question of how FoE is interpreted within the context of DSA. Contrary to the discussed research, this thesis will complement the analysis of the DSA with an analysis of additional EU documents providing solutions to combating abusive expressions in the digital sphere. Therein, this thesis attempts to paint a larger picture of the European vision and interpretation of FoE. Notably, contemporary research also lacks deeper analysis of how the proposed instruments stand in relation to freedom of expression and its boundaries, going beyond a surface level conclusion whether an instrument undermines FoE or not. Having identified the knowledge gaps of the current state-of-the-art, this research will seek to answer three sub questions, which in turn will be used to answer an overarching research question.

1.5 Research Questions

Considering the established knowledge gaps, this thesis will aim to answer the research question of:

"How does the European Union interpret Freedom of Expression in the Digital Sphere?"

Guiding the research process, three sub-questions have also been formulated, which together will contribute to answering the underlying research question of the thesis.

a) How are the boundaries of FoE discussed within the DSA?

b) How does the EU envision solutions to abusive forms of expressions in the digital sphere?

c) To what extent is AI envisioned as a solution for coping with this?

1.6 Research Approach

The formulated research questions are interpretive in nature, for this reason, an interpretive content analysis serves as a fitting methodology to answer these research questions. This method allows for an in-depth analysis of text passages within EU documents that deal with the treatment of FoE related issues, thereby providing relevant definitions, solutions, or obligations. In doing so, this thesis investigates the meanings of the boundaries set for FoE and therein provides new insights to narrow the knowledge gap and ultimately develop a picture of how the EU aims to cope with the issue of regulating the digital sphere's seemingly never ending expansion and its consequences. Here, the review of contemporary literature established several knowledge gaps, which are reflected in the presented research questions. As conflicting findings about the question of how the DSA impacts FoE have been found, this thesis aims to establish clarity by taking a step back and analyse the passages in the DSA that affect FoE and assert boundaries in the first sub question.

Therefore, it is attempted to gain insights into how the EU interprets freedom of expression in the digital sphere and how its limits are placed. In this attempt, the particular meaning of the prescriptions the DSA provides relating to FoE are of interest, thereby providing a deeper look into the functioning of the DSA. Here, the produced insights will be examined against the backdrop of a review of FoE theory in chapter two, which will guide the interpretation of coded passages in the DSA. By having investigated the EU's interpretation of FoE in the digital sphere including the question of what constitutes an abusive form of expression (which are deemed to be unprotected from any interference) in the first sub question, the second sub question aims to establish how the EU aims to deal with this issue of combating such abusive forms. Here, in the review of contemporary literature on the DSA, it was established that it is quite ambiguous whether the DSA does or does not in fact violate FoE as well as scholars mostly providing vague speculative assessments whether the DSA instruments combating abusive forms of expression have the potential to safeguard or infringe on FoE. It remains to be seen, however, how those instruments, along with other provided solutions by the EU, stand in relation to FoE theory. Moreover, answering this sub question will fill the knowledge gap in regard to theorising the provided EU solutions, thereby looking at what those solutions mean in relation to FoE and its boundaries. Additionally, as the EU provides more relevant documents discussing abusive forms of FoE and how they ought to be treated within the digital sphere, next to the DSA, a more encompassing image of the EU's interpretation of FoE in the digital sphere is aimed to be provided. Finally, for the third sub question, the already established knowledge gap concerning the prior sub question applies here as well, however, an introduction of AI provides a critical nuance deserving of a distinct focus whereas in the attempt to answer the third sub question, this thesis seeks to lay down the same interplay as in the prior sub question with a deeper look at AI solutions. Ultimately, by answering the presented sub questions in chapter four of this thesis, an answer to the overarching research question can be formulated and answered in the fifth and final chapter. Here, the overarching research question is aimed to paint a picture of how the EU foresees FoE in the digital sphere to be interpreted, as in how broad its boundaries are and therefore what forms of expressions are deemed to be protected, as well as how this status is to be achieved and how the proposed solutions stand in relation to FoE.

1.7 Preliminary Conclusion

This chapter identified the knowledge gap on how the EU interprets FoE in the digital sphere, to bridge this gap, three sub questions and an overarching research question have been formulated. In order to answer these questions, an interpretive content analysis was deemed to be a suitable method to provide an answer to these questions and finally bridge the knowledge gap. The following chapter will provide the theoretical backdrop concerning FoE in academic discussion required for this approach. This will be followed by a discussion of the approach itself in chapter three, before using the theoretical framework in accordance with this approach to provide an analysis of the DSA and accompanying EU texts on FoE in the digital sphere. Therein, answers to the underlying sub questions will be formulated, allowing to finally answer the overarching research question in chapter five.

2 Theory

2.1 Introduction

This chapter will provide a theoretical framework that, for one, informs the creation of codes relevant to answering the presented sub questions and secondly, enables the interpretation of the passages coded in the analysis of selected EU documents in chapter four. Findings including key concepts of contemporary studies for this research will be introduced and discussed, thereby providing insight into the broader state of the art. Given the formulated research questions, different interpretations of FoE and its limits as discussed in scientific literature will be explored. The chapter begins with a historical account on how FoE developed in the western world, providing an understanding of key influences, transformations and different interpretations of what FoE signified over several centuries in order to establish a first understanding of what the right entails (2.2). This will be followed by an introduction to the diverse interpretations of FoE as discussed in academic discourse. Thereby, it will be underlined how in contemporary research, different interpretations persist of what expressions the right ought and ought not to protect, ultimately highlighting a contestation of how broad FoE boundaries shall be (2.3). As this research focuses on the freedom of expression in the digital sphere, relevant dynamics as discussed by contemporary scholars will be reviewed to gain an understanding of how the right is enacted in the digital sphere. Thereby investigating potential motivations behind the EU's conviction that the regulatory regime of the digital sphere is lacking (2.4). Similarly, the subsequent section will highlight concepts conceivable as abusive forms of expression thereby seeking to develop an impression of how and why some expressions are deemed to be outside the scope of fair discourse (2.5), providing a foundation for comparison between scholarly discussion and EU legislation of those in chapter four. The key findings will ultimately be summarised against the backdrop of the research objective (2.6). As this research aims to approach the DSA and accompanying EU documents addressing the boundaries of FoE with a diverse and framework to allow for wide-ranging analysis of potential interpretations, this chapter will not choose one theoretical conception over another but highlight and compare them to finally illustrate the contestation of the boundaries of FoE. Given this approach, this chapter aims to provide a framework that allows to look at the EU's interpretation of FoE in the context of the DSA, as well as its approach on how to cope with abusive forms of expressions in the digital sphere.

2.2 The Road To Freedom of Expression

The EU specified its intent to safeguard European fundamental rights, including FoE in the digital sphere, by implementing the DSA. When aiming to regulate what content is permissible in the digital sphere, one ultimately enters the realm of FoE, raising the question of what its meaning is. To answer this question, it is sensible to start by exploring the evolution of how the concept came to be and how it was interpreted by early scholars, thereby identifying key drivers and transformations surrounding the formulation of FoE. Prior to any attempts of ratifying a guarantee to prevent interference with free thought, conscience and expression, censorship and restriction of (critical) thoughts were a common occurrence. To illustrate, in the work of John, it is described how historical thinkers from Plato to Machiavelli were proponents of censorship with Plato advocating for the banning of Homer's Iliad and Odyssey, or the Byzantines destroying religious symbols contradicting their own religious denomination. According to John, the issue of fierce religious conflict in early Modern Europe served as a key driver behind the formulation of FoE, stating:

> "it emerged in its modern form only in the seventeenth century as a by-product of generations of horrific warfare between Catholics and Protestants" (John, 2019, p.34).

While prior to the Reformation, no major attempts at allowing a free exchange of ideas was identifiable, a noticeable shift was occurring during the Reformation Era. Here, John argues that the first significant step for broadening the tolerance of religious diversity was spearheaded by Hobbes in the 17th century (John, 2019), highlighting a moving of the goal post of the boundaries to FoE, at

least in theory. What followed, with the reformation, were concrete efforts in for instance the Holy Roman Empire to accommodate more religious minorities, embedding advocacy for more freedom into actual political structures. John Locke also defended the freedom for protestants in England to express their religious beliefs, however also advocated for excluding atheists and Catholics as he believed that these groups were incapable of obeying the law and therefore were to be outcast from public life. Approaching the 18th century and the formation of the United States of America, James Madison, described as a key figure in establishing the American interpretation of FoE as still present today, published essays in which he stressed the idea of pluralism of religious sects and its potential of preventing tyranny. With the constitution adopted in 1788, FoE was enshrined as part of the first amendment, providing preconditions for its establishment in the United States. The subsequent Post Office Act of 1792 allowed citizens to make use of their right by cheaply disseminating their opinions across the country, which was mostly used for the discussion of national politics. Notably, however, in spite of any established legal protection, the enshrined protection was, for instance in slaveholding states, infringed upon. Here, these states criminalised the circulation of literature dealing with the abolition of slavery (John, 2019). For one underlining a step backwards from Madison to a vision akin to Locke, at least in part and secondly, providing that law establishing FoE protection, does not provide all encompassing safeguards of an individual's abilities to express themselves.

A century later, one of the key texts in FoE discussion was published with John Stuart Mill's 1859 essay, On Liberty (John, 2019). His work is almost ubiquitously referenced across the papers addressed in this chapter and will also be discussed in section 2.3. After freedoms began deteriorating in the early 20th century, the Universal Declaration of Human Rights (UDHR) was signed in 1948 after the Second World War. With the UDHR, FoE was enshrined within international law for the first time, with Article 19 stating that:

"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 the non-binding declaration did not specify any boundaries to these freedoms (Goddard, 2021, p.70).

Five years later, the European Convention on Human Rights (ECHR) entered into force, from which emerged the creation of the European Court of Human Rights established to deal with reviewing potential violations of European fundamental rights, including FoE (CFR Article 11) (Goddard, 2021). Ultimately, in 1966, FoE was recognised in the International Covenant on Civil and Political Rights which was adopted by the United Nations General Assembly and signed by 173 parties, addressed in Article 18-20 it provides the freedom of thought, conscience and religion, compared to the UDHR, it is legally binding and discusses limitations. Whereas FoE, shall be limited by applicable law and if necessary, for the protection of public safety, order, or health and morals or fundamental rights of others, furthermore war propaganda and any advocacy of national, racial, religious hatred constituting incitement to discrimination, hostility, or violence ought to be outside FoE's boundaries (Goddard, 2021).

In sum, it can be seen that historically, FoE and its boundaries were directly influenced by major political and religious dynamics. Be it in regard to conflicts within Christian denominations during the Reformation, or even as a reaction to the Second World War to prevent similar travesties from occurring. Here, the boundaries of what expressions were protected gradually broadened before finally being enshrined in, for instance, the US Constitution or the ECHR. On the other hand, deterioration was also seen in the form of slaveholding states or authoritarian governments in Europe, narrowing the boundaries. Where in spite of laws foreseeing to guarantee FoE, governments may still encroach seemingly protected rights, ultimately raising concerns regarding any efforts to regulate FoE, as with the DSA. Nonetheless, it is understood that the right ultimately ought to enable individuals to express their views whereas over time, the tolerance for different views gradually increased.

2.3 FoE: Contested Boundaries

An overview of how freedom of expression evolved has been provided, where it was found that from its inception the meaning of FoE was highly contested. Many factors such as culture, applicable legal frameworks, moral frameworks or ideologies provide different interpretations of what FoE is or ought to be. With the DSA, the EU aims to implement its own vision of where the boundaries in regard to what content is permissible in the digital sphere ought to be. As this research aims to investigate this vision, this section aims to develop insights into the contestation in academia concerning the different perspectives and arguments on how broad the boundaries of the right ought to be. In doing so, it is aimed at enabling the interpretation of passages in EU documents that provide insight into the European vision of how these boundaries are to be asserted in the digital sphere. The previous section already examined a key difference between Locke and Madison, whereas John explains the contrast of Locke's view that specific groups in 17th century England threatened political stability and should be exempt from having a freedom to express their beliefs, against Madison's view of multiplicity serving as "the bulwark of the republic" (John, 2019). Here, boundaries of FoE are not directly set in regard to the content of an expression, but bound by group-affiliation. The previous section also introduced John Stuart Mill's 1859 essay On Liberty. Cohen-Almagor, a contemporary scholar in the field who formulated a critique of Mill's essay, clarifies that Mill "wished to establish as far as possible unlimited freedom of speech, arguing as a general rule that it should not be subjected to state interference or control" (Cohen-Almagor, 2017, p.22). Furthermore, according to Mill's account, one can never be sure whether a given opinion is ultimately true or false. For either scenario, Mill likens the silencing of an individuals' expression with a form of robbery or evil, as it deprives an opportunity of "exchanging error for truth" or;

> "the clearer perception and livelier impression of truth, produced by its collision with error" (Mill, 1859, p.11).

Cohen-Almagor raises a notable point, whereas according to Mill, democratic agencies may be abused to exhibit intolerance to out-of-favour opinions, ultimately limiting public discourse and people's ability to promote their views through disputation with other rival opinions. Here, Mill views that free, open discussions are bound to bring truth, even going so far as to invent artificial opinions as a means to challenge concurrent ones (Cohen-Almagor, 2017). As Cohen-Almagor explained, Mill was looking for the largest sensible scope of permissible speech. As part of this large scope, Mill proclaimed how harmfulness of an expression should not be a deciding factor in a potential restriction of an expression. However, an exception to this point is provided, whereas in extreme circumstances, a restriction may be justifiable. An explanation of what would constitute such an extreme circumstance is not given. Although, Cohen-Almagor argues that Mill would for instance agree with a restriction on incitements to murder. Furthermore, while not explicitly enumerating where these boundaries should be placed, Mill does view instigation outside the scope of FoE, which for Mill refers to:

> "any speech which is intended (or if not intended then at least recklessly uttered) to lead to some mischievous action which is delivered under circumstances conducive to taking that action" (Cohen-Almagor, 2017, p.31).

Relevant here is the speaker's intention that people will follow up the expression with performing a harmful act. To illustrate, Mill provides an example where a speaker would excite a mob in front of the home of a corn-dealer, stating that this dealer is starving the poor in an effort to incite the mob. Ultimately, according to John, Mill's view of FoE can be summarised as utilitarian, foreseeing FoE as "a positive norm that can hasten the emergence of the good society" (John, 2019, p.32). From this, it can be concluded that according to Mill, boundaries to FoE should be as broad as possible, whereas a view is provided that a broad scope is inherently good. Meanwhile, restrictions for instance based on an allegation of an opinion being harmful or wrong are considered inherently evil. However, instigating a (violent) mob to commit harm or inciting murder is seen as transgressing the boundaries of FoE leading to any good.

John contrasts Mill's utilitarian view with the American civil libertarian view, where FoE is viewed as a means to check a tyrannical state (John, 2019). While Mill is arguably one of the most influential scholars on FoE, contemporary scholars such as Cohen-Almagor criticise Mill's position, calling it "unsystematic and incomplete," opening "a scope for interpretations" (Cohen-Almagor, 2017, p.43). Cohen-Almagor argues that Mill does not discuss where exactly the boundaries to FoE ought to be, although it is raised how this might have been a deliberate decision on the part of Mill, to invite further debate on FoE (Cohen-Almagor, 2017). Another scholar frequently addressed in FoE related discussion, is Thomas Scanlon, whose papers on his conceptions of FoE in 1972 and 1979 will be discussed respectively, as well as Cohen-Almagor and Restrepo's criticism. Scanlon's 1972 paper provides a libertarian interpretation of FoE while his follow-up paper restricts the scope of FoE to a liberal viewpoint (Restrepo, 2013). Restrepo, discusses Scanlon's concepts of liberal and libertarian FoE, beginning his own work by introducing Scanlon's viewpoints, followed by providing criticism by delivering his own viewpoint of what he coins democratic freedom of expression. Scanlon's 1972 paper, views

> "The content of each and every expression [as] sacrosanct. Any person has the right to express any content, and the content of expression should be unlimitedly protected against the claim that it leads people to have false beliefs and to cause people to do harm" (Restrepo, 2013, p.381).

Therefore, FoE should not be restricted based on the notion that an individual was misguided by an expression of another person, as well as that a person was incited to commit violence, or had their tendency to commit violence increased by an expression of another person. Here, Scanlon does not specifically outline all categories of expressions he deems outside the limit of FoE, but his paper deals with justifications that ought not to be used to employ limitations on FoE, which is a focal point of Restrepo's criticism. In doing so, Scanlon does follow Mill's core argument, however, maintains that there can still be acceptable restrictions on FoE. Here, Scanlon names for instance defamation laws, or laws against the dissemination of knowledge on how to easily craft bombs as boundaries to FoE. This view is criticised by Restrepo, viewing "a theory that gives us simultaneously our reasonable grounds for legitimate limits and an unrestricted protection for the flow of any and all ideas at the same time" (Restrepo, 2013, p.381) as inconsistent.

It is pointed out by Restrepo that if one were to apply Mill's principle, restrictions on expressions for instance "constituting a type of assault" (Restrepo, 2013, p.381) are not permissible, he further illustrates that a person's reputation is crucial for their livelihood and professional development, from which it is argued to have law protecting people from any type of slanderous which could take away a person's deserved respect. Restrepo explains that such a safeguard is provided by Article 12 and 22 of the UDHR, providing protection of a person's dignity and reputation, as well as Article 21 providing that elections shall be free and governments be chosen by the will of the people. Both cases however can be undermined by massive scale lies, according to Restrepo. As a result it is argued how such lies fall outside the scope of FoE protection, as for instance, a choice in the election would have been made based upon deception. Here, it is concluded that in international law, one is not provided a right to encroach on other's rights. An intricacy in this regard is pointed out by Restrepo, whereas while Scanlon's libertarian view stresses the respect for people's autonomy including FoE, additional to the view of only good resulting from absolute freedom, Restrepo illustrates how false advertisement for instance a poisoned apple, hiding the fact of the apple's toxicity would directly cause harm to those who were deceived into buying a seemingly healthy apple. Ultimately, those affected would have their opportunities to express themselves infringed upon by being deceived and effectively poisoned (Restrepo, 2013).

Cohen-Almagor equally criticises a limitation in Scanlon's 1972 (and Mill's) work in regard to the view that every expression short of incitement to inflicting physical harm to others ought to be protected. While Scanlon renounced his libertarian FoE viewpoint, the libertarian viewpoint remains today. Cohen-Almagor provides that scholars such as Dworkin and Nagel, who share a libertarian viewpoint, go even further than Scanlon's initial view. According to their outlook on extreme

expressions, restricting FoE would even be wrong when the costs of not restricting such expression would outweigh the benefits (Cohen-Almagor, 2019). As with the libertarian view, Scanlon's liberal perspective sees FoE as immune from restrictions reasoned by harmful consequences resulting from an expression, regardless of a true or false expression (Restrepo, 2013). Here, Cohen-Almagor explains that the liberal view does not ignore potential harm of an expression, and sees the societal benefits of a free exchange of ideas as outweighing any costs that come with it (Cohen-Almagor, 2019). The liberal view however becomes more nuanced than its predecessor, as Scanlon explains that there are three different categories of expression, each requiring distinct treatment. Moreover, a category is determined by participant-, audience- and by-stander interests and the available form of regulation. Scanlon explains that governments are acting legitimately in restricting information on how to make nerves gas in promotion of personal safety, however, Scanlon does not share this view regarding an intervention in political agitation with the aim of preventing widespread social conflict. According to Scanlon, the two differ, as the latter deals with political matters, whereas Scanlon sees governments as partisan and unreliable in this regard. According to Scanlon, political speech deals with electoral processes and the activities of the government, whereas "The other categories of content that are protected religious speech, sexual, and others that cannot be reliably distinguished from them, and would consequently harm our expressive interests." (Restrepo, 2013, p.383). Restrepo criticises the liberal FoE viewpoint based on Scanlon's view that it would be consistent with the status quo, as it insinuates that the status quo serves as a suitable reference for societal norms. Furthermore, the view of strictly protecting the people's FoE from the government is blurring realities of power dynamics, whereas Restrepo explains depending on the context that private powers may pose a dire threat to rights such as FoE whereas the state is intervening in a way to ensure the rights of for instance, an oppressed minority.

Lastly, Restrepo criticises that liberal FoE prevents the outlawing of specific election fraud or deceitful political information, as expressions of fraudulent political nature have the potential to distort elections or abuse by government figures. Restrepo exemplifies this in regard to holding politicians who lie to create a casus belli accountable., whereas charging Bush with murder for the death of Iraqi victims of the Iraq War would not constitute a violation of Bush's FoE according to Restrepo, who raises that in actuality the rights of Iraqis have been violated as a consequence of Bush's expression (Restrepo, 2013). Ultimately, Cohen-Almagor states that in Scanlon's 1979 paper, FoE is seen as an effort to constrain government interference, in order to protect speakers (Cohen-Almagor, 2019), resembling the core motivation of Madison's view. Restrepo criticises both Scanlon's liberal and libertarian view of FoE as they enable the domination of other individuals. Motivated by this issue, he provides his own view of democratic FoE, which Restrepo explains takes the pillars upon which FoE is founded on to heart, these being individual and collective autonomy, the right to know facts of public interests and the information necessary for effective democratic control of the government. Notably, from democratic FoE emerges that the government, if required, steps in to provide a framework allowing genuine discussion to realise the aforementioned pillars. Restrepo clarifies that democratic FoE provides individuals with

> "the right to express any view they may wish to express that does not constitute an act of domination against another" (Restrepo, 2013, p.389).

Explaining that as soon as an act of domination occurs, an expression is no longer protected. The motivation behind ascribing the potential of others to realise their rights as boundaries of FoE, motivates Restrepo by explaining that freedoms come with accountability. One is provided with FoE, but on the other hand has to equally respect the rights of others. Once a person transgresses a boundary of another individual's right, an act of domination was performed (Restrepo, 2013).

With Restrepo's criticism of the liberal and libertarian viewpoint, some major contrasts in interpretations of FoE can be highlighted. While liberal and libertarian viewpoints foresee that negative consequences of for instance misguiding somebody are not reason to place a given expression outside the boundaries of FoE protection, Restrepo raises an issue with the more absolutist notion, by highlighting that FoE also provides certain responsibilities (Restrepo, 2013). Deceptive expressions for instance are undermining others rights and therefore fall outside the boundaries of FoE

according to democratic FoE. Here, libertarian interpretations as by Dworkin and Nagel would disagree with limiting FoE based on this question (Cohen-Almagor, 2019) , whereas Scanlon focussed more on the question of keeping government's in check, whereas his conceptions of FoE ignored aspects such as those highlighted by Restrepo (Restrepo, 2013).

The now established contestation concerning the boundaries of FoE is not only relevant in academia. Different jurisdictions and their corresponding legal frameworks offer different interpretations of FoE and its limits. A comparison between for instance the location of most of the affected VLOPS addressed in the DSA, based on the First Amendment relevant for these companies in their home country the United States, versus the definition of the EU which tries to regulate these companies as laid out in the Charter or the Convention. While in prior sections, international law was already discussed, a contrast between the EU and US will be briefly highlighted. Within the polity of the EU, the freedom of expression and information is enshrined in Article 11 of the Charter of Fundamental Rights of the European Union (European Union, 2012), which states that:

"Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers", as well as that the "freedom and pluralism of the media shall be respected" (European Union, 2012).

While the condition of "without interference" may lead one to think that there are no boundaries to the dissemination of any expression, the European Convention for the Protection of Human Rights and Fundamental Freedoms, provides limitations to the FoE. Specifically those which:

"are necessary in a democratic society such as those in the interests of national security, for the prevention of disorder or crime and for the protection of the rights of others" (Cassim, 2015, p.316), any restriction however needs to be proportionate to a legitimate aim a government pursues (Cassim, 2015).

Different interpretations of FoE as highlighted above, likely provide different outlooks when these parameters are met, a final decision whether an expression is protected or an infringement was legitimate or not however ultimately lies with the European Court of Human Rights (Goddard, 2021). This interpretation is juxtaposed with the First Amendment in the United States, which was already hinted at in the first section of the theoretical framework. Here, the bill of rights simply states that_

"Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances." (Goddard, 2021, p.75).

Based on this difference, it is highlighted by Cassim how the dissemination of Nazi propaganda is illegal in both France and Germany for instance, but protected in the United States (Cassim, 2015). Furthermore, in the United States, offensiveness is not seen as a sufficient reason to infringe on FoE, whereas the European Court of Human rights states that an intention to spread racist ideas or views has to be present to deem hate speech criminal. Nonetheless, similar to the US, information or ideas which offend, shock or disturb are also protected under the FoE reasoned by the values essential to democratic society (Cassim, 2015). The comparison shows that the United States are more lenient and reflect a more libertarian conception of FoE especially in regard to the constitution foreseeing that the Congress has to refrain from placing boundaries to FoE, whereas the conception provided in the CFR reflects the notion of FoE being limited by the rights of others, therefore Restrepo's notion of domination (Restrepo, 2013). In summary, it has been established that academic discourse on the question of what constitutes FoE and where or if boundaries should be set is highly contested, which gives rise to the question of how this contestation is addressed in the DSA and if any of the provided interpretations and dynamics are reflected in the DSA. It was also found that companies and their services which are regulated by the DSA are found in a situation where the boundaries they have to enforce differ from those they are usually accustomed to in their home countries.

2.4 Freedom of Expression in the Digital Sphere

Until now different interpretations of FoE and its boundaries have been discussed, an aspect mostly left out is discussion of FoE dynamics in the digital sphere. The EU seeks to regulate what content is permissible to disseminate on social media platforms and what content is not, thereby creating a safe, predictable and trusted online environment. Against the backdrop of the DSA's goals and the quest to deal with for instance hate speech and disinformation, it is sensible to look at those dynamics. With this in mind, this section will identify notable dynamics of FoE in the digital sphere and introduce relevant concepts. Firstly, the paper by Leopoldo Garcia Ruiz discusses FoE and provides an explanation for what is observed to be a deterioration of FoE and related rights in the advent of online fact-checking and content moderation following the 2016 US Presidential Elections as well as COVID-19 (Ruiz, 2023). Similarly, Tropina also discusses content moderation during the so-called 'infodemic', a term coined by the World Health Organisation (WHO) referring to the "spread of rumours and misinformation related to the pandemic" as well as an abundance of related information during the corona pandemic (Tropina, 2023, p.229). Ruiz's article provides three distinct concepts; disinformation, misinformation and malinformation. An expression constitutes disinformation if it is indeed false (material element), the person expressing it, is aware of it being false (cognitive element) and intents to deceive (volitional element), whereas misinformation occurs if the material element applies, yet the person disseminating information is not aware of its falseness. Malinformation constitutes types of expressions that contain all three elements and additionally intend to offend or cause harm to others, which includes malicious leaks or hate speech, for instance. The terms dis- and misinformation became ubiquitous following the 2016 US Presidential election campaign and Trump's presidency, which gave rise to online fact-checking. Here, Ruiz criticises excessive paternalistic content-moderation and fact-checking during the pandemic (Ruiz, 2023).

Tropina explains that during COVID, efforts to curb dis- and misinformation accelerated, as underlined by the WHO stating that misinformation costs lives, with a growing sense of urgency causing both governments and platforms to act and restrict the dissemination of information on COVID-19. Here, Tropina stresses that these restrictions were partly disproportional, undermining FoE, ultimately blurring the lines in what separates more liberal democratic and authoritarian systems that employ criminal law to censor and restrict FoE (Tropina, 2023). Ruiz enumerates that restrictions were based on WHO guidelines for COVID-related information, as well as guidance by the EU and the United Nations, among other institutions. For instance, the Spanish Government confirmed that it was in fact monitoring social networks for the purpose of detecting potentially dangerous or criminal speech and disinformation campaigns, which were followed by efforts to combat their spread. During the pandemic, Twitter went on to consider any post that contradicted the guidance of health authorities as misleading (Ruiz, 2023), whereas in this context, boundaries to FoE were set by health authorities and enforced by a private company, as for instance also highlighted in Restrepo's criticism of Scanlon.

Ruiz stresses that fact-checkers supposedly went further than fact-checking nonsensical claims, by infringing on legitimate debates by qualified scientists from renowned universities such as Harvard, Oxford and Stanford on questions of natural immunity, the origin of the virus including the lab theory (Tropina, 2023) or potential mRNA vaccine side effects, whereas any claims deviating from those of the authorities were automatically qualified as false leading to removal of both comments and commentators from platforms such as Twitter or even shunning in academia and journalism (Ruiz, 2023). Here, according to Ruiz, the debate was censored by suppressing content which was not clearly classifiable as either dis- or misinformation. Ruiz puts forward that after Musk's takeover, internal communications of Twitter discussing these processes revealed that communication with US federal authorities occurred discussing actions against accounts that did not reflect information which the government endorsed (Ruiz, 2023). This is also underlined by Tropina who explains that topics such as the lab theory were initially deemed to be fake news but after a year, re-entered legitimate debates. Similarly, the WHO first advised against the use of face masks for healthy people only to change this opinion in June 2020, underlining how quickly during a crisis, the question of what is true, and the resulting speech restrictions can seemingly arbitrarily change. A central point by Tropina is also that due to the immense velocity and volume of content during the infodemic that had to be moderated,

platforms began to increasingly rely on automated tools for content removal instead of human moderation, which exacerbated the removal of content which did not violate platform policies (Tropina, 2023).

Ultimately, based on Ruiz's findings, during the pandemic, the limits of FoE in the digital sphere were arguably set arbitrarily by acting authorities and then adopted by private firms such as Facebook and Twitter (Ruiz, 2023), hinting at FoE in the digital sphere and its limits being flexible based on a context ascribed to by a given government or influential institution. From this emerges that during crises such as the corona pandemic, the limits of FoE in the digital sphere are at the mercy of acting governments and authorities which steer the content-moderation practices of private companies such as Facebook and Twitter. Here, Ruiz concludes that:

> "disinformation is undoubtedly a disease for democracy, but the available remedies can be even worse." (Ruiz, 2023, p.17).

Therefore, it is proposed that as FoE is a basic individual and social good in a democratic state, restrictions, even in the attempt to combat mis- and disinformation should only be permitted in a sense:

> "when the pursued communication represents a direct incitement to violence, or becomes a necessary means to violate the rights of others." (Ruiz, 2023, p.19),

reflecting the sentiment by Scanlon and Mill in regard to the limit of inciting violence, while also mirroring Restrepo's view of the rights of others serving as limits to FoE. Tropina similarly concludes that while rushed responses of restricting harmful content such as mis- and disinformation during a pandemic could be justified by a sense of urgency and public health concerns, however these also pave the way for a normalisation of vague approaches to criminalise dis- and misinformation in the future. Moreover, as a result platforms are encouraged to, on their own accord, proactively combat phenomena such as the infodemic, placing them in a position of quasi-judicial authorities, which however are not equipped to act as (Tropina, 2023).

Both scholars raise issues regarding content moderation and fact-checking of dis- and misinformation. In the context of a global crisis, Tropina argues that letting social media platforms exercise the controlling of content via terms of services, the responsibility of balancing FoE with competing acts is outsourced into private hands, whereas these are neither competent nor required to assess these interests and the underlying conflict between them, ultimately acting as an arbiter on speech restrictions, which especially becomes difficult in a global context transcending a variety of jurisdictions and cultural contexts. Namely, when platforms function as such arbiters, they are not immediately held accountable for the upholding of democratic values or human rights. Here, Tropina raises that platforms are not accountable to citizens, but to their shareholders. Ultimately, Tropina highlights the question and the accompanying uncertainty of where to draw the line between true and false information (Tropina, 2023), which is also relevant for the question of what is and is not protected by FoE therefore the meaning of FoE in itself. Tropina raises that a phenomenon such as the corona pandemic brings a lot of uncertainty especially in its early stages, when a virus is discovered, leading to changing circumstances and obscurity concerning the question of who serves as a reliable and authoritative source of information. Whereas, even governments and national health agencies are accused of supplying covid-related misinformation. In this, questions of who gets to decide what is true and what is not arise, translating to the notion of truth becoming subjective, depending on one's own convictions and interests. Especially when sufficient information on a controversial topic is absent or rapidly developing. This dynamic puts both the EU and service providers in a precarious situation, as legal obligations as foreseen by the DSA force service providers to act and potentially interfere with FoE. For this, the EU and service providers would have to be able to clearly discern what information would be dangerous medical disinformation and therefore outside the boundaries of FoE.

Contrary to the viewpoints and concerns provided by Ruiz and Tropina, Saunders proposes a more stringent approach to misinformation. While most use Mill to explore a more liberal view of FoE, Saunders relied on Mill's statement that in cases of extreme exigency, restrictions on freedoms are permissible. In his paper, Saunders argues that the recent pandemic may be describable as such an exigency, whereas even from Mill's view restricting FoE in regard to potential vaccine misinformation is permissible, as with such misinformation vaccine hesitancy is likely to increase (Saunders, 2023). If one were to adopt this view, social media platforms would have to perform the role of an arbiter of truth as criticised by Tropina and Ruiz in their respective works, underlining the previously highlighted contestation regarding the meaning of FoE even in regard to the digital sphere. Furthermore, the view presented by Saunders would interpret the suggestion of Ruiz and Restrepo of limiting FoE's by other's rights, in such a way that disseminating information causing vaccine hesitancy is in violation of other's rights and therefore not a protected expression. However, the issue previously addressed considering how for instance the EU, any governing body or a social media platform can clearly discern what is true and what is not becomes relevant again, whereas especially a libertarian view would view this critically while also from Restrepo's democratic viewpoint it is unclear how one can establish if domination occurred. In this regard, it is relevant whether vaccine hesitancy for instance is a result of disinformation which would constitute domination, while on the other information may just contradict statements by a health agency and as raised by Tropina and Ruiz information may be falsely classified as disinformation. This ultimately underlines the intricacy of Tropina's notion of having to act as an arbiter of truth in the digital sphere.

Moving away from the example of the corona pandemic, Cassim notes that the advent of the internet led to the dissemination of hate speech and cyberhate-related activities becoming widespread, with an increasing number of websites offering racist content. This gave rise to another dynamic as different interpretations of FoE clash in the digital sphere. It is raised that for instance in regard to online hate speech a given expression may be regarded as criminal in one jurisdiction but legal in another, for instance in regard to Cassim's earlier example of Nazi propaganda or paraphernalia (Cassim, 2015). Hate speech, according to Cassim, can be understood as "disparaging and abusive comments, words and phrases directed at individuals or groups representing a specific race, religion, ethnic background, gender or sexual preference." and has a propensity to injure its victims (Cassim, 2015, p.309). According to Cassim, hate speech may infringe basic human rights including privacy, human dignity or rights to freedom of religion, belief and opinion. Here, John notes that public debates in the digital sphere as "depressingly rare", further highlighting a deterioration of genuine public discussion in regard to the state of FoE in the digital sphere. John adds that whereas in a time before global communication via the internet, a provider was liable for content they disseminate, today in the United States at least, Section 230 of the Communications Decency Act prevents providers from being held accountable for the content they host. John raises that this poses a problem as platforms are able to alter media streams of their users (John, 2019). Finally, a key dynamic of FoE in the digital sphere is raised by Cohen-Almagor in his work criticising Mill. He raises a relevant point concerning Mill's example of an excited mob in front of a corn-dealer's home. Through the rise of new technologies, a mob no longer needs to be situated directly outside a person's home. Today, (incitement to) violence is transmittable over the internet and can target many people in many places (Cohen-Almagor, 2017). This challenge became increasingly grave against the backdrop of surging online terrorism, which will be discussed in the next section. To recapitulate, this section highlighted how the digital sphere enables the acceleration of disseminating misinformation and disinformation, which becomes especially relevant in times of crises. Under these conditions, boundaries of FoE can shift quickly when information is rare and susceptible to change. With the DSA the EU foresees to cope with these notions, this section found however that this is a risky task that can result in violating FoE. Likewise, just as the conception of FoE itself, the question of what is true and what is not is disputed, whereas different scholars and viewpoints provide different answers on how to deal with these dynamics. Notably, the digital sphere is also a realm where violent expressions are frequently disseminated, for the analysis it will be investigated how the EU aims to deal with identifying and mitigating such forms.

2.5 Abusive Forms of Expression

As already hinted at in the previous section, the EU wants to make the digital sphere safe and ensure the safeguarding of fundamental rights with the DSA. To investigate the EU's approach, this section will discuss the issue of (abusive) forms of expression that may or may not fall outside the scope of legal protections and can pose as drivers behind the digital sphere posing as a problematic environment. Of these, many are discussed by Bromell, who in his book explains that the livestreaming of the terror attack in Christchurch 2019 gave rise to a phenomenon of online terrorism. With this attack, live-streaming became a medium and message abused by terrorists. Bromell raises that in the case of Christchurch, the aim was not just to kill, but create a video of killing Muslims, which Bromell refers to as a performance crime. Following Christchurch multiple similar deadly shootings occurred, including in Halle, Germany, whereas prior to attacks manifestos are posted to messaging boards such as 8chan while the eventual attacks are then live-streamed. In reaction to the Christchurch attack, the Christchurch Call was formulated, asserting that despite a respect for FoE and the conviction that free internet is extremely beneficial to society, nobody has a right to create or share terrorist or violent extremist content online. To combat these forms of extreme expressions, the Christchurch Call foresees that governments and tech companies implement measures to prevent their dissemination (Bromell, 2022).

Efforts like these, again raise the question of determining when an expression is interpretable as, for instance, violent or extreme. Relatedly, in regard to hate speech for instance, Frank La Rue, the special rapporteur on the promotion and protection of the right to freedom of opinion and expression established a hate speech categorisation, which is of interest as it underlines the question of when hate speech enters a realm which warrants interference or even prosecution. Whereas in terms of regulating social media platforms, it has to be provided how to make a distinction among the three forms of hate speech. Here, La Rue explains that some expressions may be criminal under national and international law while some might only be punishable in a civil suit or justify a restriction. Notably, La Rue establishes a third category for expressions that are neither criminally nor civilly punishable, but concerning in regard to tolerance, civility, and respect for others, whereby this category is classified as lawful hate speech (Bromell, 2022). Relatedly, Cassim explains that following the Second World War, European countries began to formulate laws that restrict hate speech in an attempt to promote respect and equality (Cassim, 2015), meanwhile in the United States the first amendment remained the major point of reference for hate speech and other abusive forms of expression. Here, Cohen-Almagor is of the view that the tolerance for hate based on the first amendment comes at a price, stressing that hateful speech translates to hate crime (Cohen-Almagor, 2019).

In the discussion of hate speech, Bromell raises that the term hate speech is misleading as anything publicly expressed may incite discrimination, hostility, or violence. Furthermore, hate can be expressed in various forms next to speech, such as in form of a cartoon, henceforth, Bromell proposes to instead refer to dangerous or extreme expressions when referring to a:

> "public communication that intends or is imminently likely to incite, discrimination, active hostility, or violence" (Bromell, 2022, p.152).

Another issue in regard to the term hate speech, as raised by Strossen who is quoted by Bromell, is its overuse or misuse. The term is used often in regard to public policy issues, or by individuals who speak of assault when they are confronted by an idea offensive to them. From an equation of controversial ideas with physical violence, some feel motivated to push for the restriction of out-of-favour ideas, reflecting some of Mill's concerns of stifling unpopular opinions. Strossen provides the concept of the "Mourner's Veto", whereas emotions are used to suppress controversial expressions. This dynamic comes at the cost of the ability to make use of FoE and for instance engage in reasonable debate, motivated by this it is argued that the boundaries of FoE should not be dictated "by the most sensitive person in the room" (Bromell, 2022, p.153). Other scholars go further, such as Edwin Baker who views that laws restricting racist or hate speech are violating a speaker's autonomy, while arguing that hateful expressions do not, thereby viewing restrictions as illegitimate (unless a

person wilfully agrees to refrain from for instance making such comments in a professional position) (Bromell, 2022).

Similarly, John diagnoses that for instance hate speech bans are being issued by universities, whereas controversial speakers are disinvited, safe spaces are mandated within classrooms or triggers warnings ought to be issued for potentially sensitive material which according to John ultimately leads to an alteration of instructional dynamics. As a reaction, some professors from Harvard and Princeton issued a public statement against these phenomena, instead favouring "trust seeking, democracy, and freedom of thought and expression" (John, 2019). On the other hand, Cohen-Almagor states that ultimately:

> "in a perfect world we would respond to hate with education, not criminal laws. But our world is not perfect, and history shows that hate speech can lead to horrific crimes" (Cohen-Almagor, 2012, p.57).

These contrasts ultimately leave the question of what approach is the correct one unanswered. To conclude, this section found that the question of hate speech is very dependent on how abusive or extreme one perceives a given expression to be. On one hand, in Bromell's book for instance, it is diagnosed that terms such as hate speech are overused to infringe on genuine speech. On the other hand, there are very critical developments in the digital sphere whereas the internet and social media platforms are used to spread terrorism, putting the life of many at risk and dissemination of dangerous content. In regard to the research focus, it will be investigated how the EU aims to address these. The table below (Table 1) visualises the findings of this chapter and illustrates how they relate to the sub questions presented in this thesis and highlight the key relevant findings for the respective sub question, detailing how they will be used in the analysis in chapter four.

Table 1: Key Insights

2.2	**2.3**	**2.4**	**2.5**
John: FoE emerged from religious conflict (John, 2019). SQ1	Restrepo: FoE boundaries are asserted by the rights of others (Restrepo, 2013). SQ1&2&3	Ruiz: Disinformation is a disease for democracy, yet remedies may undermine FoE (Ruiz, 2023). SQ1&2&3	Cohen-Almagor: Education not sufficient to tackle hate speech (Cohen-Almagor, 2012). SQ2&3
John: Despite legislative safeguards, FoE can be violated (John, 2019). SQ2&3	Scanlon 1972: The content of each expression is sacrosanct, no inferences based on harmfulness of an expression (Restrepo, 2013). SQ1&2&3	Tropina, Ruiz: FoE boundaries were arbitrarily set in the digital sphere (Tropina, 2023) & (Ruiz, 2023). SQ1&2&3	Restrepo: Disinformation campaigns undermine FoE (Restrepo, 2013). SQ1

2.2	2.3	2.4	2.5
Madison: FoE as a check for tyranny (John, 2019). SQ1	Scanlon 1979: Political Speech ought to be protected from governmental interference (Restrepo, 2013). SQ1&2&3	Tropina: Service providers increasingly rely on AI content moderation (Tropina, 2023). SQ3	Saunders: Disinformation in crises deserving of interference (Saunders, 2023). SQ1&2&3
Mill: Abuse of democratic agencies to stifle unpopular opinions (Cohen-Almagor, 2017). SQ1&2&3	Restrepo: Shift in power dynamics: FoE in danger by private actors and government (Restrepo, 2013). SQ1&2&3	Tropina: Urgency to combat harmful content runs risk of normalising vague approaches, undermining FoE (Tropina, 2023). SQ2&3	
	Mill: Interferences only on expressions constituting direct incitement (to violence) (Restrepo, 2013). SQ1&2&3		

**SQ refers to the related sub question.*

2.6 Preliminary Conclusion

This chapter provided insight into the state-of-the-art of FoE and the discussion of its boundaries. With this, the theoretical findings can be connected to the underlying research questions. The first sub question of this research aims to investigate how the DSA discusses the boundaries of FoE in the digital sphere. The chapter found that different scholars provide different interpretations of where those boundaries ought to be. Predominantly, it was identified that libertarian scholars view FoE as means to keep governments in check, whereas Mill advocates for the free exchange of opinions (Mill, 1859). In regard to boundaries, Mill only foresees extreme contingencies such as incitement as outside the boundaries, while scholars like Restrepo disagree with Mill and Scanlon by providing that the boundaries to FoE shall be ascribed to by the rights of others. Ultimately, one can differentiate between the democratic, liberal and libertarian FoE boundaries. Here, boundaries are either ascribed to by the rights of others (Restrepo, 2013), in regard to political speech no limits should be set, while government interference shall be limited to for instance cases of information being evident to cause harm in the form of instruction for the creation of nerve gas (Scanlon 1979) (Restrepo, 2013), or from a libertarian or absolutist view as provided by Dworkin and Nagel, boundaries should not be placed at all (Cohen-Almagor, 2019). Additionally, the historical account of FoE highlights that it is rooted in conflict, reflecting this contestation, as FoE proved to be entirely dependent on the context of religious and political dynamics (John, 2019). For the first sub question, it will be seen how these boundaries are reflected in the DSA. Considering the overview of chapter one, it is expected to find the DSA interpret freedom of expression similar to a democratic FoE conception, given its aim to respect fundamental rights (Wilman, 2022).

Adding to the issue of contestation, which was found to be a key element of FoE, is the rise of the digital sphere. New dynamics are transforming the way the FoE is enacted, introducing new challenges such as misinformation (Ruiz, 2023) but also content-moderation (Tropina, 2023). Views such as those by Mill and Scanlon (Restrepo, 2013) of pursuing truth insinuate that for instance mis- and disinformation should be permitted as it is an objective judgement whether information is ultimately true or false is impossible (Mill, 1859). Relatedly, Tropina raises the issue that for instance during the infodemic, the question of what is authentic information or disinformation is particularly nuanced and may even change over time (Tropina, 2023). Restrepo, on the other hand, argued how disinformation, for instance in a political context, constitutes a form of domination and therefore ought to be outside the boundaries of FoE protection (Restrepo, 2013). Moreover, authors such as Cohen-Almagor (Cohen-Almagor, 2012) and Saunders (Saunders, 2023) highlight how abusive forms of expressions such as hate speech undermine the rights of others. Lastly, Bromell provides a view highlighting both fronts of the discussion, for one highlighting the immense risks of online terrorism but also highlighting criticism for hate speech regulation and restrictions on FoE both in the digital sphere as well as physical realm (Bromell, 2022). Relating this to the second sub question of this thesis, it will be analysed if and how the EU interprets issues of abusive FoE as indeed abusive, whether they warrant interference and lastly how the EU envisions to combat those. Given the selected data, particularly the codes of practice and conduct, it can be inferred that expressions such as hate speech and disinformation are in fact seen as problems which ought to be combated.

Lastly, while the second sub question already aims to address the aforementioned dynamics that may require interference by social media platforms, a question arises of how such interferences could be performed given the huge amount of data that has to be reviewed (Sulmicelli, 2023). For this, it will be looked at, how AI can deal with the identified phenomena based on how the EU addresses the technology across the analysed documents. Based on the overview established in chapter one, it is inferred that AI may be foreseen as a tool used in identifying content which ought to be combated. Having provided a theoretical framework and relevant key concepts in the discussion of FoE and its boundaries, the next chapter will highlight how this framework will be used to finally allow the interpretation of coded passages in chapter four's analysis.

3 Method

3.1 Introduction

This chapter will lay down the methodological approach of the analysis in chapter four to formulate an answer to the underlying research questions. In the context of this thesis, this chapter will thereby establish the approach that aims to lay down how the EU interprets FoE in the digital sphere. For this, the first section will provide the selected case that aims to encapsulate this interpretation. Here, the role of FoE in the DSA is described, as well as the origins of the act and how the role of FoE is developed within it are highlighted (3.2). This will be followed by a description of the collected documents and an explanation of how the collected data was identified in order to obtain a sample that allows to paint a picture of how the EU interprets FoE in the digital sphere (3.3). Finally, it will be discussed how the analytical approach in the form of an interpretive content analysis (ICA) will be utilised to analyse this data allowing insights into the methodological process of this research, ultimately aiming at establishing reliability and reproduction (3.4). The final section of this chapter will summarise the research activities conducted in this thesis and provide an overview over the coding scheme used in the eventual analytical process (3.5).

3.2 Case Selection

The aim of this study is to paint a picture of how the EU foresees the establishment of FoE boundaries in the digital sphere in the context of the DSA. With this act, the EU has taken a giant leap in the regulation of FoE in the digital sphere, whereas this step gives rise to several questions. Next to the research questions this thesis aims to answer, a question of how the EU envisions the regulation of FoE in the digital sphere emerges. With Musk's purchase of Twitter and the subsequent rebranding to X a lot of attention has been drawn to these questions. The opening of formal proceedings by the European Commission against his platform (European Commission, 2023) underlined a conflict between his personal vision (PR Newswire, 2022) and the regulatory regime prescribed by the EU. As established in the introductory chapter, prior to the DSA, internet service providers were mostly unregulated by the EU and exempt from any obligations such as monitoring user consent under the ECD. Therefore, before the introduction of the DSA, the question of FoE in the digital sphere remained entirely on the member state level whereas countries either decided to implement specific laws regarding FoE in the digital sphere or not. Enter; The DSA, this act can arguably be presented as the most relevant paper concerning FoE in the digital sphere, as it adds numerous obligations to service providers which ultimately relate to the freedom of expression in the digital sphere by laying down how to moderate content in accordance with EU law. As laid down by Heldt, policymakers in the EU realised their perceived discontent with the shortcomings of the regulatory regime under the ECD, including von der Leyen, who called for efforts to address the issues of disinformation and online hate messages (Heldt, 2022). With this sudden realisation of needed change in the digital sphere on the part of the EU, a relevant case is constructed.

Now that the DSA has entered into force, the aforementioned questions have to be answered. In the DSA, it is written that the ongoing digital transformation is bringing risks for individuals, companies and society in general. Furthermore, it is explained that in reaction to this, member states are beginning to tackle issues such as illegal content or disinformation. Here, however, the EU sees a problem in diverging national laws undermining the freedoms upon which the EU has been established, with a free market at the forefront. Service providers and users shall benefit from a free market which requires intervention in the form of the DSA according to the EU (European Union, 2022). As the DSA aims to mitigate these challenges it directly regulates FoE and its boundaries, thereby especially in the case of VLOPs and VLSEs, lays down what content has to be removed and makes for instance service providers as well as users liable to interference or even criminal prosecution (European Union, 2022). Looking at this and the question of how the EU foresees boundaries of FoE and solutions to the aforementioned challenges, more questions arise whether the EU further specifies answers to these questions. Thereby, this thesis aims to investigate EU documents that refer to these. Here, the DSA serves as a suitable starting point for an analysis to find answers to

the question of FoE interpretation in the digital sphere. Additionally, EU documents (next to the DSA) dealing with abusive forms of expression and how the EU envisions to deal with them in the digital sphere are of interest. Notably, as the EU published several documents on how to deal with abusive expressions in the digital sphere, including codes of conduct, codes of practice and a regulation on combating terrorist content, insights from these documents are paramount to ultimately paint a picture of how the European Union interprets FoE in the digital sphere. Ultimately, to establish a concrete case for this picture, it is aimed to analyse documents that are still relevant in regard to legislative influence and dealing with contemporary dynamics regarding FoE in the digital sphere. Finally, relevant to answering the final sub question, the EU presented its landmark regulation for AIA, the Artificial Intelligence Act (AIA) therefore inviting an analysis of it as well. Having discussed the case that will be investigated, whereby relevant passages for each sub question will be coded, the following section will lay down the method of how these documents have been collected.

3.3 Method of Data Collection

This section will provide the approach of how the necessary data to perform an analysis that allows to paint a picture of how the EU draws the boundaries of FoE in the digital sphere and cope with arising challenges has been collected. To ultimately answer the underlying research questions, data was collected in two phases with two distinct aims. This choice was motivated as only after having established the state-of-the-art, its knowledge gap and insights into FoE related theory, it was deemed suitable to be able to identify the relevant EU documents that allow the analysis of how the EU interprets FoE in the digital sphere. Additionally, as the aim of this research is to analyse FoE documents with a framework of FoE literature free from as much bias as possible, this choice was made in an attempt to avoid that insights gained prior to the analysis, from either category (FoE theory and EU documents) could sway the formulation of the theoretical framework to finally ensure reliable interpretation of coded passages. The first phase of data collection began with collecting data in the form of news articles, to inform the general context of the research and inform the topic choice. Subsequently, academic literature has been collected, whereas on the basis of this literature, a foundation to conduct interpretations of relevant passages in EU documents (collected in the second phase) is created. For this particular data collection, libraries including Scopus, FindUT by the University of Twente and Google Scholar were used to collect secondary data. Here, academic literature on European and human rights law, FoE theory and FoE in the digital sphere including hate speech and phenomena such as disinformation has been collected. In this process, the DSA has been identified as a major object of interest, as it was frequently addressed in news articles and academic sources alike. Ultimately, by aiming for reliable sources, such as established academic publishers, biassed and faulty research was aimed to be minimised ultimately decreasing threats to reliability and validity by aiming for conceptualisations based on (the most possibly) reliable sources.

In the second phase of data collection, data that will be subject to the actual analysis has been collected. These being documents published strictly by the EU and its respective institutions, which have been collected via purposive sampling. Here, EU papers dealing with FoE in the digital sphere, disinformation, hate speech, illegal or extreme speech and forms of expressions such as terrorism and a document specifically on AI have been collected. In doing so, it is aimed to have gained a sample of documents that provide information on how FoE is interpreted, how abusive forms are interpreted and how the EU foresees to cope with those issues in the digital sphere, including the use of AI mechanisms to do so. To ensure the collection of strictly relevant documents, which is integral to the data collection process in an ICA (Drisko & Maschi, 2015), purposive sampling allowed for the targeted research and identification of documents that in fact discuss FoE in the digital sphere for instance. In this stage of the data collection, it was aimed to discard no longer relevant sources that are not directly influencing the current regulatory regime in the EU. For instance, the ECD or outdated versions of codes of conduct or practice were avoided to prevent a distortion of the findings of analysing current regulatory regime, for this reason non-purposive sampling methods were not utilised, as purposive sampling allowed for deliberate inclusion and exclusion of documents. In this regard, to safeguard that findings will be of relevance, only papers published within the last ten years have been collected. Furthermore, as in the first data collection phase, it was attempted to ensure the

reliability of the sources and as this research aims to strictly analyse official EU papers, the EUR-Lex service was utilised. This service enables searching the official library of the EU and provides direct access to EU law documents, moreover the official websites by the European Commission was used to gain access to documents such as codes to address a given FoE related phenomenon. This was done to ensure that only relevant and informative data in the form of EU documents, central to the analysis of FoE in the digital sphere, will be collected as foreseen by Drisko & Maschi (Drisko & Maschi, 2015). In this search, the terms "Artificial Intelligence Act", "Digital Services Act & Freedom of Expression","Freedom of Expression & European Union", "Freedom of Expression on the Internet", "Digital Freedom of Expression European Union" and "Hate Speech & European Union" have been used. With this approach, besides the DSA (2022), six more documents have been collected and will be subjected to an ICA.

1.EU Human Rights Guidelines on Freedom of Expression Online and Offline (2014)

2.The EU Code of Conduct on Countering Illegal Hate Speech Online (2016)

3.Tackling online disinformation: A European Approach (2018)

4.Regulation to Address the Dissemination of Terrorist Content Online (2021)

5.The Strengthened Code of Practice on Disinformation 2022 (2022)

6.Artificial Intelligence Act (Corrigendum) (2024)

These documents have been chosen as they are papers by the EU dealing with the respective objects of interest of this research and allow gaining insights for answering the underlying research question of this thesis. Therefore, this thesis will in total analyse seven documents, with a combined number of 695 pages. The publication date of the identified documents ranges from 2014 to 2024, providing a fairly recent time span allowing a provision of FoE interpretations which are still relevant or applicable and capture current technological developments relevant for FoE in the digital sphere. Finally, by selecting purposive sampling, a risk however arises of missing integral data that would have provided necessary insights relevant to the research questions. In an attempt to mitigate this issue, the aforementioned sources have been scanned multiple times, as well as documents having been checked for any references or mentions of other potentially overlooked EU documents after the analysis. Having laid down how the necessary data was collected, the following section will explain how this data will be used to gain insights relevant to answering the presented research questions.

3.4 Method of Data Analysis

For the analysis of the aforementioned documents, an interpretive content analysis, making use of a coding scheme will be conducted as the method is "used to interpret text data from a predominately naturalistic paradigm" (Hsieh & Shannon, 2005, p.1278), posing as a suitable fit for this effort. Drisko & Maschi discuss the methodological approach of an ICA in chapter three "Interpretive Content Analysis" of their book Content Analysis from the Pocket Guides to Social Work Research Methods (Drisko & Maschi, 2015). In this chapter, the scholars explain that some scholars view interpretive and qualitative content analysis as being synonymous, Drisko and Maschi however subscribe to the idea that an ICA represents a unique form of analysis. An ICA, according to the authors, provides more attention towards "the contexts of communication and meaning making" whereas a qualitative content analysis focuses more on a summative aspect of data analysis. Here, the authors refer to Krippendorff who specifies that an ICA is:

> "A research technique for making replicable and valid inferences from texts (or other meaningful matter) to the context of their use" (Drisko & Maschi, 2015, p.59).

Krippendorff is quoted as saying that an ICA is more useful than other forms of content analysis when the unit of analysis is a rich understanding of the content's meaning, which applies to this analysis. Nonetheless, an ICA shall be grounded in empirical data according to the chapter (Drisko & Maschi, 2015). To comply with this requirement, the theoretical framework established in chapter two shall serve as a suitable foundation to interpret coded passages found in EU documents relating to freedom of expression in the digital sphere.

By going beyond describing what a document provides, an ICA is used in an attempt to establish the criteria of why, for whom and to what effect the analysed text prescribes (Drisko & Maschi, 2015), furthermore it:

> "may be used to inform, describe, evaluate, and summarise, as well as to provide a basis for advocacy and action." (Drisko & Maschi, 2015, p.67).

Hence, the approach of an ICA will be used to analyse the DSA to establish the boundaries of FoE in the digital sphere in the EU context. By also analysing the remaining documents, it will ultimately be established how these boundaries are interpreted, therefore how they should be implemented by service providers as well as how they potentially restrict users in their FoE. The choice behind the method is further motivated, as Drisko & Maschi provide that an ICA serves as a suitable method when direct access to original sources is limited or impossible, whereas in the case of this research, the documents do not specifically discuss the boundaries and their meanings of FoE. While access to the relevant documents is provided with the EUR-Lex service, interpretations and deeper discussions of FoE boundaries are not provided, calling for research into the topic to answer the related questions. Hence, to mitigate the issue of lacking information concerning the boundaries and corresponding meanings, an ICA will be conducted. The authors similarly note that texts often lack crucial information; here, this analysis attempts to bridge the missing information concerning FoE interpretations and how its boundaries are set in the digital sphere by conducting the ICA grounded in the theory established in chapter two, to finally conceptualise the understanding the of role of FoE in the EU context following the DSA entering into force. Notably, Drisko & Maschi enumerate that texts not only consist of meaning but receive meaning by having a perspective on it provided, furthermore, with interpretive research, personal bias can infringe on reliability and validity of the research (Drisko & Maschi, 2015). To mitigate these issues, it is aimed that by relying on academic research and theories established by other scholars, personal biases will be minimised in the interpretation process when for instance investigating the deeper meaning of underlying questions regarding hate speech or right violations.

In regard to the actual analysis in terms of coding the documents, the authors provide that in the coding process, an ICA focuses on both manifest and latent content including the context which allows understanding. Thereby, by applying connotative content categories, an ICA looks past just the words used by looking at the overall or symbolic meaning of the analysed text passage. Simultaneously, the analysis also explores the more manifest definitions of phenomena such as hate speech and disinformation, as well as scans the DSA for words such as "boundaries" or "limits" to establish what the documents entail in this regard. By illustrating both how the coding and the analysis were conducted, it is aimed to improve the research credibility (Drisko & Maschi, 2015). Section 2.6 connected the main findings of the theoretical framework to the presented research questions, from these emerge the codes, which are formulated in a way that enables them to summarise passages using descriptive narratives, as foreseen for ICA's by Drisko & Maschi (Drisko & Maschi, 2015), while also reflecting the theoretical insights gained in chapter two. Ultimately, given the nature of the formulated research questions, an interpretive content analysis serves as a perfect fit, as it has been established that contemporary research failed to provide necessary findings of an interpretive nature. Specifically, as contemporary research does not analyse the wording used in EU documents, but focus on a more general understanding of proposed instruments. Considering this, combined with the insights gained from Drisko & Maschi's work, the ICA will be conducted in six steps.

1. The first step of the analysis is the creation of the coding scheme that enables the ICA in the first place, where codes will be created based on the theoretical framework and in relation to how they could help deliver insights to the sub questions. These will be created taking both manifest and latent codes into consideration.

2. The documents will be scanned for the use of the code words in the manifest coding process. This will be done using the search function in Atlas.ti, which is a coding software that enables a more efficient way of coding documents such as those analysed in this research.

3. To avoid missing integral passages relevant to answering the research questions by for instance strictly focussing on the occurrence of a specific code word in the second step, the third step will focus on the latent coding procedure, which is more resource intensive. Nonetheless, by manually reading and coding the documents, it will be ensured that no relevant passages addressing the object of interest that may not have been directly mentioned have been missed.

4. In the fourth step, all codes and the instances where they have been utilised will be gathered to allow the actual interpretation process.

5. In the fifth step, the coded passages will be interpreted based on the findings established in the theoretical framework in chapter two.

6. In the sixth and final step, the findings will be illustrated in chapter four in a manner to answer the underlying sub questions. This will be done in sequence of the sub questions laid down in the first chapter.

To answer the first sub question, the DSA-FoE Boundaries code category and its respective codes will be used to identify how the DSA interprets FoE and if the boundaries as described in chapter two are identifiable. Additionally, it will be analysed what function the DSA ascribes to FoE by looking at the context of how it is utilised. The code Boundaries will be used to code for one, passages that directly mention the word boundaries or limit in relation to FoE boundaries, as well as to code passages that address the issue, even if these words directly occur. A manifest code of “FoE” will be used to code passages that directly reference FoE. Relevant here are not only passages that indicate where, for instance boundaries are placed, but further, what explicitly is presented as a boundary or expression that is placed outside those boundaries. In an effort to answer the second sub question and to investigate how the EU seeks to cope with abusive forms of FoE, the code categories of Abusive Expressions and Solutions will be utilised to first identify how the EU defines these abusive forms of expression to identify what expressions fall outside the scope of FoE and why, while also aiming to analyse the EU’s their interpretation against the findings of chapter two. Secondly, the Solutions codes will be used to explore how the EU aims to tackle them. Finally, to answer the third sub question the EU documents will be investigated for strategies on how AI is addressed to mitigate the challenges that have been highlighted in chapter two and in the answers to the first two sub questions. Based on the interpretive nature of this research, here the focus will lie on how AI is presented in the documents. Given this, the coding scheme in Table 2 has been created to perform an ICA, which finally allows the interpretation of the texts based on the given categorisations in chapter four (The coding scheme with additional examples of how a code from each code category was utilised can be found in the appendix).

Table 2: Coding Scheme

Category	Codes	Explanation
DSA-FoE Boundaries	Boundaries; FoE;	The codes will be utilised in regard to how the boundaries are interpreted and to code passages addressing FoE in general.
Abusive Expression	HS-Definition; DI-Definition; IC-Definition; Illegal Content; Hate Speech; Disinformation;	The codes will be utilised to code sections to identify how hate speech, disinformation and other forms of illegal content are conceptualised and described.
Solutions	AI-Solutions; HateSpeech-Solution; Disinformation-Solution; Illegal Content-Solution; AI-Content Moderation;	The codes will be utilised to identify the solutions to the respective challenges.

3.5 Preliminary Conclusion

Summarising the methodological approach of this thesis, after firstly having identified a research topic, the current state of the art was established to create a framework allowing the interpretation of the to be coded passages in chapter four's analysis. But most importantly, by informing about relevant concepts and dynamics in the field, it allowed the creation of the necessary codes and subsequent identification of relevant passages in the provided EU documents. Secondly, by performing an ICA consisting of coding EU documents dealing with FoE in the digital sphere, particularly with the DSA at the forefront, it is envisioned to be provided with relevant findings that allow to answer the underlying sub questions and present them in chapter four. Lastly, having provided a foundation to formulate an answer to the overarching research question of this thesis in the fifth and final chapter, it is expected to have delivered and presented insights into how the EU envisions a European digital sphere in regard the question of how FoE and its boundaries are interpreted and how this interpretation is envisioned to be executed.

4 Analysis

4.1 Introduction

Having performed the coding of EU FoE-related documents as laid out in the prior chapter, this chapter will provide an analysis of the relevant passages which have been identified in the coding process by applying a lens based on the findings from the theoretical framework. Section 4.2 to 4.5 will analyse how the DSA discusses the boundaries of FoE. In this process, it will be investigated how FoE boundaries ought to be placed in the digital sphere as well as what forms of expressions are discussed with a focus on how these are interpreted. With this, findings will be gathered that enable the answering of the first sub question. Section 4.6 seeks to investigate how the EU foresees to enforce this vision of FoE boundaries by looking at how the EU plans to combat abusive expressions, thereby providing the necessary findings to answer the second sub question. Section 4.7 will provide a focus on how the EU portrays the role of AI in this process in order to answer the final sub question. In total, this section aims to provide the necessary findings to paint a picture of how the EU interprets FoE in the digital sphere, thereby, section 4.8 will recapitulate the main findings of this chapter and formulate answers to each sub question. By firstly showing what the analysed documents discuss and secondly analysing these passages in an interpretive manner, this approach ultimately enables the presentation of patterns uncovered in the analytical process which will finally be used to answer the overarching research question in chapter five.

4.2 The Role of FoE: Between Preventing Domination & Superficiality

Before establishing how the EU foresees the boundaries of FoE in accordance with the DSA, it was investigated how FoE itself is discussed in an EU regulation dictating what type of content ought and ought not to be permissible. Thereby, it was aimed to extract a status the EU asserts to FoE within actual regulation by exploring the capacity in which FoE is directly addressed. In the search for passages codable in a manifest nature, the term “Freedom of Expression” was found to be occurring in a total of 18 instances, whereas the coding of the DSA amounted to 89 passages having been coded in total. In sum, the DSA clarifies how FoE should be taken into account when enacting its instruments such as when service providers are moderating content in accord with DSA guidelines, moreover the DSA stresses how safeguarding FoE is amongst its aims. To illustrate, Article 14 states how:

> “Providers of intermediary services shall act in a diligent, objective and proportionate manner in applying and enforcing the restrictions referred to in paragraph 1, with due regard to the rights and legitimate interests of all parties involved, including the fundamental rights of the recipients of the service, such as the freedom of expression, freedom and pluralism of the media, and other fundamental rights and freedoms as enshrined in the Charter.” (European Union, 2022, p.49).

This method of addressing FoE is repeated in each instance where FoE is mentioned either directly or when referencing European fundamental rights in their entirety, across all documents in this analysis, not just limited to the DSA (with one exception). By specifically naming FoE while only generally referring to other fundamental rights, one can infer that the EU acknowledges that FoE is particularly relevant in the context of the DSA, inviting deeper investigation. Notably, it is not only raised that FoE shall be safeguarded, but the 153rd provision of the DSA stresses that in regard to fundamental rights including FoE:

> “all public authorities involved should achieve, in situations where the relevant fundamental rights conflict, a fair balance between the rights concerned, in accordance with the principle of proportionality” (European Union, 2022, p.40).

For one, DSA instruments and its provisions translated into content moderation practices in the digital sphere are foreseen as being limited in their power by not transgressing the rights of users, especially

FoE, however, the second passage highlights that those rights are set to clash. In this regard, a nuance can be observed. The EU foresees the safeguarding of FoE and thereby presents service providers with obligations to adhere to, here however, it is kept ambiguous how service providers can in practice clearly achieve a balance in accordance with the principle of proportionality and safeguard fundamental rights.

In regard to how the DSA inserts FoE and its function, parallels to the findings of the theoretical framework are observable. By repeatedly stressing that the instruments of the DSA and the resulting actions of service providers shall not transgress the rights of citizens, the importance of fundamental rights and FoE is underlined. Moreover, unwarranted right transgressions are envisioned to be averted, mirroring Madison's view of preventing tyranny (John, 2019) by safeguarding FoE. In this case, tyranny could also take up the form of unwarranted content moderation, going far beyond tackling abusive expressions outside the boundaries of FoE and instead entering a territory of censorship reminiscent of more authoritarian views such as those by Locke. It can also be raised that while Restrepo used the notion of domination to establish FoE boundaries (Restrepo, 2013), the notion is reflected in limiting the powers of DSA instruments and obligations by repeatedly highlighting how FoE and fundamental rights shall not be transgressed. The aspect of rights-balancing however brings a nuance worthy of deeper investigation and will be addressed in the following section, here a key question will be which forms of expressions are deemed to infringe on the rights of others. Finally, from this passage and the initial act of passing the DSA itself, it emerges that the EU's vision of FoE deems it necessary to safeguard FoE by establishing requirements for service providers to adhere to, further reflecting Restrepo's view. Nonetheless, it can be concluded that the DSA inserts FoE as an attempted countermeasure to prevent right transgressions, yet it was also found that these assumed attempts do not go deeper than recurring reminders when addressing obligations laid upon service providers. This ultimately raises a question of the authenticity behind passages calling for FoE safeguarding, in particular against the backdrop raised by John, whereas safeguards of FoE laid down in law, do not pose as absolute safeguards preventing encroachments on FoE. Here, the analysis of the actual EU solutions related to content moderation will further explore the EU's vision, including whether these calls go deeper than mere (vague) mentions by word and are genuinely reflected in the foreseen instruments.

4.3 Manipulative Expressions: Overinclusive Definitions as Threats to FoE

To develop an understanding of the European vision of FoE in the digital sphere, as asserted by the DSA, it was investigated how the boundaries of FoE are actually discussed in the regulation. Contrary to the prior section, the DSA lacks passages discussing the boundaries of FoE by word, preventing manifest coding, requiring a deeper analysis. Therein, in the latent coding procedure, the Boundaries code was utilised to code 23 passages. Having identified how according to the DSA fundamental rights of different individuals including FoE have the potential to clash, a parallel to Restrepo's conceptualisation of democratic FoE and the notion of domination can already be established (Restrepo, 2013), as the boundaries of FoE are primarily asserted by rights of others. In this regard, the DSA lays out that a framework in the digital sphere ought to be created where FoE is interpreted in a way where content which trumps the rights of others can not be freely disseminated. To provide a full picture of this interpretation, however, it is necessary to investigate when expressions meet the parameter of domination according to the EU. Relatedly, chapter two's investigation of how the boundaries of FoE and potential right conflicts are discussed in academic research highlighted various concepts that, particularly in the digital sphere, are of relevance. Based on this, the question arises of where the concepts of misinformation, malinformation and disinformation, as raised by Ruiz (Ruiz, 2023), are placed in terms of the DSA's foreseen FoE boundaries and how they are interpreted. Looking at the act, it was established that the first two are not explicitly addressed in the DSA, while disinformation is mentioned in 13 instances. In spite of this, the DSA fails to articulate an understanding of what constitutes disinformation, instead it is raised how:

"Member States are increasingly introducing, or are considering introducing, national laws on the matters covered by this Regulation, imposing, in particular, diligence requirements for providers of intermediary services as regards the way they should tackle illegal content, online disinformation or other societal risks." (European Union, 2022, p.1).

In this and the twelve remaining instances the DSA addresses disinformation, disinformation is likened to a societal risk which ought to be addressed, suggesting that disinformation lies outside the boundaries of any protection under FoE. It remains unclear, however, what expressions are ultimately addressed as a clear definition is missing. The DSA references the Code of Practice (COP) on Disinformation explaining how it was extended, culminating in the Strengthened Code of Practice on Disinformation 2022. This COP, including one of its predecessors, the communication of the European Commission "Tackling online disinformation: a European Approach" provides more insights into the EU's interpretation of FoE and its boundaries in regard to disinformation. Interestingly, the communication raises how in the digital sphere, the issue of disinformation gains a new dimension as disinformation is disseminated:

"on a scale and with speed and precision of targeting that is unprecedented, creating personalised information spheres and becoming powerful echo chambers for disinformation campaigns.". Furthermore, it is stressed that disinformation "erodes trust in institutions and in digital and traditional media, and harms our democracies by hampering the ability of citizens to take informed decisions.", as well as "supports radical and extremist ideas and activities (…)" and "impairs freedom of expression", ultimately underlining how the EU views disinformation as "a major challenge for Europe" (European Commission, 2018, p.1).

This passage illustrates the gravity that the EU ascribes to the issue of disinformation in the digital sphere. Not only is the issue enabled by the nature of the digital sphere, but a view is shared whereas anybody suffers from the mere presence of disinformation as it undermines trust in institutions and has the potential to distort decisions in democratic processes therefore interfering with individual's FoE. This view directly resembles Restrepo's thoughts on deceptive expressions whereby such expressions constitute acts of domination, entertaining his concept of democratic FoE. Based on this, the EU (and Restrepo) view interference as necessary in order to balance conflicting rights as foreseen by the DSA. Similarly, the presented passage is inline with Ruiz view of disinformation as a disease for democracy (Ruiz, 2023). Contrarily, however, a notable difference is seen between the EU's vision of freedom (in the digital sphere) and Mill's perspective on FoE. With manipulative behaviours such as disinformation, the EU sees a threat that prevents freedom, requiring interference with the "open discourse" to establish a framework that enables to make use of rights such as FoE in stark contrast to Mill's notion of a discourse without interferences particularly in regard to opinions, even inventing artificial opinions, posing as the polar opposite of combating disinformation. Here however, one can not safely say whether Mill would also hold his view with the dynamics of the digital sphere in mind as raised in the prior passage. Conversely, Mill did raise a potential of democratic institutions being abused to undermine unpopular opinions (Cohen-Almagor, 2017), while Tropina equally noted how the regulation of disinformation may easily take up the form of undermining FoE (Tropina, 2023), inviting further analysis of the EU's approach and interpretation.

A missing definition for disinformation is found in the Commission's communication, whereas it is clarified that:

"Disinformation is understood as verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm." (European Commission, 2018, p.3).

Here it is enumerated that such threats to public harm relate to any threat to democratic or policymaking processes and public goods such as health, environment and security within the EU. Conversely, reporting errors, satire and parody as well as clearly identified partisan news and commentary do not constitute disinformation. The clarification regarding clearly identified partisan

news and commentary resembles Scanlon's 1979 liberal FoE view, whereas political speech shall be exempt from any interference (Restrepo, 2013). The communication does not address further how to address political speech in regard to FoE boundaries, however it is inferred that such speech is deemed permissible as it does not constitute disinformation and is referenced next to common forms of expressions such as satire and parody, nonetheless a clear distinction is missing, leaving relevant information ambiguous. Viewing the provided definition in the light of the findings of Tropina and Ruiz papers, an intricacy has to be pointed out. While the presented definition partly overlaps with the one established by Ruiz (Ruiz, 2023), the interpretation provided by the European Commission sets some arguably vague parameters, inviting an arbitrary restriction of FoE boundaries. After highlighting a material element, by addressing content which is verifiably false, a major deviation from Ruiz understanding is seen, as the Commission also includes misleading information in the definition.

Moreover, by stating how either economic gain or intentional deception are both in themselves qualifying criteria to determine if an expression constitutes disinformation, a larger pool of content would be classified as disinformation compared to Ruiz understanding. This deviation alone does not pose a serious issue, however the definition by the Commission would allow classifying content which is created, presented and disseminated for economic gain as disinformation if it is argued to be misleading. Here, however, a question of how to determine what constitutes misleading information is left unanswered and seems to be a highly subjective parameter, susceptible to potential abuse. This becomes critical in the light of Ruiz and Tropina's findings of how in the attempt to combat disinformation, the process seems to easily become arbitrary. Additionally even in regard to supposedly more concrete cases where information may seem verifiably false, both Tropina and Ruiz highlighted issues of where it is either impossible to fully verify any truth or consensus is quickly changing, whereas some information that once was understood as false, can become true or vice versa, as illustrated by Tropina raising the changing of WHO guidelines for facemasks (Tropina, 2023). In addition, the EU's provided understanding yet again contradicts the Millian view, whereas the passage provides that information or an opinion can be verifiably false, as opposed to Mill's view how one can never have absolute certainty in this regard, advocating against any inferences based on this notion (Mill, 1859). After defining disinformation, the communication sets the stage for how the EU interprets its genesis and explains that the dissemination of disinformation occurs in societies facing rapid change rife with economic insecurity and soaring extremism for instance, ultimately fostering preconditions for disinformation campaigns increasing underlying tensions. In this regard, the Commission enumerates how:

"The primary obligation of state actors in relation to freedom of expression and media freedom is to refrain from interference and censorship and to ensure a favourable environment for inclusive and pluralistic public debate. Legal content, albeit allegedly harmful content, is generally protected by freedom of expression and needs to be addressed differently than illegal content, where removal of the content itself may be justified. As the European Court of Human Rights has concluded, this is particularly important in relation to elections." (European Commission, 2018, p.1).

The beginning of the presented passage reflects the ideal of free and open discussions as pioneered by Mill), further underlined by stressing how harmfulness itself is not disqualifying content from being protected under FoE as does Mill and the general the call to refrain from censorship. However, the EU's vision of a favourable environment as opposed to Mill's, entails that interferences are inherently necessary, for instance by specifying that particularly in the context of elections, content removal may be justified. While Mill viewed that such an environment would bring about truth and that it shall be refrained from stifling opinions, it is established that in the context of elections, some opinions and given the focus of the documents and the portrayal of disinformation, some information, or opinions have to be removed. This is a direct contradiction of Mill's idea to even invent artificial opinions to further debate in the quest of bringing about truth (Mill, 1859). Also, paradoxically, this passage seems rather contradictory to the exemption of partisan news and commentary highlighted earlier, whereas this exemption seemed to coincide with allowing a broader scope of FoE in regard to political speech, in this case however, it is stressed that political speech and the political context surrounding

elections should receive particular attention and a more narrow scope of FoE, therefore contradicting Scanlon's 1979 view of political speech under liberal FoE. Furthermore, it remains unclear where a clear line is drawn in this context between legal and even "allegedly harmful content" and illegal content justifying removal.

Notably, with this passage, a question of whether a given expression lies within or outside FoE boundaries moves from a binary dimension to a non-binary one, by introducing a quasi-legal category which "requires a different approach" while still enjoying some protection (whereas here it is assumed that this passage addresses harmful content), in contrast to entirely illegal content that justifying complete removal. How to differentiate between harmful and illegal content remains unanswered, including in the Code of Practice on Disinformation of 2022. Therefore, no clarity is provided in regard to the concerns raised by Tropina and Ruiz thereby having delivered an unexpected finding that again a key issue in determining what constitutes a protected expression remains ambiguous, supporting the idea that FoE boundaries can be set arbitrarily under some circumstances as per Ruiz (Ruiz, 2023). Lastly, based on the hereby introduced ambiguity resulting from not addressing how to differentiate between categories, particularly in regard to what constitutes misleading information, it is unclear how to differentiate between a service provider acting according to the DSA by removing disinformation or engaging in actual censorship. The same conclusion can be drawn in regard to the issue of misinformation, as the code of practice defines it as:

> "false or misleading content shared without harmful intent though the effects can still be harmful" (European Commission, 2022, p.1).

Strikingly, the issue of ambiguity is exacerbated in this case by lowering the threshold to classify any given expression as misinformation by removing the parameters of deception and economic incentives. Here, the strengthened code further foresees that:

> "In order to limit impermissible manipulative behaviours and practices across their services, Relevant Signatories commit to put in place or further bolster policies to address both misinformation and disinformation across their services" (European Commission, 2022, p.15).

As the discussion of misinformation in EU documents in regard to the boundaries of FoE is limited to passages such as this one, the question of where in particular misinformation stands in the light of FoE boundaries is not explicitly answered. From this passage emerges, that misinformation and disinformation are impermissible manipulative behaviours, therefore not entirely protected, considering the EU invites signatories to tackle these issues. Approaches on how to tackle manipulative behaviours are not disclosed however, moreover it remains if impermissible manipulative behaviour differs from harmful content, or for instance is of an equal status as illegal content. Thereby, it was found that the EU establishes many categories of expressions without enumerating to what extent they are protected under FoE or how they deviate from one another. In sum, it can be seen how the COP is focussed on providing solutions to the laid out issues, a discussion of the relevant concepts lying outside its focus, therefore the document will become more relevant from section 4.6 onwards. Strikingly, it fails in establishing a clear understanding of the concepts it seeks to regulate. This section ultimately found that while at first a clear framework of FoE boundaries being asserted by the rights of others akin to Restrepo is foreseen by the EU, a combination of ambiguity and overinclusive definitions modifies this framework in a way that sets the stage for arbitrary interpretation of FoE in the digital sphere. What the EU deems manipulative expressions ought to be addressed as they undermine fundamental rights, by providing overinclusive definitions; however, the categorisation of any expression as manipulative is enabled. Combined with the obligation to combat abusive expressions, the vision provided by the EU is for one interpretable as an attempt to tackle disinformation while safeguarding genuine information, while also interpretable as enabling the infringement of FoE by restricting any expression in arbitrary manner, reflective of what Tropina and Ruiz criticised in their respective works.

4.4 (Il)legal & Harmful Expressions in the Digital Sphere: Blurry Boundaries

Having analysed the concepts of both mis- and disinformation as seen by the EU, necessary to portray how the DSA discusses and foresees the boundaries of FoE, the DSA itself can be analysed further. Next to expressions classifiable as for instance manipulative behaviours, Article 3 provides another category of expressions outside the boundaries of FoE, laying down expressions categorised as illegal content which in the context of the DSA refers to:

"information that, in itself or in relation to an activity, including the sale of products or the provision of services, is not in compliance with Union law or the law of any Member State which is in compliance with Union law, irrespective of the precise subject matter or nature of that law;" (European Union, 2022, p.42).

While content classified as illegal based on individual Member State law lies outside the scope of this research, a thorough understanding of the concept can still be developed by again consulting a wider range of EU documents. Firstly, found during the latent coding phase, page four and five of the DSA broadly define the concept of illegal content by referring to various applicable forms where the EU deems interference suitable. Besides an extensive list of information of content in various forms which is also illegal offline, the DSA names illegal hate speech, terrorist content, unlawful discriminatory content as well as, if not in itself illegal, information that "the applicable rules render illegal in view of the fact that it relates to illegal activities" (European Commission, 2022, p.4). Analysing these forms in sequence as presented, the term illegal hate speech is used in six instances in the DSA, however an explanation or definition is missing as in the case of disinformation. Similarly, a look at other EU documents provides necessary insights, as the Code of Conduct (COC) On Countering Illegal Hate Speech Online provides an agreed upon definition, whereas illegal hate speech refers to:

"all conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin." (European Commission, 2016, p.1).

On one hand, this definition partly reflects Mill's notion of instigation which he viewed outside the boundaries of FoE, whereas Cohen-Almagor added how since the advent of the expanding digital sphere, this notion becomes more relevant as instigation or incitement to violence as in the case of illegal hate speech, easily reaches people on a global scale (Cohen-Almagor, 2017). Moreover, the aspect of hatred towards a specific group reflects Cassim's definition of hate speech, who noted how such expressions ultimately undermine a person's rights (Cassim, 2015), which according to Restrepo and the notion of domination would also warrant an interference. On the other hand, and as previewed in the discussion of the role of FoE in the DSA, the code of conduct stresses that companies need to safeguard and consider FoE when tackling illegal hate speech. Here a particular nuance can be raised, as the COC stresses that besides ideas and information of inoffensive nature, FoE also protects "those that offend, shock or disturb the State or any sector of the population" (European Commission, 2016, p.1). This passage underlines an intricacy, whereas not only has it been established that the question of FoE boundaries is highly contested in academic discourse, but the issue of ambiguity of how to adhere to the DSA while enabling fundamental rights, as raised in the prior sections, is exacerbated. Particularly when having to distinguish between an expression that is shocking, disturbing, or offensive to a sector of the population versus an expression which constitutes hatred directed against a group of persons, whereas a given expression may be interpretable as either one of these categories, requiring further guidance on how to differentiate between those, which however is missing.

Illustrating this ambiguity, from a view of hate speech undermining rights of others as shared by Cassim for instance, the definition may be interpretable as allowing the tackling of threats or insults and thereby being aimed at safeguarding fundamental rights. Conversely, by not establishing what expressions constitute hatred for instance, any given critical expression may be conceivable as hatred as raised by Strossen in regard to the mourner's veto, instead allowing to suppress controversial

expressions (Bromell, 2022). Looking past this ambiguity, by stressing how offensive views are generally protected under FoE, the EU does at least to some extent, share Madison's view and valuation of plurality, acknowledging views which are not widely shared, as opposed to explicitly advocating for stifling unpopular or offensive information. Here, however, the framework established by the EU arguably allows for diversions from what is proclaimed in the COC. Ultimately, the DSA's 12th provision, lays down that forms of expression applicable under the definition for illegal hate speech are out outside the boundaries of FoE and represent illegal content, a difficulty remains in deciding between tackling those and safeguarding FoE in regarding forms which "stretch the boundaries" of FoE. Relatedly, the COC on Illegal Hate Speech further adds that:

"The spread of illegal hate speech online not only negatively affects the groups or individuals that it targets, it also negatively impacts those who speak out for freedom, tolerance and non-discrimination in our open societies and has a chilling effect on the democratic discourse on online platforms." (European Commission, 2016, p.1).

With this passage, it can be seen that the negative effects the EU foresees regarding illegal hate speech are virtually the same as with disinformation, moreover they reflect the concerns provided by Cassim in regard to the effects of hate speech. Additionally, an argument reflective of Restrepo's notion of domination is again provided, whereby regulating content in the digital sphere is justified in a 'circular' manner as according to the EU, if uninterfered with, a scrutinised form of expression itself would undermine FoE. Lastly, by specifically addressing 'illegal hate speech', it is implied that according to the EU, hate speech itself is not automatically unprotected under FoE, reflective of the categorisation by Frank La Rue (Bromell, 2022). However, it is not laid out how to differentiate between legal and illegal hate speech or if this category is even considered by the EU. In sum, with the DSA, in regard to illegal hate speech, the EU subscribed to the view of applicable expressions violating rights and posing risks, justifying interference reflecting Restrepo's view of democratic FoE, however a difficulty arises in regard to having to safeguard FoE in itself as it is unclear where to objectively draw a line between (illegal) hateful and (legal) disturbing or offensive information.

Next to illegal hate speech, terrorist content was also addressed in the discussion of illegal content, here, the EU introduced the Regulation on Addressing the Dissemination of Terrorist Content Online (RADTC). The RADTC joins the by now established trend of not specifically defining crucial terms, as no definition for terrorist content is provided. However, the regulation does provide numerous indicators allowing the development of an understanding of the concept, in light of FoE boundaries. Hereby, the RADTC calls for establishing a definition that encompasses:

"material that incites or solicits someone to commit, or to contribute to the commission of, terrorist offences, solicits someone to participate in activities of a terrorist group, or glorifies terrorist activities including by disseminating material depicting a terrorist attack." (…)"material that provides instruction on the making or use of explosives, firearms or other weapons or noxious or hazardous substances, as well as chemical, biological, radiological and nuclear (CBRN) substances, or on other specific methods or techniques, including the selection of targets, for the purpose of committing or contributing to the commission of terrorist offences. Such material includes text, images, sound recordings and videos, as well as live transmissions of terrorist offences, that cause a danger of further such offences being committed." (European Union, 2021, p.3).

Subsequently, it is expressed that material disseminated for the purpose of education, journalism, art, or research as well as to raise awareness are not classified as terrorist content and would therefore be placed within the boundaries of permissible expressions. Similarly to the case of illegal hate speech, it is reminded how:

"the expression of radical, polemic or controversial views in the public debate on sensitive political questions should not be considered to be terrorist content" (European Union, 2021, p.3).

While the notion of incitement and solicitation may arguably also allow vague interpretations, more details necessary to meet the parameters to be classifiable as terrorist content are provided (compared to illegal hate speech and disinformation). Earlier, it was found how Scanlon's view of political speech is partly reflected, yet also contradicted among analysed passages. While for one, it was stressed how in the context of elections, interferences may be justifiable, the COP on disinformation raised how partisan news or commentary are protected from interferences. Here, the latter notion is again entertained, in regard to radical views for instance, fuelling the ambiguity of where a line is ultimately drawn and what content is protected under FoE. Similarly, ambiguity can arguably be found between protected radical political views and unprotected glorification of terrorist activities. In regard to scientific discourse encaptured in chapter two, only the views by the most libertarian scholars such as Dworkin and Nagel would potentially place the described terrorist content within the boundaries of FoE based on a view where the act of restricting FoE is inherently wrong even when societal benefits would be greater (Cohen-Almagor, 2019). Still, an intricacy can be raised, whereas Article 1 of the regulation raises how an:

"assessment shall determine the true purpose of that dissemination and whether material is disseminated to the public for those purposes" (European Union, 2021, p.11).

Firstly, at no point in the regulation is it enumerated by whom or how the assessment would be conducted. More notably however, with this assessment, the RADTC introduces the notion of a given entity having to act as an arbiter of truth, which Tropina raised as being problematic given a private company's interests being more akin to satisfying stakeholders than upholding fundamental rights (Tropina, 2023). Notably, the passage stresses that the true purpose of an expression will be determined, whereby the true purpose behind an expression does not lie with its actual originator, but somebody else. Lastly, with the role of an arbiter of truth, the boundaries of FoE arguably depend on the acting entity.

Next to terrorist content, the category of unlawful discriminatory content was raised in the DSA, this category however is mentioned only in two instances across the analysed documents and neither provide information on what this category entails (in the DSA and AIA respectively), preventing an in depth analysis of it. Moreover, one could assume that according to the DSA not all discrimination would be illegal based on explicitly addressing unlawful discrimination, the RADTC however states how any discrimination is prohibited, making this distinction redundant, as well as withholding a definition for discriminatory content (European Union, 2021). The final category referenced in the initial passage is the category of content relating to illegal activities. While naming all relevant types of content is beyond the scope of this research, the DSA raises "illustrative examples" (European Union, 2022, p.4), highlighting content regarding child sexual abuse or non-consensual sharing of photos, as well as unlicensed use of copyrighted material. Here, it is explained that for instance merely depicting an illegal act as in the case of eyewitness videos, content would not automatically be classified as illegal, given the act of recording itself was not illegal. Relatedly, page 24 also adds that content in the form of cyber violence and illegal pornographic content is also outside the boundaries of FoE (European Union, 2022). These examples again underline how FoE boundaries are set in the attempt to protect rights of others, however, cyber violence is yet another category addressed by in the DSA without having a definition provided. Another relevant document to the discussion of illegal content is provided by the Council of the European Union. In the EU Human Rights Guidelines on Freedom of Expression Online and Offline, the Council establishes forms of expressions outside the boundaries of FoE, worthy not only of content removal but even warranting criminal punishment. Besides foreseeing that member states outlaw the already discussed form of content constituting illegal hate speech, the guidelines add:

"publicly condoning, denying or grossly trivialising crimes of genocide, crimes against humanity and war crimes (...) against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin when the conduct is carried out in a manner likely to incite to violence or hatred against such a group or a member of such a group" (Council of the European Union, 2014, p.27).

Here again, the boundaries between a legal controversial view on a sensitive political question and an illegal trivialisation of the listed crimes, remains blurry. Some guidance on the distinction between legal and illegal hate speech is provided however, whereas, based on European Court of Human Rights case-law, a distinction is made between hate speech that can "negate the fundamental values of the Convention" or "not apt to destroy the fundamental values of the Convention" (Council of the European Union, 2014, p.23). Yet it is not raised how one can explicitly distinguish between what is assumed to be legal and illegal hate speech, particularly in regard to the DSA and how service providers should determine whether European values are destroyed based on a given expression. These findings are again in line with the concerns raised by Tropina regarding an entity acting as arbiters of truth (Tropina, 2023). Notably, if the key distinction is only determinable by the European Court of Human Rights and is not in the realm of decisions made by service providers, a question has to be raised how service providers could be able to correctly moderate content when having to decide whether an expression is placed within or outside the boundaries of FoE. As in prior analysed documents, the guidelines highlight the notion of expressions that, figuratively speaking, stretch the boundaries of FoE, but enjoy full protection, specifying that FoE:

"includes the freedom to express and impart information and ideas of all kinds that can be transmitted to others, in whatever form, and regardless of media. Information or ideas that may be regarded as critical or controversial by the authorities or by a majority of the population, including ideas or views that may "shock, offend or disturb", are also covered by this." Here it is added that these can include "Commentary on one's own or on public affairs, canvassing, discussion on human rights, journalism, scientific research, expression of ethnic, cultural, linguistic and religious identity and artistic expression, advertising", including "political discourse and advertising during election campaigns. " (European Union, 2014, p.4).

On the other hand, these protected expressions are directly opposed with the definition of illegal hate speech as well as:

"denial or gross trivialisation of certain international crimes when carried out in a manner likely to incite to violence or hatred" and "genuine and serious incitement to extremism" (European Union, 2014, p.17).

This passage further underlines the issue of blurry boundaries, as well as the ambiguity concerning the breadth of protection surrounding political speech. Lastly, the guidelines provide additional insights regarding interferences with FoE, reflecting notions of both Restrepo and Mill, whereas it is enumerated how:

"No person may be subject to the impairment of any rights on the basis of his or her actual, perceived or supposed opinions.", furthermore, "All forms of opinion are protected, including opinions of a social, political, scientific, historic, moral and religious nature. States may not impose any exceptions or restrictions to the freedom of opinion nor criminalise the holding of an opinion" (Council of the European Union, 2014, p.3).

In the first part, Restrepo's notion of domination is reflected yet again, this time in regard to FoE itself, not only as a boundary to one's FoE. In the second part, Scanlon's 1972 notion of sacrosanctity of any opinion is entertained. However, as Mill points out himself, once an opinion is expressed, it begins to affect others. Still, his viewpoint foresees that resulting harm from expressions only serve as a justification in the most extreme cases (Mill, 1859). Here, the guidelines go on to discuss boundaries of FoE, beginning with yet another statement reflective of Restrepo's conceptualisation of democratic FoE regarding accountability accompanying the exercise of one's own rights (Restrepo, 2013). It is stated how:

"The exercise of the rights (…) carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are

necessary: (a) For respect of the rights or reputations of others; (b) For the protection of national security or of public order (ordre public), or of public health or morals" (Council of the European Union, 2014, p.3).

From this passage emerges that any restriction foreseen in the DSA is ultimately necessary, conversely, it is assumed that any of the expressions outside FoE boundaries have the potential to undermine the affected areas. It can be noted however that the issue of ambiguity is reflected here again by providing an empty parameter in the form of protecting public morals, which can be arbitrarily filled by a responsible entity. Such an entity would have to act as an arbiter of truth by assuming that a given set of morals is held by the public and ascribe it to the public, ultimately inviting the arbitrary setting of FoE boundaries depending on how a responsible institution is interpreting this passage and the given set of public morals. Moreover, based on the analysed documents a clear-cut understanding of what public morals are according to EU vision (of the digital sphere) can not be provided. In sum, this section found that to fully grasp the boundaries of FoE as foreseen by the DSA, one is required to investigate a number of substitute EU documents, potentially undermining the DSA's goal of establishing legal certainty, particularly in regard to the recurring issue of ambiguity concerning distinct concepts and their placement within or outside FoE boundaries. In the light of this ambiguity, the persistent notion of requiring an entity to act as an arbiter of truth is most striking, illustrated for instance by the foreseen external assessment in regard to terrorist content determining the true purpose of an expression. Moreover, this section further underlined how a general framework reflective of democratic FoE is provided, however in regard to how relevant abusive forms of expressions are discussed, the EU vision provides an understanding that allows interpretations of all kinds, ultimately providing ambiguity concerning what the actual EU vision is.

4.5 FoE: Flexible Boundaries & Systemic Risks

Having explored the concept of illegal content as raised in the DSA and adjacent EU documents, two additional nuances in regard to FoE boundaries in the digital sphere can be highlighted. Both of these notions are of interest as contrary to the assumption held until now, FoE boundaries as laid down in the DSA are more flexible than rigid in two distinct ways.

The first of these was identified in the latent coding process, in Article 36 of the DSA specifically, which sets up a crisis response mechanism. Once invoked by the Commission, the mechanism can obligate service providers to initiate a crisis response. Here, the DSA clarifies that this includes:

"adapting content moderation processes and increasing the resources dedicated to content moderation, adapting terms and conditions, relevant algorithmic systems and advertising systems, further intensifying cooperation with trusted flaggers, taking awareness-raising measures and promoting trusted information and adapting the design of their online interfaces" (European Union, 2022, p.25).

Thereby, it is provided that if the Commission declares a crisis, it is able to initiate a response where content is moderated in a manner which decreases any risk related to a crisis. However, against the backdrop of Tropina and Ruiz, particularly the findings concerning the corona pandemic, this mechanism invites speculation concerning its susceptibility to be used to arbitrarily set the boundaries of FoE. Especially as the Commission, upon recommendation by the European Board for Digital Services can decide if a crisis response is required, thereby the sole decision whether to modify FoE boundaries lies entirely within EU institutions. Ultimately, if it is determined that:

"extraordinary circumstances occur that can lead to a serious threat to public security or public health in the Union or significant parts thereof" (European Union, 2022, p.25),

the mechanism can be initiated. Notably, the prior sections highlighted how the mere existence of disinformation or illegal hate speech are seen as immense risk to the public, giving rise to whether the mechanism in question may be usable arbitrarily and ultimately allow the EU to determine what

content is permissible in a given situation, and thereby act as an arbiter of truth. Here, Provision 108 adds that the Commission may also create voluntary crisis protocols, while service providers are not obliged to implement these contrarily to the initial crisis response, doing so would further allow the Commission to set the boundaries of FoE with regard to a circumstance where it deems that rapid spread of illegal content or disinformation is occurring. The DSA proposes these protocols in situations:

"where the need arises for rapid dissemination of reliable information" (European Union, 2022, p.30).

This passage further underlines the provided concerns as it suggests that the Commission may interfere with the lesser regulated information exchange and supplies its own, favoured information. Surprisingly this mechanism lays the groundwork for what Ruiz observed during the corona crisis with private firms arbitrarily moving FoE boundaries based on influence by a governmental institution, whereas any remedies to combat the crises and related disinformation run the risk of undermining FoE (Ruiz, 2023) despite the DSA stressing that a crisis response shall take into account fundamental rights.

The final notion to be analysed concerning the EU's interpretation of FoE boundaries in the digital sphere is the notion of systemic risks. Provision number 79 of the DSA stresses that VLOPs and VLOSEs can strongly influence safety online or even shape public opinion discourse thereby sharing the vision of Hohmann & Kelemen raised in chapter one, whereas online platforms can act as gatekeepers (Hohmann & Kelemen, 2023). To mitigate this, the DSA foresees service providers to address four systemic risks. Those related to the dissemination of illegal content, those directly related to the issue of fundamental rights, including abusing platform mechanisms to silence speech, those related to democratic processes or public security and lastly those related to public health, for instance based on the way a service is built up, similar to risks from disinformation campaigns. However, a final intricate notion is highlighted whereas the DSA also foresees that:

"When assessing the systemic risks identified in this Regulation, those providers should also focus on the information which is not illegal, but contributes to the systemic risks identified in this Regulation. Such providers should therefore pay particular attention on how their services are used to disseminate or amplify misleading or deceptive content, including disinformation" (European Union, 2022, p.23).

With this passage, the notion of expressions not being clearly categorizable is again underlined. In this case, it is difficult to interpret and categorise related expressions in the dichotomy of within or outside FoE boundaries, as a category of legal expressions which nonetheless shall receive attention by service providers is introduced. What information in particular shall be addressed is again left ambiguous, intensifying the intricacy of this categorisation and the general finding of a framework that provides the possibility for ambiguous interpretation. Ultimately, given the view of expressions being either within or outside FoE boundaries, this finding comes as a surprise, especially as yet another form of ambiguity is introduced by the DSA. Moreover, what constitutes sufficient addressing by service providers is not enumerated, fuelling the ambiguity concerning where information contributing to systemic risks stands in regard to FoE boundaries even more. In sum, this section found that not only can the Commission actively influence "the marketplace of ideas" by obligating undisclosed forms of content moderation in cases where a crisis is declared, but further, it was found that the DSA provides categories of expressions that prevent a complete illustration of where the EU foresees to set the boundaries of FoE in the digital sphere.

4.6 Combating Abusive Expressions: A Framework of Ambiguity & Autonomy

By now, it has been established that the EU views expressions classifiable as disinformation, illegal hate speech and other various types of illegal content as immense risks for society which need to be combated as they undermine fundamental rights of others. Chapter one already previewed the major

instruments of the DSA, this section will now provide an in-depth analysis of those focused on combating abusive forms of expressions, including instruments derived from the remaining EU documents, thereby establishing how the EU aims to cope with the prior analysed issue. In the analysis, three overarching approaches related to combating abusive expressions have been identified. The first approach is characterised by vague laying out what ought to be addressed, thereby making service providers aware of a potential issue, which are then provided with considerable autonomy on how to deal with said issue. Here, requirements are restricted to communication and transparency about what intervention was eventually designed by service providers. A second approach deals with the EU attempting to provide an anchor of trusted information or sources in the digital sphere; these however run a risk of entities acting as an arbiter of truth, as for instance criticised by Tropina (Tropina, 2023). The third approach deals with the potential drawbacks that may arise when combating abusive expressions in the digital sphere, whereas potential infringements of fundamental rights, with FoE at the forefront, are attempted to be mitigated with distinct countermeasures. Each of these approaches is ultimately related to a given form of content moderation, which the DSA defines as:

"the activities, whether automated or not, undertaken by providers of intermediary services, that are aimed, in particular, at detecting, identifying and addressing illegal content or information incompatible with their terms and conditions, provided by recipients of the service, including measures taken that affect the availability, visibility, and accessibility of that illegal content or that information, such as demotion, demonetisation, disabling of access to, or removal thereof, or that affect the ability of the recipients of the service to provide that information, such as the termination or suspension of a recipient's account" (European Union, 2022, p.43).

This definition highlights multiple relevant aspects, besides mentioning the possibility of utilising AI to combat abusive expressions, it is added how FoE boundaries themselves are not the only factor influencing what expressions ought to be permissible, as compatibility with terms and conditions also influences whether content is subject to an interference. Moreover, this passage also entertains the established notion of quasi-legal content which does not warrant legal consequences, however according to the EU's vision this category has to be addressed in some form. In regard to the first approach and the corresponding autonomy for service providers referenced earlier, many passages of the DSA revolve around ascribing responsibility to service providers concerning aspects which have to be addressed, in particular those relevant for VLOPs and VLOSEs. To illustrate, looking at the notion of systemic risks and the final passage raised in section 4.5, the DSA draws attention towards a specific risk and potential contributors, such as in Provision 79 whereby service providers:

"should take appropriate mitigating measures in observance of fundamental rights." (European Union, 2022, p.22).

Even in cases where more guidance is provided, the DSA tends to ascribe a high degree of autonomy for how to achieve a given obligation, such as in the form of opting to implement COCs, therefore providing a form of self-regulation (European Union, 2022). COCs are proposed for systemic risks such as manipulative behaviours as well as illegal content (European Union, 2022), therefore, a noticeable amount of trust is laid upon service providers, as they are trusted with dealing with issues which within the EU's vision of the digital sphere constitute risks of a systemic scale. From this emerges that the DSA foresees that addressed service providers need to interfere with the exchange of expressions in an attempt to filter out those which are abusive and undermine rights of others according to the EU's vision, while also safeguarding FoE of those potentially affected. A further overlap between the EU's vision and Restrepo's concept of democratic FoE is underlined here, as it is provided how interferences are required to enable the exercise of fundamental rights including FoE (Restrepo, 2013). Provision 84 draws attention towards one aspect deserving of interference, particularly, towards the way a service is built up and its susceptibility for coordinated manipulation. From this emerges that according to the EU's view, the digital sphere itself is prone to amplify the risks of abusive expressions, by enumerating how:

"such risks may arise, for example, through the inauthentic use of the service, such as the creation of fake accounts, the use of bots or deceptive use of a service, and other automated or partially automated behaviours, which may lead to the rapid and widespread dissemination to the public of information that is illegal content or incompatible with an online platform's or online search engine's terms and conditions and that contributes to disinformation campaigns." (European Union, 2022, p.23)

Applying an analytical lens to the notion of assessing systemic risks, it invites a constant state of caution surrounding what content is shared on a platform, for one, such caution could be helpful in addressing actual threats such as incitement to violence, relating to the committing of crimes and terrorism. On the other hand, it may invite an excessive amount of precaution leading to interferences with for instance those opinions which can be categorised as offensive, or disturbing, ultimately leading to FoE infringements. Here, however, the DSA strictly stresses how monitoring obligations are not foreseen (European Union, 2022), decreasing the risk of the latter interpretation. Nonetheless, the DSA predominantly views the digital sphere as an inherent risk factor, as opposed to more optimistic ideals such as envisioned by Musk raised in chapter one (PR Newswire, 2022). This fear of risks is also extended, reflected in the discussion using AI when combating abusive expressions, painting a picture of seeing AI as a double-edged sword. Hereby AI is for one discussed as a potential solution, yet simultaneously raised as a risk source in itself, an image which will be deepened in the analysis. Here, Article 23 (Measures and protection against misuse) specifies how:

"When conducting risk assessments, providers of very large online platforms and of very large online search engines shall take into account, in particular, whether and how the following factors influence any of the systemic risks referred to in paragraph 1: (a) the design of their recommender systems and any other relevant algorithmic system; (b) their content moderation systems; (c) the applicable terms and conditions and their enforcement" (European Union, 2022, p.57).

In this regard, Article 35 adds that any algorithmic systems have to be tested and if necessary adapted in order to mitigate systemic risks (European Union, 2022), further underlining a general level of caution in regard to using AI as a potential tool. Once again however, detailed requirements are withheld ultimately leaving the question of what measures are deemed appropriate and to safeguard fundamental rights with service providers, against the concerns raised by Tropina criticising the capability of private companies in sufficiently safeguarding fundamental rights (Tropina, 2023).

A similar approach to regulating the digital sphere is provided in the RADTC, surprisingly, it was found that the aforementioned level of autonomy is also foreseen when combating terrorist content. Relatedly, the view of the digital sphere as being inherently risk-prone based on its susceptibility for abuse is highlighted on the first page of the regulation. Here it is stated how:

"services of hosting service providers are in certain cases abused by third parties for the purpose of carrying out illegal activities online. Of particular concern is the misuse of those services by terrorist groups and their supporters to disseminate terrorist content online in order to spread their message, to radicalise and recruit followers, and to facilitate and direct terrorist activity." (European Union, 2021, p.1).

Based on this vision, the regulation foresees service providers to play part in protecting public security while simultaneously having to construct "appropriate and robust" safeguards that protect fundamental rights, including FoE (European Union, 2021, p.2). As discussed, service providers are provided with autonomy when deciding on what measures meet these criteria while being able to identify and remove terrorist content. Once again underlining the notion of trust put onto the service providers, however accompanied by Tropina's critical account of private companies being tasked with safeguarding FoE, while their primary responsibility lies with satisfying their stakeholders (Tropina, 2023). On the other hand, a fear of (financial) punishment may incentivise attempts to safeguard FoE and other rights, however, this fear may also lead to infringements of FoE when not having laid out a

clear path to achieve the regulation's objectives. For instance, if this fear leads to platforms opting to implement overly restrictive mechanisms, ultimately undermining user's rights. With the high degree of autonomy service providers are trusted with, it can be concluded that contrary to the concerns by Tropina, the EU does see private companies well-equipped to deal with the challenges as raised in the analysis until now. Alternatively, a possible outcome arises whereas prior sections identified an issue with the EU providing a framework that enables ambiguous interpretations based on overinclusive definitions, for instance. Here, the issue of ambiguity is combined with autonomy on the part of service providers, thereby exacerbating the initial issue. To illustrate, in order to establish a well-rounded mechanism that addresses abusive expressions while safeguarding FoE, one arguably requires clearly identifiable categories that allow distinguishing between what is permissible and what is not. With the established combination of autonomy and ambiguity, different interpretations of what for instance is appropriate and able to satisfy both ends are possible, which ultimately can come at the cost of undermining FoE or for instance not sufficiently addressing terrorist content in the case of the RADTC.

Regardless of the provided ambiguity, section 4.4 laid down how the DSA categorised terrorist content as illegal content, whereas content considered to be illegal ought to be expeditiously removed, or access to it being disabled within 24 hours after being made aware of it, as foreseen by the DSA (European Union, 2022). Notably, according to the RADTC, if service providers however receive a removal order concerning terrorist content (Article 3), the RADTC requires this process to be undertaken within one hour of having received the order, moreover, it is foreseen that the related data is stored for six months for the purpose of further investigation. Generally, non-compliance with the regulation can lead to penalties and is seen as a necessity for effective implementation according to the RADTC (European Union, 2021) Given these narrow time frames and a looming threat of punishment a question arises concerning potential safeguards for miscategorisation and unwarranted interferences in case of miscategorisation. Section 4.2 however established that despite repeated calls to safeguard FoE, predominantly within the DSA, information on how to achieve this is rarely provided (across all analysed documents). One exception (as addressed in 4.2) to this pattern is provided in the RADTC, which specifies that:

> "Complaint procedures constitute a necessary safeguard against the erroneous removal of or disabling of access to content online where such content is protected under the freedom of expression and information." (European Union, 2021, p.7).

DSA Article 20 (Internal complaint-handling system) also foresees such a mechanism, requiring that any interference ought to be explained to the affected users, as well as enabling users to lodge a complaint against such an interference. Provision 58 states how:

> "Recipients of the service should be able to easily and effectively contest certain decisions of providers of online platforms concerning the illegality of content or its incompatibility with the terms and conditions that negatively affect them" (European Union, 2022, p.15).

These passages can be interpreted in two distinct ways, again underlying the intricacy concerning ambiguity in the analysed passages. For one, the mechanisms can be interpreted as providing users the possibility to defend themselves from arbitrary interferences and potentially overturn decisions that undermine their FoE. On the other hand, given that the discussion of FoE safeguards in a more explicit manner is severely lacking, the passage invites an interpretation whereas the protection of FoE in the digital sphere mostly plays out in a manner where a user is assumed guilty and has to defend themselves for making an expression. Notably, Provision 58 provides that the aim of this mechanism is a non-arbitrary or fair outcome, whereas the process itself is not addressed, which in turn raises questions concerning the fairness of the initial process. For one, if an arbitrary decision to stifle an opinion for instance is taken back, FoE has already been undermined which in political contexts as similarly raised by the EU across documents, may have huge influence over elections and individual's ability to make informed decisions. Here, users are ultimately envisioned to justify making use of their rights in the digital sphere. Assuming any arbitrary interference is in fact being

overturned, this interpretation still deviates from the definition of FoE laid down in CFR Article 11, foreseeing no interference when holding, seeking, receiving or imparting information (Goddard, 2021). Additionally, when considering the issue of ambiguity concerning where to draw a line between illegal and shocking content for instance, the problem is exacerbated, providing a considerable amount of blurriness between the fronts of safeguarding FoE and combating abusive expressions as obligated by the EU, especially as Article 23 dresses to suspend users who frequently spread illegal content. With this, a question arises how the EU envisions to prevent arbitrary interferences in prospect. The latent coding process allowed the identification of a limited number of such mechanisms.

One applicable requirement is provided in the RADTC, which also provides insights for the question of how the EU envisions AI to combat abusive expressions. Provision 23 states how:

> "hosting service providers should act with due diligence and implement safeguards, where appropriate, including human oversight and verifications, to avoid any unintended or erroneous decision leading to the removal of or disabling of access to content that is not terrorist content." (European Union, 2021, p.5)

Thereby, FoE ought to be protected by preventing AI from falsely classifying content as illegal information. Again, however, the notion of ambiguity and autonomy interplay in this requirement. Firstly, accountability is shifted to service providers by delegating whether AI is a suitable and appropriate tool to the service providers, here however, it may be difficult to determine appropriateness without specific parameters and especially the issue of how to clearly distinguish between illegal expressions and those which stretch FoE boundaries. Moreover, feasible alternatives are never addressed or proposed by the EU, leaving little evidence that the EU seriously considers those. With this, it is highlighted how AI itself, in terms of fully automated decision-making, ought not to be trusted according to the EU, requiring human oversight. Notably, erroneous decisions would potentially have tremendous repercussions besides just FoE infringements, such as in cases where authorities are alerted and quickly respond to assumed terror threats. The RADTC continues by requiring service providers to report any use of AI to competent authority, including any measures ensuring erroneous decisions for instance do not occur, allowing oversight and potential scrutiny. If an assessment (by the authority) determines that the reported rules are insufficient, the DSA obligates improvements; however, the regulation does specify that in this case no monitoring or an implementation of AI is foreseen (European Union, 2021). Additionally, it can be added that the DSA foresees to suspend users who abuse flagging mechanisms, whereas unfounded flagging would constitute an attempt of domination, whereas with this approach incentives to act as a watchdog on online platforms are removed, in relation to what was raised as the mourner's veto in Bromell's book whereas some may attempt to undermine reasonable debates on the basis of emotional appeals (Bromell, 2022). A different dimension of safeguarding FoE is based on the EU's view of manipulative or hateful expressions themselves undermining FoE, here, the DSA foresees the (ability for) participation of users in the effort to combat abusive expressions, whereas Article 16 establishes a mechanism allowing anybody to notify the service providers of suspected illegal content and therefore ultimately aid in safeguarding FoE as viewed by the EU.

Roughly 10 years before the DSA entered into force, the Human Rights Guidelines on FoE presented by the Council of the European Union also discussed the issue of safeguarding FoE in the attempt to combat abusive expressions in the digital sphere. Yet, here, a mostly general call to adhere to fundamental rights reflective of the findings in section 4.2 is provided. In spite of highlighting the need to fight illegal hate speech, it is stressed for example how:

> "Hate speech legislation should not be abused by governments to discourage citizens from engaging in legitimate democratic debate on matters of general interest" (Council of the European Union, 2014, p.17).

The Council foresees that in its quest to combat abusive expressions, hate speech legislation does not prevent legitimate democratic debate, here it is unclear however what differentiates legitimate democratic debate from illegitimate debate for instance, especially where the exact boundary would be, again in the light of ambiguity between offensive and illegal expressions. A similar notion is entertained in regard to defamation laws not being abused in order to censor criticism surrounding public issues. The guidelines provide another vague yet grandiose promise, although it misses concrete resolutions to safeguard FoE, advocating against infringements by promising how the EU will:

"Advocate for the application of all human rights, including the right to freedom of opinion and expression, both offline and online." (Council of the European Union, 2014, p.9).

With this call, in combination with providing not to abuse hate speech legislation, the leitmotif of ambiguity and autonomy can be illustrated once more. It is unclear how, based on the provided definitions and instruments, rights can be safeguarded, including FoE, while combating abusive expressions which in the European vision represent right transgressions if being expressed. Furthermore, adding to the findings of section 4.2, FoE is seen as an essential foundation for democracy or rule of law and valued as essential for development of individual's identities, here, the guidelines stress how:

"Free, diverse and independent media are essential in any society to promote and protect freedom of opinion and expression and other human rights. By facilitating the free flow of information and ideas on matters of general interest, and by ensuring transparency and accountability, independent media constitute one of the cornerstones of a democratic society. Without freedom of expression and freedom of the media, an informed, active and engaged citizenry is impossible." (Council of the European Union, 2014, p.1).

This passage highlights a free flow of information thereby is directly at odds with the DSA and particularly the 2022 COP on disinformation, which foresee interference with the flow of information by boosting specific "established" or "trusted" media forms, while decreasing the visibility of content flagged as disinformation (European Commission, 2022) as the analysis will further show. Instruments of informational warfare such as bots or manipulation require interference according to the EU, however here a free flow of information clashes with an idea of genuine or authentic information marketplace. With regard to the theoretical framework, the quoted passage reflects Mill's conception with a scope of FoE boundaries as broad as possible (Cohen-Almagor, 2017), simultaneously the EU argues for a repeated need to interfere with the flow of information and tighter boundaries as opposed to Mill. From this it can be inferred that according to the EU's view in order to achieve a free, diverse and independent media interferences are a necessity.

In light of this, the DSA addresses some notions not thoroughly touched upon in the theoretical framework. This includes the issue of what the DSA deems dark patterns:

"Dark patterns on online interfaces of online platforms are practices that materially distort or impair, either on purpose or in effect, the ability of recipients of the service to make autonomous and informed choices or decisions." (European Union, 2022, p.18)

Here, the DSA Article 25 prohibits service providers from nudging or deceiving users, distorting their decision-making through the design, structure, or functionalities of their service, in the attempt to prevent individuals from engaging in behaviour which goes against their interest or will. As before, the EU provides a view akin to Restrepo's notion of domination (Restrepo, 2013), whereas manipulation is not only feasible through expressions such as disinformation, but through the service itself. In this regard, Provision 70 notes how recommender systems can influence the extent to which individuals are able to retrieve, interact with and disseminate information, therefore have a huge influence on their ability to make use of the right to FoE, requiring platforms to inform users on how information and based on what parameters is presented to them. Similarly, Provision 69, addresses the issue of online advertisement explaining how disinformation campaigns or discrimination may find

their way into these services, whereas platforms are particularly sensitive environments for these practices. As a result, the EU aims to minimise the risk of exposing individuals to disinformation by outlawing target advertisement based on profiling. By preventing user exposure to disinformation, for instance, the potential of users themselves spreading disinformation may decrease.

Another type of approach is provided in the COC on countering illegal hate speech, which introduces the notion of developing counter narratives, which was also vaguely referenced in the RADTC (European Union, 2021). In the COC, the issue of illegal hate speech is explained to be rooted outside the digital sphere, while the dissemination of these forms of expressions are merely a symptom. In sum, the document appears as a general agreement to tackle the issue of illegal hate speech, providing little tools to combat it. The COC begins by implicating the broader society in having a role in preventing illegal hate speech online, which ought to be done:

> "by developing counter-narratives promoting non-discrimination, tolerance and respect, including through awareness-raising activities" (European Commission, 2016, p.1).

Here, a particular role is ascribed to civil society organisations (CSO) to partake in this approach. Therein, to combat hateful rhetoric, page three explains how IT companies (among the signees) and the European Commission recognise:

> "the value of independent counter speech against hateful rhetoric and prejudice, aim to continue their work in identifying and promoting independent counter-narratives, new ideas and initiatives and supporting educational programs that encourage critical thinking." (European Commission, 2016, p.3).

For this, a cooperation between IT companies and CSOs is stressed, whereas CSOs should act as a trusted partner without however enumerating what this role entails. Ultimately, the cooperation should lead to providing best practices to combat hateful rhetoric and counter speech campaigns (European Commission, 2016). While being presented with those best practices more insights would be possible, this approach can still be thoroughly analysed. Reflective of the general ambiguity, the EU again presents only general points which will be addressed. With this passage, it can be seen that according to the EU's vision of the digital sphere, offline approaches are also relevant to ensure a safe engagement with the digital sphere. Interestingly, the theoretical framework mentioned Cohen-Almagor's pessimistic view of strictly using education to combat hateful expressions, here, the EU partly follows this notion by not relying on just one approach, but envisioning a mix of approaches, both online and offline. Relatedly, the RADTC claimed how terrorist content is a proven catalyst for radicalisation of individuals (European Union, 2021) mirroring Cohen-Almagor's view of hateful expression being followed by hate crimes (Cohen-Almagor, 2012). Ultimately, to prevent hatred from arising, counter narratives and speech are part of the European vision of combating abusive expression, posing as a strategy reflective of providing an anchor of trustworthy content, particularly in regard to the promotion of "new ideas". As little information is provided of what these concepts entail questions are begged whether these concepts aim to enlighten potential perpetrators with what is deemed to be an objective truth or for instance approved positions reflective of what Tropina (Tropina, 2023) and Ruiz (Ruiz, 2023) raised in their discussions.

Hereby, the third overarching approach of establishing an anchor of trustworthy information is entertained, which is particularly relevant when tackling manipulative behaviours such as dis- and misinformation. Reflective of this notion is for instance the highlighting of a given set of information with a respective dilution or restriction of other information. On one side trusted information will be promoted while information deemed untrustworthy will be demoted, applying the notion of Tropina (Tropina, 2023), these cases necessitate an institution to act as an arbiter of truth. An example for such an approach, as already explained in section 4.5 for instance, is the DSA's crisis protocol mechanisms. Here, it can be added that the crisis protocol mechanism proposes the displaying of information relevant to a crisis (European Union, 2022), thereby providing a specific set of information approval which as a result may be accepted as objective truth. As by now established as a clear pattern, this

aspect also allows two distinct interpretations. On one hand, credible information which in a crisis may have the potential to save lives and limit a crisis' impact can be displayed to quickly reach as many people as possible. On the other hand and as criticised by Ruiz, genuine discussions not directly in line with a given approved set of information is at risk of being encroached upon, ultimately undermining user's FoE (Ruiz, 2023). In this case, content moderation of information that contradicts the approved set of information is invited, whereas an entity ought to act as an arbiter of truth and thereby decide what information is abusive or manipulative and therefore warrants interference. Here, decisions would not be made based on objective truth and the right to FoE, but on the decision of what ought to be true, for instance made by the European Commission, ultimately inviting a possibility of arbitrarily establishing FoE boundaries when attempting to combat abusive expression of FoE.

Outside any crisis, this notion is also relevant regarding DSA Article 22 (and the 2018 Communication by the Commission), establishing the role of trusted flaggers. Here, the Digital Service Coordinator of a Member State awards the status based upon application to entities (not individuals) which:

> "have demonstrated, among other things, that they have particular expertise and competence in tackling illegal content and that they work in a diligent, accurate and objective manner" (European Union, 2022, p.16).

Notes regarding potential illegal content by trusted flaggers are meant to receive priority (European Union, 2022). Just as before, this approach provides a possibility to interpret it in at least two distinct ways. On one hand, disinformation can easily be identified by an expert, flagged and be subject to content moderation. On the other hand, as trusted flaggers are considered experts, a risk may arise where platforms take their input at face value, appealing to their authority, deciding to follow their recommendation and interfere with an expression. Thereby, Trusted Flaggers would function as an entity placing boundaries to FoE in their given field of expertise, as well as acting as an arbiter of truth. To mitigate potential risks from trusted flaggers, the DSA foresees investigations in case a flagger is frequently engaging in misconduct. However, if it were to occur, damage would only be minimised by reversing decisions as FoE was already undermined.

The aspect of an arbiter of truth is particularly noticeable in EU documents focussing on combating manipulative behaviour. Of those, the 2018 communication by the Commission precedes the strengthened code of practice of 2022 on disinformation, whereby there is a notable overlap between the two approaches of these, however both provide approaches not found in the other document. In highlighting the susceptibility for abuse of the online platforms, the 2018 communication diagnoses how these failed to act proportionately in addressing the respective challenges and risks of disinformation. The communication goes on to stress how the digital sphere as whole is being manipulated to spread disinformation, hereby underlining how algorithms in their essence are built to amplify sensational content as well as that advertisement models reward sensation. Thereby disinformation is incentivised by the very way the digital sphere functions and people engage in it, whereas purveyors of disinformation are motivated by monetary benefits. In order to tackle this aspect, it is foreseen to remove incentives to create and disseminate disinformation following four guiding principles, transparency, diversity, credibility and inclusivity. In this, similarly to the DSA mostly general directions to tackle are provided, here it is explained that:

> "it is necessary to promote adequate changes in platforms' conduct, a more accountable information ecosystem, enhanced fact-checking capabilities and collective knowledge on disinformation, and the use of new technologies to improve the way information is produced and disseminated online." (European Commission, 2018, p.7)

With the act of fact-checking, the EU again entertains an approach that requires an entity to act as an arbiter of truth. Similarly, the COP lays down 44 requirements (which are in accordance with the obligations and aims of the DSA) for its signatories. Of these, some remain vague, while others more explicitly lay out what service providers ought to follow. A notable finding here as well is the view

that a major driver behind disinformation is a commercial aspect, whereas to tackle this monetary dimension, the COP foresees to:

"significantly improve the scrutiny of advertisement placements, notably in order to reduce revenues of the purveyors of Disinformation" (European Commission, 2022, p.4).

The 2018 communication prohibits targeted political advertisement, and requires more transparency concerning advertising and sponsored content in political contexts as well as sets a general target to reduce the revenue of those abusing advertisement services for the dissemination of disinformation. (European Commission, 2018). The first fourteen commitments of the 2022 COP pursue the same goal, adding how harmful disinformation ought to be barred from being placed within advertisements. Relatedly, the focus of commitment six lies with transparency, requiring a sufficient level of labelling of political and paid content, moreover, information surrounding sponsors of such content should be made known. Here, to prevent any biases when manual reviewing a potential candidate for advertising, signatories agree to pursue neutrality regardless of political orientation or issue, although it is not provided how to achieve or measure neutrality. Importantly, aiming to evade a verification would lead to a ban from access to advertisement services. Related to verification, the 2018 foresees the closing of fake accounts, however more information besides stating that fake accounts are used to amplify disinformation is left out. Moreover, both documents introduce an intricate notion reflective of concerns regarding a role of an arbiter of truth. For one, the 2018 communication requires to:

"Facilitate users' assessment of content through indicators of the trustworthiness of content sources, based on objective criteria and endorsed by news media associations, in line with journalistic principles and processes, transparency regarding media ownership and verified identity" (European Commission, 2018, p.8).

More strikingly, however, the 2022 COP stresses that:

"Signatories recognise the importance of diluting the visibility and permeation of Disinformation by continuing to improve the findability of trustworthy content, enhance the safe design of their services and empower users with dedicated tools to identify disinformation and empowering users with tools to detect and report these types of content" (European Commission, 2022, p.18).

These commitments pose as a prime example of how a passage is interpretable in two very contrasting ways, underlining the intricacy of the arbiter of truth type approaches . On one hand, if objectivity is achievable, trustworthiness indicators could easily help to discern disinformation or its purveyors, on the other hand by having news media associations endorse and therefore influence relevant criteria, private companies are given influence over these indicators, which given the findings of Tropina seems rather concerning (Tropina, 2023). Moreover, it is not addressed what these indicators ultimately entail or how they ought to be designed. In either interpretation, however, the EU foresees strong interference with the exchange of information on online platforms. In one vision it is seen as necessary to enable rights of others by combating illegal hate speech and disinformation, akin to Restrepo's view (Restrepo, 2023), in another vision however, it can be seen as directly undermining FoE and genuine debates, instead aiming to boost favourable opinions, as criticised by Mill for instance (Mill, 1859). The latter passage depends on the objectivity of what is laid out in the prior passage, whereas with the dilution of specific expressions or content, while increasing the prominence of what is deemed trustworthy content, the strategy of simply tackling abusive expressions, is complemented by actively deciding what content to promote and what content to demote. With the findings of section 4.3 in mind, issues arise in terms of the understanding of dis- and misinformation, whereas any given content could virtually be interpreted as mis- or disinformation, leading to its dilution, fuelling prior raised concern. This notion is further specified in regard to a "safe design practices" as a general principle when developing systems, policies, features or recommender systems, as these are supposed to be created in a way that leads to authoritative information receiving more prominence over disinformation (Commitment 18) (European Commission, 2022).

With regard to the already mentioned fact-checking, another approach reflecting the notion of establishing an anchor of trustworthy information or arbiter of truth is underlined. Notably, fact-checkers are portrayed as a key part in the EU's vision to combat abusive expressions, stating that:

"Fact-checkers have emerged as an integral element in the media value chain, verifying and assessing the credibility of content based on facts and evidence. They also analyse the sources and processes of information creation and dissemination. Fact-checkers credibility depends upon their independence and their compliance with strict ethical and transparency rules."(European Commission, 2018, p.9)

By stressing the significance of fact-checkers, they are cemented as part of the European vision of a favourable digital sphere. As with the approaches before, the question of their success depends on how well they work in practice with regard to the issues of ambiguity concerning clear boundaries between legal and illegal categories of expressions, as well as Ruiz findings on excessive fact-checking (Ruiz, 2023). Notably, the communication precedes Ruiz's account, providing a primarily negative outlook in this regard, including on the question of how well the mentioned ethical rules are formulated or are adhered to. Here, the 2022 COP adds how fact-checkers ought to be verifiably independent of partisan institutions and transparent regarding their finances and methods (European Commission, 2022). A final notable aspect is mentioned in the 2018 Communication, which stresses how Russia is actively engaging in disinformation campaigns, thereby naming a country as a direct threat to the EU's vision of a safe digital sphere. Motivated by this, it is stressed how EU institutions have been established to:

"monitor and address hybrid threats by foreign actors, including disinformation, aimed at influencing political decisions inside the EU and in its neighbourhood.",

including a cooperation with NATO for "strengthened European response" and improved resilience (European Commission, 2018, p.16).

While more details are left out, it is underlined how disinformation is also seen as a foreign threat or even a form of warfare, thereby requiring cross border alliances to ensure a safe digital sphere.

In sum, this section found that to combat abusive expressions in the digital sphere, the EU does not lay out clear pathways on how to sufficiently tackle an issue related to abusive expressions while simultaneously providing safeguards to FoE. Instead, service providers are provided with a considerable amount of autonomy on how to satisfy obligations for both dimensions. Notably, the EU also envisions making use of several instruments that aim to provide a form of approved information which ought to receive priority over other forms of content. In combination with the issues of ambiguity concerning an understanding of abusive expressions requiring interference this framework allows interpretation and execution of a digital sphere in an arbitrary manner ultimately giving rise to concerns by overlapping with issues raised in the theoretical chapter.

4.7 AI In the European Digital Sphere: Fighting Fire with Fire

Up until now, the analysis already hinted at how the EU envisions the role of AI in combating abusive forms of expression in the digital sphere, whereby an image of AI as a necessary evil is entertained. This section will flesh out this vision by looking at the remainder of passages addressing AI, including a look at the AIA. Firstly, the DSA requires VLOPs and VLOSEs service providers to enable the monitoring of DSA compliance, which is done by providing relevant data. For AI, this includes data concerning any algorithms, particularly in terms of any risks and harms they emit. Relevant here are the accuracy and functioning of a system when utilised in content moderation or recommendation (Provision 196). Moreover, as discussed earlier, the DSA foresees that service providers need to be aware and address any systemic risk concerning algorithmic amplification, especially, concerning the second category of systemic risk regarding fundamental rights, which according to the European

vision also extends to the design of utilised algorithmic systems (European Union, 2022). To illustrate, Provision 94 of the DSA requires service providers to:

"assess and, where necessary, adjust the design of their recommender systems, for example by taking measures to prevent or minimise biases that lead to the discrimination of persons in vulnerable situations, in particular where such adjustment is in accordance with data protection law and when the information is personalised on the basis of special categories of personal data" (European Union, 2022, p.26),

Here it can be seen that whenever AI is addressed, the EU equally addresses its potential risks or safeguards to prevent for instance the undermining of fundamental rights. Moreover, the notion of service provider's being provided with autonomy is again underlined. What poses as a notable finding here, however, is that algorithmic decision-making always ought to be accompanied by human review in decisions such as the restriction of a user's access to a service. Hereby, likely attributable to the fear of potential risks, distrust in AI as a tool is insinuated. Moving from the DSA to the AIA, eight relevant passages have been coded in regard to using AI as a solution. In doing so, it was found that the AIA does not provide a significant amount of information regarding combating abusive expressions using AI, as passages mostly revolve around requiring that AI in the EU ought to be trustworthy and its use does not undermine fundamental rights (European Parliament, 2024). In regard to content moderation this would refer to AI not miscategorising content, leading to unjust interferences ultimately undermining fundamental rights, similar to Provision 94 of the DSA. In regard to the research focus, the AIA does refer to content moderation when addressing general-purpose AI (AI usable for a variety of purposes), however it merely addresses how AI can be utilised for this task. More broadly, the AIA's risk-based approach sets requirements for so-called high-risk AI systems and outlaws unacceptable AI practices, whereas risk is viewed in terms of risks to fundamental rights, again reflective of the DSA. To mitigate these risks, high-risk AI systems ought to be designed in a way that rights are not undermined (European Parliament, 2024). Here, the AIA raises how attention shall be paid towards risks of general purpose systems in regard to:

"Any actual or reasonably foreseeable negative effects on democratic processes, public and economic security; the dissemination of illegal, false, or discriminatory content." (European Parliament, 2024, p.99).

This passage once more highlights the immense risk that AI is ascribed to in the European vision, raising a potential of it contributing to the dissemination of abusive expressions. Relatedly, a unique solution on tackling abusive forms of expressions is in fact provided in the AIA, particularly in regard to manipulative content. Here however, AI is again presented as a source of the issue and not as an immediate solution, whereby AI providers and deployers are required to enable the detection of AI generated or manipulated outputs as well as need disclose the artificial nature of the output, by for instance labelling deep fakes (manipulated content). In spite of this risk-laden image of AIA, the 2018 Commission as opposed to any other analysed document, provides how AI is in fact seen as a tool when combating abusive expressions, hereby stating how AI will become:

"crucial for verifying, identifying and tagging disinformation" (European Commission, 2018, p.11).

Again however, in the very same sentence it is raised how AI ought to be subjected to human oversight and reflective of the general ambiguity, it is not specified how exactly AI may be utilised in this regard. In sum, AI is on one hand valued as a crucial part of combating abusive expressions, on the other hand and in virtually every instance AI is addressed, the EU discusses potential risks that accompany its use or provides hints of distrust thereby requiring human oversight. With this, the extent to which the EU envisions AI as a tool can be summarised with the following passage:

"it should be possible for hosting service providers to use automated tools if they consider this to be appropriate and necessary to effectively address the misuse of their services for the dissemination of terrorist content." (European Union, 2021, p.6).

4.8 Preliminary Conclusion: Answering The Sub Questions

The last six sections provided and discussed the findings of the coding procedure applied to a selection of EU documents discussing FoE in the digital sphere. Thereby an interpretive outlook on the EU perspective of FoE boundaries, abusive expressions and their placement in regard to these boundaries was provided. Moreover, it was explored how the EU envisions to cope with the provided issues, as well as to what extent the EU foresees the implementation of AI in this strategy. Having presented the findings, this section can finally formulate answers to the sub questions presented in chapter one, beginning with the first sub question of

"How are the boundaries of FoE discussed within the DSA?".

Ultimately, it was found that the actual discussion of FoE boundaries in the DSA is largely fruitless, both in passages that explicitly discuss the notion by word and those that indirectly address it. In the rare instances where FoE boundaries are indeed discussed, a leitmotif was uncovered, whereas FoE boundaries as well as the limits to the powers laid out in the DSA are asserted by the rights of others. With this, on the surface level, the DSA discusses FoE boundaries reflective of Restrepo's concept of democratic FoE, whereby the permissibility of an expression is determined by whether it constitutes an act of domination and transgresses the rights of others (Restrepo, 2013). Here, the DSA raises how abusive expressions such as dis- and misinformation, illegal content including illegal hate speech and terrorist content ought to be combated as they inherently undermine the rights of others. A striking finding however was made, whereas the DSA misses passages clarifying an understanding of those types of expressions (European Union, 2022). Here it was found that to deepen the understanding of the barely discussed FoE boundaries, more documents needed to be consulted. Surprisingly however, the deepening of this understanding led to the finding of the EU providing a framework that allows ambiguous interpretation of categories of expressions that warrant an interference as the EU provides either highly subjective parameters or paradoxically juxtaposes categories of legal, shocking or offensive information with illegal categories in a way that does not allow establishing a clear distinction between those two fronts. This was illustrated by for instance the notion of misleading content in regard to dis- and misinformation, whereas the question of what constitutes misleading is for one highly subjective and secondly not explained. Notably, the DSA also discusses aspects which undermine the dichotomous understanding of FoE boundaries, whereas for one, quasi-legal expressions are introduced which are not illegal yet ought to be addressed in some form, whereas these are subject to an undisclosed interference therefore not protected. These were discussed for instance in relation to systemic risks, whereas this category in itself is not provided with any parameters allowing an explicit identification, moreover it is left ambiguous how they ought to be addressed in relation to FoE boundaries. Lastly, the boundaries themselves in the DSA are found to be flexible, whereas this aspect is not explicitly discussed, but external conditions such as crises allow the Commission to influence content moderation practices, thereby influencing the boundaries of FoE. Notably, Cohen-Almagor criticised Mill's work for providing an incomplete conceptualisation of FoE and its boundaries (Cohen-Almagor, 2017), based on the analysed discussion of FoE and its boundaries, this criticism is also applicable to the provided EU vision. With the finding of ambiguity concerning how to discern protected from unprotected speech, the question arises,

"How does the EU envision solutions to abusive forms of expressions in the digital sphere?".

With the DSA at the forefront, the EU envisions a uniform approach for combating abusive expressions in the digital sphere by laying down the harmonisation on the EU level. It was found that to combat abusive expressions in the digital sphere, the EU primarily envisions service providers to design solutions that simultaneously enable the safeguarding of FoE. Here, the EU foresees service providers to for instance take into account risks of systemic scale, provide and boost the availability of

some form of approved information, and make sure content moderation is contestable and transparent. A finding in this regard, however is, the DSA and adjacent documents largely however refrain from laying out how to satisfy obligations. Thereby, service providers receive a considerable degree of trust, as inferred from the level of autonomy they are provided when designing mechanisms. Strikingly, the prior raised faulty framework of determining whether expressions are abusive or not, combined with ambiguity of how to sufficiently satisfy either end (safeguarding FoE and combating abusive expressions) as well as the degree of autonomy service providers receive, ultimately enables a realm dictated by arbitrary interpretation and mechanisms which enforce a blurry conception of FoE boundaries in the digital sphere. Another notable finding was made in regard to the recurring notion of approaches reflecting the creation of an anchor of trustworthy information, whereby the EU aims to establish figurative safe harbours for approved information. Authoritative information is envisioned to receive more favourable treatment by recommender systems and therefore receive more visibility, whereas contradictory information likely is set to be demoted or even removed. These findings raise the intricacy of the final sub question, which asks

> "To what extent is AI envisioned as a solution for coping with this?".

Contrary to the first two sub questions, the final sub question can be answered in a more definitive manner, whereas the aspect of ambiguity is mostly limited to the question of how service providers will ultimately choose to implement AI on their own accord to satisfy obligations. Similar to most of the strategies analysed in this research, service providers are provided with autonomy in deciding if and how to use AI as a solution to combat abusive expressions. From this it can be concluded that the EU envisions AI as a potential tool and in the case of addressing disinformation even discusses it as inescapable (European Commission, 2018). This view is underlined by the fact that the EU does not propose any alternatives to the use of AI and refrains from prohibiting its use. Contrarily, a noticeable amount of distrust of AI solutions is found, whereas the EU equally refrains from obligating or explicitly suggesting AI use, thereby merely permitting its use. Based on the AIA's risk-based approach and numerous transparency obligations as well as requiring AI to be error-free, AI is not only viewed as part of a solution to combat abusive expressions, but also as a significant driver of the problems and concerns as well as carrying a risk profile when being used to combat abusive expression. Thereby, the EU envisions AI as a helpful and likely inevitable tool, however based on its accompanying risks, a vision reflective of fighting fire with fire is provided, whereby AI is ultimately seen as a double-edged sword.

5 Conclusion

5.1 Introduction

The previous four chapters attempted to illustrate the European vision of FoE in the digital sphere by conducting an interpretive content analysis using EU documents that discuss issues relating to freedom of expression. This chapter will formulate a final answer to the overarching research question and discuss what insights have been made in relation to the knowledge gap identified in chapter one. Finally, the practical implications of the attained findings will be discussed, thereby extrapolating how the theoretical findings of this research relate to the realm of EU policymaking.

5.2 The European Vision of FoE In The Digital Sphere

The overarching research question of this thesis was

> *"How does the European Union interpret Freedom of Expression in the Digital Sphere?".*

Initially, it was expected to gain findings which allow categorising how the EU, with the passing of the DSA, interprets FoE in the digital sphere in regard to potential overlaps with interpretations provided in academic research. Thereby, it was expected to gain an understanding of how to identify and discern abusive expressions from those that ought to be protected from any interferences. Moreover, it was expected to acquire themes within the EU's strategy to combat abusive expressions. While some findings along these expectations have been made, whereby expressions such as disinformation, illegal hate speech and terrorist content undermine the rights of others and therefore ought to be addressed, or even removed from online platforms. As well as a general picture being painted whereas FoE is interpreted in a manner where boundaries are asserted by whether an expression undermines European fundamental rights, it was uncovered how relevant parameters that ultimately allow completing this picture are kept ambiguous across all analysed documents.

Here it was found that overinclusive definitions which depend on subjective interpretations are among the roots of the identified problem. This issue is further exacerbated as information on how to clearly discern expressions which figuratively speaking are seated at the fringe of FoE boundaries, such as controversial (political) statements, from those which are outside FoE boundaries as they constitute a form of hatred is missing. Moreover, as guidelines for service providers when designing solutions mostly remain undisclosed, an arbitrary interpretation of what constitutes for instance dis- and misinformation, or illegal hate speech and therefore how the boundaries of FoE are interpreted in the digital sphere is enabled. This notion is equally exacerbated as the EU provides service providers with complete autonomy on how to design solutions that meet these vague parameters on their own accord. Of the select findings which are clearly highlightable as elements of the EU's interpretation of FoE in the digital sphere, some raise even more concerns against the backdrop of concerns raised by contemporary scholars. A frontrunner of these concerns is the notion of instruments which require an entity to act as an arbiter of truth (Tropina, 2023). In this regard, the EU envisions fact-checking, boosting authoritative information while interfering with information contradictory to an approved set, or in the case of terrorist content aims to determine the true purpose behind an expression being disseminated. Whereas in each of these cases raised in the analyses, an entity ought to draw a line somewhere between what content is to be trusted and what content is deserving of interferences, however, all parameters of this decision are kept ambiguous. Similarly, it was found that FoE boundaries are viewed as flexible, contrary to the understanding prior to the analysis, further underlining the finding of how the EU provides a framework enabling arbitrary interpretation. Here, with the Commission being able to obligate changes to content moderation, coinciding with the concerns raised by Tropina, whereas during the corona pandemic, FoE boundaries were arbitrarily set ultimately undermining FoE, (Tropina, 2023) insights have been gained which allow explanation how the raised concerns were able to materialise.

Even here, the issue of ambiguity reappears as little details on how content moderation ought to be altered are provided. From this, it can be concluded that the flexibility provides the Commission with the power to move the "Overton Window" by deciding what expressions ought to be moderated, again underlining prior concerns. Therefore, it was found that, not only is a framework provided which enables arbitrary categorisation of expressions in regard to the question of their legality, but the approach to combat abusive expressions itself, based on the high level of autonomy and again vague parameters, solutions could be implemented arbitrarily. Whereby it can be concluded that the EU's interpretation of FoE in the digital sphere is a framework that enables to resolve the provided ambiguities of when rights are being transgressed by an expression, with parameters based on any given interpretation of FoE. As a result, mechanisms enforcing these (ambiguous) parameters ultimately enforce this interpretation of freedom of expression in the digital sphere.

5.3 Contributions To The Knowledge Gap

Having answered the overarching research question and presented the main finding of this research, this section can formulate what has been added to the identified knowledge gap laid down in chapter one. This thesis aimed to establish clarity, for instance in regard to whether the DSA is compliant with FoE or not, as contradictory findings were found in contemporary research. Firstly, it was found that both a conclusion akin to Sulmicelli, whereas the DSA may lead to content moderation practices that undermine the rights of minorities (Sulmicelli, 2023), as well as a conclusion akin to Paige, whereas FoE may ultimately still be protected by the act (Paige, 2023) can be correct. Here, the potential for arbitrary interpretation of the EU's vision ultimately allows for both a potential compliance with FoE and one that undermines it. With this finding, a possible explanation for why scholars reach contradictory findings can also be provided. Again, as the framework the EU provides ultimately allows each scholar to insert their individual interpretations and thereby determine whether the laid out parameters would ultimately be compliant with FoE, contradictory viewpoints can be established.

The question of how broad FoE is interpreted in the digital sphere was found to be unaddressed in contemporary research. Here it was found that in the context of the DSA, FoE boundaries are flexible. For one, given the established ambiguity and autonomy allowing arbitrary interpretations which adjust the breadth of these boundaries, but also as a result of distinct mechanisms providing the possibility to move FoE boundaries depending on external conditions, such as crises. Turilazzi and peers also criticised an ambiguity, hindering the possibility to discern between harmful and illegal content (Turillazzi et al., 2023). With this thesis, it was able to add that on top of an unclear differentiation between these categories, virtually any category of expressions addressed in the DSA and adjacent documents appears indistinguishable as a result of subjective parameters. Examples such as the use of the word "misleading", or having to juxtapose between information constituting "hatred" and that which is "shocking" or "offensive" were raised. Notably, it was added how not only is it impossible to objectively discern between these, but this issue runs the risk of arbitrary categorisation, enabling service providers or the EU to ultimately undermine fundamental rights. Regarding AI in particular, Sulmicelli stressed how service providers are becoming increasingly reliant on AI tools (Sulmicelli, 2023). This study found that the EU equally views AI as becoming integral to combating abusive expressions, particularly in regard to disinformation. However, the EU also remains sceptical of the technology as a result of a view of it being tremendously risk-laden. Lastly, Hohmann & Kelemen described the DSA as a

> "set of rules shielding individuals from abuse of power in the digital environment" (Hohmann & Kelemen, 2023, p.226).

This statement is particularly interesting in the light of the findings presented in chapter four. It is understandable how such a conclusion may be reached, however, as it was found that the EU provides a framework that equally allows the opposite of this notion, thereby at least in theory, one can equally conclude that the DSA and adjacent EU documents enable the abuse of power in the digital environment, contradicting this image entirely.

5.4 Practical Implications

The DSA was approved and entered into force in 2024, therefore has received approval by the EU across all institutions involved in the legislative process. However, by providing a framework that not only allows ambiguous interpretation of FoE boundaries and the classification of what content is protected, potential for friction arises when different service providers aim to implement strategies based on DSA guidelines. Here, service providers may easily interpret relevant questions and concepts differently than the EU, therefore based on the missing disclosure on how to differentiate between legal and illegal content, conflict between providers and the EU is bound to occur. Relatedly, it was raised how the EU began investigating Twitter in regard to not complying with DSA guidelines (European Commission, 2023), the insights of this research give rise to speculation that more conflict is bound to follow. As a result, legal certainty, therefore an aim of the DSA is arguably undermined as providers face incredible amounts of ambiguity while also facing a threat of potential punishment when failing to comply with the DSA, potentially shying away from the European market. A similar conclusion was raised in the work of Hohmann & Kelemen who raised how the DSA's operating costs alone may shy away especially start-ups (Hohmann & Kelemen, 2023), the issue of ambiguity likely exacerbates this risk. With this, at least two possible scenarios also arise. Platforms may either insufficiently tackle abusive content, such as direct calls to inflict violence. Or, on the other hand, content moderation may become too restrictive, infringing on legitimate debates, hindering an individual's personal development ultimately undermining FoE, as already observed by Ruiz (Ruiz, 2023). In either case, however, the Commission would likely resort to penalising service providers for not adhering to DSA obligations. Next to service providers, also users themselves may be uncertain whether what they can ultimately express without having to fear repercussions, potentially shying individuals away from utilising their fundamental rights.

With this in mind, some form of clarification by EU policymakers on how to achieve an equilibrium between the FoE of an individual and rights of others, thereby the EU's actual interpretation of FoE, would be helpful, if not necessary. Based on this clarification, content moderation practices and ultimately clarity on what expressions are legal to disseminate and what expressions ought to be combated could be provided. Here, an intricate question can be raised however, specifically, how or if one could ultimately establish such an equilibrium. In some conceptions, this question seems rather simple, while in a realm subordinate to the rule of law many aspects need to be considered, whereas the gravity of this challenge is reflected in the tremendous amount of ambiguity encountered in the analysed documents. When subscribing to the framework of right transgressions serving as boundaries to FoE, one needs to establish when this parameter is ultimately satisfied. Here, however, different individuals, with varying moral frameworks, likely perceive the boundary of a right transgression to be met at varying levels of expressions. The author of this research generally agrees with the notion as provided by Restrepo (Restrepo, 2013), however, as conceded, clarification is needed where one considers rights to be transgressed, be it an actual threat of violence, insults, defamation or offensive opinions and notably when any of these categories is actually satisfied. On the other hand, it is to be expected that when an institution makes use of its sovereign power to influence what content is permissible, it also ought to clearly lay out how to discern between legal and illegal expressions in the digital sphere. Thereby, agreement with the scholars Tropina and Ruiz is found which highlight how abusive expressions such as disinformation are akin to diseases for democracy (Ruiz, 2023), however, the remedy to consequently act against such expressions opens a possibility for arbitrary interferences (Tropina, 2023), as underlined in this thesis.

References

Academic Sources

Bromell, D. (2022). *Regulating Free Speech in a Digital Age: Hate, Harm and the Limits of Censorship*. Springer International Publishing. https://doi.org/10.1007/978-3-030-95550-2

Cassim, F. (2015). Regulating hate speech and freedom of expression on the Internet: Promoting tolerance and diversity. In *South African Journal of Criminal Justice* (3rd ed., Vol. 28). https://uir.unisa.ac.za/handle/10500/21722

Chen, G. (2022). How equalitarian regulation of online hate speech turns authoritarian: a Chinese perspective. In *Journal of Media Law* (1st ed., Vol. 14, pp. 159-179). https://doi.org/10.1080/17577632.2022.2085013

Cohen-Almagor, R. (2012). Is Law Appropriate to Regulate Hateful and Racist Speech: The Israeli Experience. In *Israeli Studies Review* (2nd ed., Vol. 27). Association for Israel Studies. https://www.researchgate.net/publication/233852311_Is_Law_Appropriate_to_Regulate_Hateful_and_Racist_Speech_The_Israeli_Experience

Cohen-Almagor, R. (2017). JS Mill's Boundaries of Freedom of Expression: A Critique. In *Philosophy*. https://ssrn.com/abstract=2992211

Cohen-Almagor, R. (2019). Racism and hate speech – A critique of Scanlon's Contractual Theory. In *First Amendment Studies* (1-2 ed., Vol. 53, pp. 41–66). https://doi.org/10.1080/21689725.2019.1601579

Drisko, J., & Maschi, T. (2015). Content Analysis. In *Pocket Guides to Social Work Research Methods*. Oxford Academic. https://doi.org/10.1093/acprof:oso/9780190215491.001.0001

Goddard, C. (2021). Freedom of Speech and Freedom of Expression – where are the boundaries? In *Pravosudie/Justice* (4th ed., Vol. 3, pp. 68–93). 10.37399/2686-9241.2021.4.68-93

Heldt, A. (2022). EU Digital Services Act: The White Hope of Intermediary Regulation. In *Digital Platform Regulation: Global Perspectives on Internet Governance* (pp. 69-84). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-95220-4_4

Hohmann, B., & Kelemen, B. K. (2023). Is There Anything New Under the Sun? A Glance at the Digital Services Act and the Digital Markets Act from the Perspective of Digitalisation in the

EU. In *Croatian Yearbook of European Law and Policy* (Vol. 19, pp. 225 - 248). https://www.cyelp.com/index.php/cyelp/article/view/542

John, R. (2019). Freedom of expression in the digital age: a historian's perspective. In *Church, Communication and Culture* (1st ed., Vol. 4, pp. 25-38). https://www.tandfonline.com/doi/full/10.1080/23753234.2019.1565918

Jørgensen, R. F., & Zuleta, L. (2020). Private Governance of Freedom of Expression on Social Media Platforms: EU content regulation through the lens of human rights standards. In *Nordicom Review* (1st ed., Vol. 41, pp. 51 - 67). Sciendo. https://sciendo.com/article/10.2478/nor-2020-0003

Jozwiak, M. (2016). Balancing the Rights to Data Protection and Freedom of Expression and Information by the Court of Justice of the European Union: The Vulnerability of Rights in an Online Context. In *Maastricht Journal of European and Comparative Law* (3rd ed., Vol. 23, pp. 404 - 420). SageJournals. https://journals.sagepub.com/doi/abs/10.1177/1023263X1602300302

Lange, B., & Lechterman, T. (2021). Combating disinformation with AI: Epistemic and ethical challenges. In *2021 IEEE International Symposium on Technology and Society (ISTAS)* (pp. 1-5). IEEE. 10.1109/ISTAS52410.2021.9629122

Leerssen, P. (2023). An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation. In *Computer Law & Security Review* (Vol. 48). https://www.sciencedirect.com/science/article/pii/S0267364923000018

Mill, J. S. (1859). *On Liberty*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R2065

Paige, T. (2023). *The Digital Services Act: Does it Respect the Freedom of Expression, and Is It Enforceable?* eRepository @ Seton Hall. https://scholarship.shu.edu/cgi/viewcontent.cgi?article=2413&context=student_scholarship

Restrepo, R. (2013). Democratic Freedom of Expression. In *Open Journal of Philosophy* (3rd ed., Vol. 3, pp. 380-390). http://dx.doi.org/10.4236/ojpp.2013.33058

Ruiz, L. G. (2023). Disinformation, Misinformation and Limits on Freedom of Expression During the Covid-19 Pandemic: A Critical Inquiry. In *The Age of Human Rights Journal* (Vol. 21). https://doi.org/10.17561/tahrj.v21.8149

Saunders, B. (2023). A Millian Case for Censoring Vaccine Misinformation. In *Bioethical Inquiry* (Vol. 20, pp. 115-124). https://doi.org/10.1007/s11673-022-10226-3

Sulmicelli, S. (2023). Algorithmic content moderation and the LGBTQ+ community's freedom of expression on social media: insights from the EU Digital Services Act. In *BioLaw Journal* (2nd ed., pp. 471-489). https://doi.org/10.15168/2284-4503-2717

Tropina, T. (2023). Pandemics and Infodemics: How COVID-19 is Reshaping Content Regulation. In *Beyond the Pandemic? Exploring the Impact of COVID-19 on Telecommunications and the Internet, Emerald Publishing Limited, Leeds,* (pp. 229-243). Emerald Publishing Limited. https://doi.org/10.1108/978-1-80262-049-820231011

Turillazzi, A., Taddeo, M., Floridi, L., & Casolari, F. (2023). The digital services act: an analysis of its ethical, legal, and social implications. In *Law, Innovation and Technology* (1st ed., Vol. 15, pp. 83–106). Routledge. https://doi.org/10.1080/17579961.2023.2184136

Wilman, F. (2022). The Digital Services Act (DSA) - An Overview. https://dx.doi.org/10.2139/ssrn.4304586

European Union Sources

Council of the European Union. (2014, May 12). *EU Human Rights Guidelines on Freedom of Expression Online and Offline*. https://eeas.europa.eu/sites/default/files/eu_human_rights_guidelines_on_freedom_of_expression_online_and_offline_en.pdf

European Commission. (2016, June 30). *The EU Code of conduct on countering illegal hate speech online*. European Commission. https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

European Commission. (2018, April 26). *Tackling online disinformation: a European Approach*. EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018DC0236

European Commission. (2022, June 16). *2022 Strengthened Code of Practice on Disinformation | Shaping Europe's digital future*. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation

European Commission. (2023, December 18). *Commission opens formal proceedings against X under the DSA*. European Commission. https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6709

European Union. (2012, October 26). *Charter of Fundamental Rights of the European Union*. EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A12012P%2FTXT

European Union. (2021, April 29). *Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online (Text with EEA relevance)*. EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX%3A32021R0784

European Union. (2022). Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance). In *Official Journal of the European Union* (1st ed., Vol. L277, pp. 1–102).

News Articles

Browne, R. (2022, April 26). *Elon Musk's Twitter takeover sets him on a collision course with Europe*. CNBC. Retrieved January 20, 2024, from https://www.cnbc.com/2022/04/26/elon-musks-twitter-takeover-sets-up-a-potential-clash-with-europe.html

Datta, T., & D'Silva, A. (2022, November 30). *EU warns Musk that Twitter faces ban over content moderation -FT*. Reuters. Retrieved January 20, 2024, from https://www.reuters.com/technology/eu-warns-musk-that-twitter-faces-ban-over-content-moderation-ft-2022-11-30/

Milmo, D. (2022, April 14). *How 'free speech absolutist' Elon Musk would transform Twitter*. The Guardian. Retrieved January 21, 2024, from https://www.theguardian.com/technology/2022/apr/14/how-free-speech-absolutist-elon-musk-would-transform-twitter

PR Newswire. (2022, April 25). *Elon Musk to Acquire Twitter*. PR Newswire. Retrieved January 20, 2024, from https://www.prnewswire.com/news-releases/elon-musk-to-acquire-twitter-301532245.html

Appendix

Table 2: Coding Scheme with Examples

Category	Codes	Explanation	Examples
DSA-FoE Boundaries	Boundaries; FoE;	The codes will be utilised in regard to how the boundaries are interpreted and to code passages addressing FoE in general.	"all public authorities involved should achieve, in situations where the relevant fundamental rights conflict, a fair balance between the rights concerned, in accordance with the principle of proportionality" (European Union, 2022, p.40).
Abusive Expression	HS-Definition; DI-Definition; IC-Definition; Illegal Content; Hate Speech; Disinformation;	The codes will be utilised to code sections to identify how hate speech, disinformation and other forms of illegal content are conceptualised and described.	"Disinformation is understood as verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm." (European Commission, 2018, p.3).
Solutions	AI-Solutions; HateSpeech-Solution; Disinformation-Solution; Illegal Content-Solution; AI-Content Moderation;	The codes will be utilised to identify the solutions to the respective challenges.	"hosting service providers should act with due diligence and implement safeguards, where appropriate, including human oversight and verifications, to avoid any unintended or erroneous decision leading to the removal of or disabling of access to content that is not terrorist content." (European Union, 2021, p.5)