UNIVERSITY OF HELSINKI

Regulating Lies: Measuring the Risk to Freedom of Expression in Government Responses to Disinformation Across Media Systems

Master's Programme in Contemporary Societies, Social and Public Policy track

Master's thesis

Author:
Yenlik Dairova

Supervisor(s):
Dr. Minna van Gerven
Dr. Kari Karppinen

26.05.2025
Helsinki

Faculty: Social Sciences
Degree programme: Master's Programme in Contemporary Societies
Study track: Social and Public Policy
Author: Yenlik Dairova
Title: Regulating Lies: Measuring the Risk to Freedom of Expression in Government Responses to Disinformation Across Media Systems
Level: Master's
Month and year: May 2025
Number of pages: 77 (84)
Keywords: Disinformation; Policy Analysis; Human Rights; Freedom of Expression; Media Systems; Digital Governance; Democracy; Censorship
Supervisors: Dr. Minna van Gerven, Dr. Kari Karppinen
Where deposited: University of Helsinki Library
Additional information: -
Abstract:

The growing concerns over the harms of disinformation to democracy have prompted governments to issue a wide range of policy responses. Disinformation regulation presents a dilemma from a human rights perspective. Under-regulation may lead to citizens being left unprotected from the harms of disinformation, while over-regulation is linked to increased censorship. With both extremes equally perilous to freedom of expression, researchers have turned their attention to inspecting whether responses to disinformation conflict with international human rights standards. However, there is a dearth of literature that conceptualizes and measures the risk to freedom of expression.

Against this backdrop, this thesis develops a tool for systematic assessment of the Risk to Freedom of Expression (RFE), a novel composite score measure that is grounded in international human rights principles. The score consists of three theoretically grounded concepts: (1) clarity of definition of disinformation, (2) centralization of regulatory authority, and (3) strictness of regulation. The sample includes 50 government responses to online disinformation issued by 12 democratic states of four different media systems between 2010 and 2024.

The results reveal wide variation in RFE scores, even among countries with high levels of democracy and press freedom. Countries with the lowest risk— Finland, the Netherlands, and Portugal — tend to work collaboratively with digital platforms, clearly outlining illegal disinformation and avoiding the role of arbiters of truth. Higher-risk countries — among which are Germany and the U.S. — tend to delegate censorship to platforms, leading to the concentration of power at the digital platform level and the increased risk of censoring legitimate expression. Furthermore, this thesis shows that policies with multidimensional and educational focus are better at safeguarding freedom of expression.

Given that disinformation issue is deeply intertwined with institutional weaknesses, future research can quantitatively inspect which factors lead to the increased risk to freedom of expression. Such insights can be used to move from diagnosis of the problem to prevention. As disinformation governance continues to evolve, it requires constant monitoring and evaluation, so the concepts developed in this thesis can also be used to understand how dire the risk is and call out disproportionate and chilling responses. This thesis contributes to the burgeoning research on balancing speech regulation and democratic accountability in disinformation governance.

Table of contents

Acknowledgements

This thesis has allowed me to combine my fascination with the issue of disinformation and passion for human rights with the rigor of quantitative methods. This has been the greatest academic challenge and adventure I have yet had.

I am grateful to my supervisor, Dr. Kari Karppinen, for his academic guidance, invaluable insight into the topic of media governance, and support that has left a profound impact on the completion of this thesis. I would also like to extend my sincere thanks to my supervisor Dr. Minna van Gerven for her practical suggestions and helpful advice. Her meticulous attention to detail has significantly shaped this thesis. I am also grateful to Dr. Kimmo Vehkalahti for providing insightful comments and suggestions for the methodology and analysis sections. My appreciation goes out to the University of Helsinki for rewarding me with a generous tuition fee scholarship and master thesis grant, which enabled me to wholeheartedly focus on my research. Thank you to my peers for their support and the stimulating discussions that created an environment where we nurtured our research ideas into fleshed out academic works.

Finally, I dedicate this thesis to family, to whom I am immensely grateful for their unwavering support. To my mother Sholpan, for being the one who sent me the UH scholarship opportunity and encouraging me to always go for it. Thank you to my father Talgat for always listening – from my retelling of entire book plots when I was 11 to brainstorming ideas with me for this thesis. To my grandparents, Kadisha and Nurlan, for being my academic role models and greatest cheerleaders. Thank you to my sister Malika and brother Shona for calling and checking in when I have been lost to the abyss of my research process. And to my partner Charly, thank you for being there for me at every step of the way.

1 Introduction

An insightful remark, "A lie will fly around the whole world while the truth is getting its boots on," is often attributed to Mark Twain (Standard Player Monthly, 1919). Ironically, the adage itself is an example of misinformation since Twain died nine years prior. The wording of this phrase evolved throughout the 19th and 20th centuries while being attributed to historical figures like Jonathan Swift, Winston Churchill, and Thomas Franklin (Chokshi, 2017). Most probably, people misquoted it unintentionally, which is a case of misinformation. However, if created and shared deliberately to mislead and increase its persuasive weight by attaching a famous name, it would be closer to disinformation. The difference hinges on intent, making disinformation a communicative act as much as any persuasive information.

From the communication theory perspective, disinformation is not just false content or one innocent lie; it is strategically constructed, persuasive messaging designed to deceive individuals (Smarandache et al., 2014). As such, disinformation is a form of psychological aggression, exploiting cognitive biases and socio-political fault lines (Ibid). False narratives have been paramount in political communication, which accelerated with the invention of the press, radio, and television. Targeted lies produced on a mass scale became state propaganda, which populated communication channels during wars. Scholars such as Ogden and Richards (1923) explored how meaning is made and can be manipulated through language, which was crucial to studying propaganda in the context of World War I and its effects on populations. In the 20th century, such campaigns enabled large-scale manipulation to raise public morale, as well as to justify mass-scale violence via patriotic sacrifice (Fox & Welch, 2012). The adoption of digital technologies brought this trend to new heights, allowing disinformation to spread globally within seconds. Global crises like the COVID-19 pandemic or conflicts like Russia's attack on Ukraine have highlighted how disinformation has become a strategic tool of hybrid warfare with dangerous implications for democracy and civil stability (Dov Bachmann et al., 2023).

Disinformation goes hand in hand with authoritarianism, with the former linked to increased autocratization even in democratic states (Sato & Wiebrecht, 2024; V-Dem, 2025). The level of democracy enjoyed by an average person in 2024 plummeted to 1985 levels, with 5.7 billion people worldwide living in autocracies, while *freedom of*

expression appeared to be the worst deteriorating component of democracy (V-dem, 2025). V-Dem Institute also reports that North America and Western Europe are not immune to those alarming backslides either: by country averages, the levels of democracy in this region are back to those in 1983 (Ibid). Democracy rests upon the principle of equal participation in culture, ensuring everyone exercises their right to express themselves (Balkin, 2004). Today, the Internet has become one of the principal means by which individuals exercise that right (Ibid). Access to information is crucial for democratic debates in the public sphere, but the Internet also makes it that much easier to instantly share disinformation that hinders healthy public discourse, fueling the democratic decline (Tenove, 2020; Colomina et al., 2021; Allegri, 2024; Rucinska, 2023; Humprecht, 2023; Evangelista & Bruno, 2019; Maati et al., 2023). Manipulation of information is far from new, yet the usage of artificial intelligence-powered tools and algorithms amplified the ability to create and distribute mis-/disinformation to an unimaginable extent (Mansell et al., 2025).

Governments around the world were tasked with the challenge of developing solutions to dissuade people from creating and disseminating disinformation. Regulation of lies presents many challenges from legal perspective, particularly in context of freedom of speech (Lidsky, 2008). The first challenge was understanding what disinformation is and how to circumscribe it within general false information to avoid censorship. The second challenge was adequately sharing regulatory responsibilities with digital platforms that have control over digital communication spaces and users. Inconsistencies in defining disinformation and how to assign regulatory responsibilities lead to both under- and over-regulation that detrimentally affect freedom of expression. This tension has drawn the attention of policymakers and academic scholars to international human rights frameworks, particularly Article 19(3) of the International Covenant on Civil and Political Rights (ICCPR), which protects freedom of expression while allowing limited, clearly defined restrictions (UN General Assembly, 1966).

Thus, inspecting novel legislation from a human rights perspective became imperative. Several researchers have reported inconsistencies between these regulations and human rights standards when analyzing platform regulation of hate speech (Hatano, 2023; Banchio, 2024) or national legislative responses (Shepherd, 2017; Shattock, 2023). While many studies criticize regulatory measures for failing to

comply with international human rights standards, few offer a systematic framework to assess the individual government responses that have a prominent effect on the functioning of the whole information ecosystem. One notable example is the LEXOTA project that tracks laws and government actions against disinformation in Sub-Saharan Africa and assesses them based on the ICCPR Article 19(3) tripartite test (LEXOTA, 2023). Such tools enable cross-country comparison and provide insights into the regulation of disinformation, yet few have designed measurable indicators to evaluate national responses of the Global North. This thesis addresses the gap by seeking an answer to the following research question (RQ): *How can government responses to disinformation be systematically assessed for their risk to freedom of expression?*

This thesis adopts a human rights lens to assess the Risk to Freedom of Expression (RFE) within government responses to disinformation. The resulting RFE composite score consists of three indicators: clarity of definition of disinformation, centralization of regulatory authority, and strictness of regulation. The sample includes 50 disinformation-related responses adopted between 2010 and 2024 across 12 countries, representing four distinct media system types of the Western countries (Finland, the Netherlands, Germany, the United States, the United Kingdom, Australia, France, Portugal, Spain, Hungary, Latvia, and Slovakia).

Then, we empirically test both policy and country-average scores through two sub-questions. Countries sharing a media system model, which reflects historical relationships between political regimes, media institutions, and civil society, may shape how they approach new regulatory challenges like online disinformation (Hardy, 2021). Drawing on Hallin and Mancini's (2004) typology, this study investigates whether long-standing institutional arrangements continue to shape policy responses or whether novel, hybrid-system patterns are emerging. The first sub-question (SQ1) asks: *How well do media systems explain variation in countries' Risk to Freedom of Expression (RFE) scores compared to empirically derived clusters?*

While aggregating a country-level risk score may help compare national trends, it can obscure important differences between individual policy measures. Some policy types, like fact-checking initiatives, may pose fewer risks to expression than punitive

content controls. More crucially, a singular policy can be multifaceted, standing out from narrow policy measures that might be too myopic for a complex issue like disinformation. To capture this variation, this study shifts focus to the policy level by posing the second sub-question (SQ2): *To what extent are particular types of responses, or the diversity of approaches used within a single policy, associated with the risk to freedom of expression?*

The contributions this thesis aims to make are threefold: first, to crystallize the existing scholarly insights from human rights jurisprudence and disinformation literature into three conceptual dimensions that help capture the risk to freedom of expression; second, to empirically show which regulatory approaches can protect citizens against the harms of disinformation without compromising human rights; third, to offer pathways for future policy research by proposing theoretically-driven tools and providing evidence-informed data.

This thesis is structured as follows. Chapter 2 reviews existing literature on disinformation governance and articulates the research gap concerning the risks to freedom of expression, ending with the rationale for using human rights frameworks as a theoretical base and media systems as an analytical lens. Then, it presents the theoretical framework, outlining the human rights approach used to outline three conceptualized dimensions that circumscribe the risk of restricting free expression. It also positions the discussion of disinformation governance within the context of media systems in the digital era.

Chapter 3 describes the methods used to answer the research questions of this thesis. The methodology chapter starts by presenting the data sources and country selection, proceeds with outlining the assessment framework, and finishes with reflecting on the mixed-methods research approach and the scoring system. Chapter 4 presents the results of data analysis. It begins by examining and validating the constructed RFE score. Then, it empirically shows why media system typologies fall short in predicting country groupings while offering an alternative clustering solution and how certain policy types better safeguard speech. Chapter 5 concludes by positioning the findings in the context of existing literature, discussing limitations, and offering directions for future research.

2 Literature Review, Theories, and Concepts

2.1 Literature Review

2.1.1 Defining disinformation

This chapter sets out to identify disinformation and related notions in current research and contextualizes further discussion of the perils of online disinformation governance. Manipulating information has been a powerful tool in shaping public opinion and orchestrating social, political, and economic narratives since the start of human civilization itself (Kaminska, 2017; Banchio, 2024). In the Roman Empire, coins were used as tools of propaganda during a conflict between Octavian and Mark Antony (Kaminska, 2017; Posetti & Matthews, 2018). Inscriptions featured slogans that smeared Antony as a 'drunk,' 'womanizer,' and 'Cleopatra's puppet,' which allowed Octavian to enrage the Senate and subsequently become Augustus — the first Roman Emperor who 'hacked the republican system once and for all' (Kaminska, 2017). The spread of disinformation and misinformation especially skyrocketed with the invention of the Gutenberg press and was weaponized during World War I and World War II (Herzstein, 1978; Welch, 2014). Coins, print media, radio, and now the Internet — all exhibit how the channels of general communication have been historically used to disseminate disinformation and sow discord.

Thus, disinformation has been a pervasive part of conflicts in the 20th century and has been studied as a multifaceted phenomenon (Warldle & Deraskhan, 2017; Bennett & Livington, 2018; Appelman et al., 2022). From a communication mechanics perspective, disinformation is an 'intentional failure of communication' (Smarandache et al., 2014). Essentially, Smarandache and Vlăduţescu stipulated that "disinformation is a type of persuasive information," which is used to target "individuals or social groups on which persuasive influence is exercised either directly, or through intermediaries" (Smarandache et al., 2014, pp. 111-112). Digitalization prompted the reformulation of these theories, especially paying attention to the veracity with which lies can now be spread. Still, the core elements of these earlier communication theories remained the same: intent and harm, both of which are central to the definition of online disinformation.

Intent to deceive is an integral part of the minimal working definition of disinformation (Bennett & Livingston, 2018; Dan et al., 2021; Freelon & Wells, 2020; Hancock & Bailenson, 2021). Furthermore, scholars emphasize that disinformation involves the intent to cause harm to an individual or social group, organizations, or even whole countries (Chadwick & Stanyer, 2022; Hancock & Bailenson, 2021; Wardle & Derakhshan, 2017; Stahl, 2006). However, a statement can be reshared with the person believing its sincerity (misinformation) or may contain partially true elements (malinformation) (Appelman et al., 2022). Thus, the growing need to differentiate various forms of falsehood gave rise to a new body of research (Warldle & Deraskhan, 2017; Bennett & Livington, 2018). As part of the report for the Council of Europe (CoE Report), Wardle and Deraskhan constructed the seminal three-fold categorization framework of mis-, dis-, and mal-information that collectively comprises information disorder based on key definitions (Wardle & Deraskhan, 2017, p.20):

1. **Disinformation** is information that is false and deliberately created to harm a person, social group, organization, or country.
2**. Misinformation** is information that is false but not created/shared with the intention of causing harm.
3. **Malinformation** is information that is based on reality and used to harm a person, organisation, or country.

However, some scholars have criticized the consolidation of various forms of falsehoods due to their fixation on the arbitrary concept of truth. Stahl (2006) claimed that the criterion to prove the intent to lie complicates the governance of mis-/disinformation. He utilized a Foucauldian perspective that problematizes the very categorization since proving the intent of lying "presupposes the existence of a universal truth and the self-reflectiveness of the speaker to know her own intentions when speaking," suggesting that power structures end up deciding what is considered misinformation or disinformation (Ibid, p. 91-92). In other words, Stahl (2006) argued that in a futile effort to seek out the ultimate truth, government structures may be given hard power to define the truth themselves.

Thus, the human rights-respecting laws seek not to reinforce the true/false binary system but to focus on the rights that are being infringed upon instead. Appelman et

al. (2022) suggested that the decision to apply legal measures to limit the spread of disinformation should be made by balancing various contested human rights from relevant legal domains. For example, the law that punishes the spread of disinformation should not scare people into silence since it inhibits their right to speak freely.

This is not to say that defining disinformation is entirely futile from a legal perspective. On the contrary, the lack of clarity on what is even considered 'false' in legislation may lead to ambiguous terms like "fake news" to be weaponized by politicians as a mechanism with which "the powerful can clamp down upon, restrict, undermine, and circumvent the free press" (Wardle & Deraskhan, 2017, p. 16). Research on disinformation formulation is crucial to guide policymaking. Early EU frameworks on action against disinformation omitted the explanation due to the assumption that the concept of falseness is self-explanatory (Cavaliere, 2022). In due time, the European Commission polished the definition in both the 2018 and 2022 versions with the help of the aforementioned CoE report by Warldle and Deraskhan (2017), where the concept of falseness to be verifiably false and purposefully misleading is clarified (Ibid., p. 6). The key is the formulation that focuses not on dictating or fixing the narrative but on capturing the forms of false information that deserve regulatory attention while safeguarding freedom of expression.

This is precisely the focus of this thesis: the way disinformation is defined has the potential to infringe upon fundamental human rights like freedom of expression by attempting to cease the narrative. Hameleers (2023) conceptualized the following definition: "Disinformation refers to all practices of **intentionally** creating or disseminating **deceptive** content to **cause harm**, sow discord, or create financial and/or political gain." By focusing on falsity, harm, and intentionality, this conceptualization disentangles disinformation from other forms of falsehoods and will guide the construction of the concept of risk to freedom of expression.

2.1.2 Disinformation Governance

From 2011 to 2022, 78 countries around the world designed policy measures to limit the spread of online disinformation (Lim & Bradshaw, 2023). In democratic states, these regulations were introduced to protect the public from the dangers of

disinformation, commonly framing it as a 'threat to democracy'. This claim is widely supported in scholarly literature, investigating how disinformation operates strategically to exploit socio-political issues in various national contexts and political regimes (Hameleers, 2023), undermining legitimacy (Tenove, 2020; Colomina et al., 2021; Allegri, 2024), institutional trust (Rucinska, 2023; Humprecht, 2023), and producing inequity (Evangelista & Bruno, 2019).

While the harms of disinformation to democracies have been outlined, not all policies that state they 'protect democracy' do so in reality. Critics argue that when policies define disinformation too broadly as 'false information' or 'narratives that threaten democracy', they undermine its very principles (Phiri, 2023; Katsirea, 2018). Suppose it is up to governments to decide what is false, and the citizens are left to wonder what falsehoods are punishable. In that case, the reasoning of doing it 'for the sake of democracy' becomes a façade to censorship. On the other hand, disinformation researchers caution against completely handing over control to digital platforms (Ruiz, 2023). This fear stems from market research that shows how digital platforms financially benefit from “circulating (or ignoring)” online mis-/disinformation (Giansiracusa, 2021; Ruiz, 2023). This cleavage must be addressed by outlining what disinformation governance is, who is involved, and the boundaries of the government's involvement in the information ecosystem.

Since disinformation tends to exploit the weaknesses of other social institutions (Hameleers, 2023), it cannot be inspected and 'solved' in isolation. Disinformation is an evolving and multifaceted issue that calls for "meaningful and coordinated sociotechnical responses from platforms, communities, educators, policymakers, organizations, researchers, and individuals" (Sanfilippo et al., 2024). Thus, digital media governance is a system of interdependent statutory (government-imposed) and non-statutory (platform- or civil society-driven) responses (Puppis, 2010; Raboy & Padovani, 2010). Since civil participation increasingly happens on digital platforms, governments and social media platforms must work collaboratively to tackle disinformation (Calo et al., 2023). This way, top-down statutory legislation increasingly gave way to co-regulatory and self-regulatory models (Finck, 2017).

As Perlman (2021) explained, “digital platforms have assumed state-like powers without the accountability that comes with democratic elections”, which spurred the debates about holding platforms accountable for the content their users publish. While platforms essentially hold an intermediary position, they have been urged by supranational actors like the UN and the EU to step up and protect their audiences from hate speech and disinformation by issuing appropriate regulations (Romanova et al., 2020). The EU Commission issued several initiatives to coordinate the fight against disinformation on the national, institutional, platform, and individual levels, including Action Plans and two versions of the Code of Practice on Disinformation in 2018 and a strengthened version in 2022 (European Commission, 2022; Bayer, 2024). However, the adoption of these recommendations is still context dependent. In the US, for example, the approach is different due to the immunity granted to platforms inspired by the First Amendment's scrutiny over speech restrictions (Goldman, 2024).

Even considering supranational actors like the EU that endorse the best approaches, the states are ultimately tasked with adapting these approaches and values to their national contexts. States actively shape the regulatory environment by steering organizations involved in media governance (Bevir, 2013, p. 56; Kooiman, 2003; Raboy & Padovani, 2010). In his doctoral dissertation, Phiri (2023) explicated how governments not only have a duty to protect their public from disinformation but also can do so without undermining speech. The present thesis argues that the potential risk of restricting expression depends on the government's vision of its own role in fighting disinformation, which is expressed through policy text.

We have established that there are multiple layers to disinformation governance: supranational entities, states, digital platforms, and citizens themselves. While the literature review in this section has exhibited how the state is in no way a sole actor responsible for disinformation governance, there is a consensus that a state has a duty to protect its citizens and navigate the fine line between under- and over-regulation. An empirical approach is necessary to identify patterns and weaknesses in the formulation of disinformation responses. This study does not seek to assess the entirety of the disinformation governance scene but rather to analyze how state responses are designed from a human rights perspective.

2.2 Theoretical Framework

Having defined what this study means by disinformation and how it understands online disinformation governance, this section outlines the theoretical framework guiding this study's examination of government responses to online disinformation. It begins by situating freedom of expression as a foundational human right upon which democracy itself rests. In the digital age, however, ways to express oneself have fundamentally changed. While digital platforms have facilitated the reach of information flow, they have also facilitated the rapid spread of disinformation. This change has forced governments to rethink whether freedom of expression differs in digital spaces and which actions to take to counter disinformation without suppressing legitimate speech.

To navigate this conundrum, I discuss international human rights frameworks based on the International Covenant on Civil and Political Rights (ICCPR), which circumscribes the government's power by providing legal standards for permissible reasons to restrict speech (UN General Assembly, 1966). Drawing from this legal basis, the chapter introduces three key dimensions of disinformation legislation: clarity of definition, centralization of regulatory authority, and strictness of regulation. These dimensions represent policy's potential risk to restrict expression. Finally, the chapter discusses media system theory as an analytical lens to assess whether countries with similar media systems tend to respond to disinformation in comparable ways while also reflecting on critiques that challenge the explanatory power of this framework in the context of digital regulation.

2.2.1 Freedom of Expression and Disinformation

Freedom of expression is not only a right in itself but a cornerstone for the enjoyment of other fundamental rights, such as freedom of assembly, association, and electoral participation (Bychawska-Siniarska, 2017). These rights collectively form the civil and political foundation of democratic societies, which are now increasingly exercised online. Balkin (2004) highlighted the benefits of digitalized communicative spaces, namely how they have enabled new forms of civil participation and collective action, promoting democratic culture. However, significant ease of distributing information came at the cost of its quality, giving rise to disinformation that can undermine the

integrity of public debate and electoral processes (Phiri, 2023). European Convention of Human Rights stipulates that political expression deserves the highest level of protection, which necessitates state action against disinformation (Bychawska-Siniarska, 2017). Since disinformation legislation involves restrictions on speech, it requires careful examination from an international human rights perspective.

Article 19(3) of the International Covenant on Civil and Political Rights (hereinafter, ICCPR) was formulated on UN-recommended interpretations of the permissible legal limitations to restrict speech (UN General Assembly, 1966). The ICCPR Article 19(3) permits restrictions on freedom of expression if they pass the tripartite test: (1) "it is formulated with sufficient precision to allow individuals to regulate their conducts accordingly"; (2) "it has a "legitimate aim" among those established by the Article"; (3) it is "necessary and proportionate to the aim". Legitimate reasons for restriction include: (a) For respect of the rights or reputations of others; (b) For the protection of national security or of public order or public health or morals (Ibid).

All three of these components must work in tandem in a policy limiting freedom of speech. However, some human rights experts state that the second legitimate reason (protection of national security or public order) is broad enough to leave much discretion to the states on how they frame threats to national security or order, which risks majoritarian infiltration (Callamard, 2008; Gunatilleke, 2021). Building on that argument, Gunatilleke (2021) criticized the tripartite test for prioritizing majoritarian interests that leave the freedom of expression of minorities and political dissidents unprotected and targeted.

This vulnerability manifests when disinformation-related legislation is analyzed using the ICCPR Article 19's tripartite test. During the COVID-19 pandemic, many states issued counter-disinformation legislation in the name of public order and national security (Lim & Bradshaw, 2023). Due to the struggles over defining disinformation and its harms, these laws against "fake news" often failed on the very first component of the tripartite test requiring "sufficient precision". In 2020, UN Special Rapporteur David Kaye stated that such laws often empower the authorities to become the arbiters of truth, a role incompatible with democratic freedom (Kaye, 2020). However, many of these legislations over-emphasize the legitimacy component and

claim these restrictions are necessary for the sake of public order and national security, when, in reality, these laws often clamp down on dissent and persecute legitimate expressions as well (Lim & Bradshaw, 2023). Appelman and her coauthors provided a legal commentary on “awful but lawful” content that remains protected expression and "cannot simply be restricted by statutory means only because it is false" (Appelman et al., 2022). Vague restrictions are illegitimate by proxy, be it for the sake of "public order" or even "public peace". The UN Special Rapporteurs, legal scholars, and human rights jurisprudence unequivocally contend that disinformation should be regulated in proportionate, rights-respecting measures, not by empowering states to decide what is true.

As a result, international human rights bodies have increasingly turned their attention to adapting the ICCPR framework to the context of disinformation policies. Amnesty International's report entitled "A Human Rights Approach to Tackle Disinformation" detailed recommendations to prioritize "credible, reliable, objective, evidence-based and accessible information is disseminated to all", instead of silencing and criminalizing mis-/disinformation (Amnesty International, 2022). The Global Partners Digital (2023) framework similarly suggests that demonstrable harm is a key determinant of the regulation of disinformation and that determinations of harmful content be decentralized and fall under the review of independent judicial authorities, not political actors. Political communication and legal scholarship echoed their sentiments (Jacobs, 2022; Shepherd, 2017; Hatano, 2023; Cavaliere, 2022; Helm & Nasu, 2021).

Several studies have used a human rights framework to assess the compliance of disinformation-related legislation with the ICCPR tripartite test on national and digital platform levels (Shepherd, 2017; Jacobs, 2022; Shattock, 2023). For instance, Vese (2022) critically analyzed legislative and administrative responses to disinformation across nine countries around the world (China, the US, Russia, the UK, Australia, Canada, Burkina Faso, Singapore, and India) and across four EU Member States countries (Germany, France, Italy, and Spain) in terms of their adherence to freedom of expression standards and argued for the effectiveness of self-regulation and empowerment tools utilized by platforms.

The most seminal effort has been made by a consortium of civil society organizations that developed a LEXOTA project, which analyzes the laws and government responses to disinformation in the Sub-Saharan Africa region using the ICCPR Article 19(3) as a part of its assessment (LEXOTA, 2023). Their methodology uses a 3-question framework for government responses and a 6-question framework for laws by providing "yes/no/potentially" answers. LEXOTA illustrates its findings on an interactive map and provides a detailed analysis of every law or government action. The present thesis enriches these endeavors by offering measurable conceptual dimensions that serve as a base for a composite score measure of risk to freedom of expression. Stemming from human rights frameworks, this thesis identifies three key conceptual dimensions of disinformation policies that may conflict with the legal principles designed to protect speech.

Clarity of Definition of Disinformation

One of the key conceptual dimensions used to assess the risk of disinformation policy to restrict freedom of speech is the clarity of the definition of disinformation. As stated above, restriction of speech in disinformation governance is permissible if it is "provided by law", meaning the government response to disinformation should be formulated precisely and narrowly enough for the person to understand what is prohibited. For example, under the UN Human Rights Committee (General Comment No. 34) and Global Partners Digital (2023) frameworks, policies targeting vague or ambiguous concepts like “fake news”, “dangerous lies”, or “non-objective content” would fail this standard. As outlined in the literature review, disinformation scholarship broadly agrees that effective legal definitions should include three core elements: provable falsity, intent to deceive, and demonstrable harm (Wardle & Derakhshan, 2017; Hameleers et al., 2023; Banchio, 2024).

Crucially, some government responses to disinformation focus on prevention rather than restriction, for example, by imposing responsibilities on journalists, mandating that they publish only lawful content known to them as truth. Nevertheless, this reverse formulation still covers the same triad: falsity, intent, and harm. Detailed articulation of harm is important for satisfying the necessity and proportionality requirements under human rights law and the ICCPR tripartite test. From a

theoretical standpoint, clarity safeguards against arbitrary enforcement by limiting interpretive discretion (Farinho, 2021). Simply put, the clearer the illegal disinformation is defined, the lower the risk that mere falsehoods with no demonstrable harm will be criminalized. Under such a framework, false information in itself does not constitute a legal violation, as the act of lying is not inherently criminal.

Centralization of Regulatory Authority

The second key conceptual dimension is the centralization of regulatory authority — who holds power to determine what counts as disinformation. While Article 19(3) of the ICCPR does not directly address the regulatory source, human rights frameworks inspired by ICCPR raise the questions of legitimacy and impartiality of restrictions on speech. For example, Global Partners Digital's (2023) framework emphasizes that decisions on speech legality should be made by independent and impartial judicial authorities rather than political actors. The concept of centralization of power refers to the concentration of decision-making power within a single source of authority, which can be conceptualized as centralization of power to decide on acts of disinformation (Porter & Olsen, 1976; Bowman & Krause, 2003; Spina, 2014; Arayankalam et al., 2024; Lim & Bradshaw, 2023).

Researchers commonly distinguish between three models of media governance positioned on a spectrum: self-regulation, co-regulation, and statutory regulation (Puppis, 2010; Baldwin, Cave, & Lodge, 2011; Stasi & Parcu, 2021). Yet, this regulation framework is more of a “spectrum of possibility”, across which the regulation types may not fully be distinct (Stasi & Parcu, 2021, p. 422). For example, under self-proclaimed co-regulatory models, as Stasi & Parcu (2021) defined, “rules are negotiated by the regulator and those subject to them”. More often than not, digital platforms are ultimately the ones who decide on the acts of disinformation and censor content, which would be closer by definition to self-regulation. I aim to enrich the regulation conceptual model by restructuring it from the perspective of the centralization of power concept.

The more the power is centralized (e.g., governments as arbiters of truth), the greater the potential for abuse, censorship, and political manipulation (Sperry, 2024; Lightfoot & Wisniewski, 2014). De-centralization of power can be achieved by delegating the role to independent judiciary bodies. This logic echoes long-standing critiques of statutory regulation that may erode civil liberties and foster surveillance when not balanced by judicial oversight or public accountability (Lim & Bradshaw, 2023; Stasi & Parcu, 2021).

Finck (2017) argues that both statutory and self-regulatory design options raise problems due to the centralization of power, while the co-regulatory model combines the platforms' involvement with external oversight from the state, which allows for "more informed decision-making, easier enforcement, and continuous review and assessment" (Finck, 2017, p. 29). Crucially, this dimension is not concerned with who regulates disinformation but with who, if anyone, is an arbiter of truth. By viewing regulatory models through a lens of centralization of power, we can inspect how different institutional designs may influence the risk of restricting expression under the guise of combating disinformation.

Strictness of Regulation

Lastly, the strictness of regulation dimension focuses on how state responses restrict the spread of false information and the severity of those responses. By extrapolation, we can deduce one of the reasons Article 19(3) of the ICCPR stipulates the necessity and proportionality of restrictions on freedom of expression — to prevent policing of disfavoured content. The strictness dimension focuses on the control of false but lawful information regulation: how and to what extent false content is suppressed, removed, or sanctioned. Thus, assessing strictness is useful to understand how heavily states intervene in the information ecosystem. I utilize Lim and Bradshaw's (2023) typology of legal penalties for spreading disinformation: administrative burdens, content censorship, monetary fines, and imprisonment.

The concept of strictness of regulation utilized in this thesis focuses on identifying instances where false but lawful expressions are targeted. Hence, it does not report on the issued punishment for illegal disinformation. To illustrate, let's review an

example of now repealed Germany's NetzDG law. The German law prompted social media platforms to remove “illegal content”, as defined by the German Criminal Code (2017), and fined platforms that failed to make a decision in 24 hours. While fines were only applied when illegal content was not removed, legal scholars have argued that the law incentivized platforms to remove legal but controversial speech out of fear of financial penalties (Tworek & Leerssen, 2019). This thesis classifies this policy as “content censorship” on the dimension of strictness of regulation.

From a human rights perspective, the implicit effects of regulation matter as much as its explicit ones. The goal of this measure is not to assess the necessity of disinformation regulation, which would be challenging to measure empirically but to investigate more apparent violations of free speech with such regulation. Therefore, the dimension of strictness captures the level of control over the spread of false but legal information, which is incompatible with the fundamental right to freedom of expression principles.

In sum, literature exhibits how specific formulations of government responses to disinformation can threaten freedom of speech. This section has conceptualized three dimensions of disinformation policy content that will be used to assess this risk. Such an approach allows for the systematic analysis of policy-level risks, and buttressing international legal standards makes it possible to assess and compare disinformation governance across different national contexts. The following section now turns to the analytical lens of media systems, asking whether countries with similar media system characteristics also exhibit similar patterns in regulating disinformation and safeguarding or undermining freedom of speech.

2.2.2 Disinformation and Media Systems

The way the state treats regulatory issues like disinformation is the product of broader socio-political, media, and cultural environments. One state's policies may be distinct from the policies of another. Zooming out even more, countries may form clusters based on their disinformation governance approaches. In their groundbreaking book “Comparing Media Systems,” Hallin and Mancini conducted an empirical comparative analysis of media systems of North America and Western Europe by analyzing the development of media and identifying characteristic patterns

of relationship between system characteristics (Hallin & Mancini, 2004 p. 11). The authors outlined four key dimensions to compare the systems: press development, political parallelism, journalistic professionalism, and the role of the state (Ibid). The analysis identified three ideal models: the Mediterranean or Polarized-Pluralist model, the North/Central European or Democratic-Corporatist model, and the North Atlantic or Liberal model. This thesis uses media systems theory as an analytical lens since it reflects the factors that shape how states define, regulate, and enforce information policies.

Media systems research seems particularly useful when analyzing the relationship between media and politics, but in the twenty years since its original conception, the theory has been extensively scrutinized and reformulated. Hardy (2012) outlines three key areas of criticism: normativity and legacy media focus. Firstly, by "ideal-typing" the media systems, Hallin and Mancini assume the normative interpretation of the world and simplify the models from which real media systems will always diverge (Brüggemann et al., 2014). Secondly, contemporary analysts of media systems question the lack of discussion of new forms of digital media communication and reliance on outdated traditional media like newspapers. Hallin and Mancini's later work (2017) acknowledged these shortcomings and re-iterated that media systems are not static but evolve in response to internal factors, like governance and journalistic culture, and external forces, such as globalization and technological advances.

Despite the limitations, researchers have explored the heuristic value of media systems in studying human rights issues, such as press freedom. Maniou (2023) used the updated typology by Brüggemann et al. (2014) to investigate how internal (e.g., media self-censorship, harassment of journalists) and external (e.g., political polarization, political stability, economic growth) factors shape press freedom across different media systems. This study also identified the hybridization of media systems: traditionally Polarized Pluralist media systems became more liberal due to growing press freedom, while Liberal countries like the US, Australia, and Canada exhibited the characteristics of a Post-Communist media system due to high media self-censorship. Similar studies supported the hybridization of ideal types of media

systems, which is an intriguing starting point for comparative analyses (Brüggemann et al., 2014; Chadwick, 2013; Mattoni & Ceccobelli, 2018; Humprecht et al., 2022).

Even though disinformation is highly relevant to politics and media, applying media system theory to mis- and disinformation remains relatively limited. A notable exception is a study by Humprecht et al. (2020), who explored the dimensions of media systems that form resilience to disinformation and classified countries into three clusters. Interestingly, the most resilient cluster was a mix of states belonging to Democratic-Corporatist and Liberal types, while Southern European countries and the United States showed low resilience. Considering the value of media systems in studying political communication, Hardy (2021) argued for the integration of misinformation into media systems analysis by emphasizing the role of governance in shaping information environments. He proposed a comparative normative framework with six variables, among which is the role of the state that can be studied through "laws and policies affecting information and communications" (Hardy, 2021, p. 16).

Since the COVID-19 pandemic, political communication scholars have increasingly adopted the media systems lens to examine the role of the state in disinformation formation and dissemination dynamics. For instance, Janjić & Kleut (2022) discuss the effects of four key variables of media systems on disinformation resilience in the Western Balkans, while Arguelles and Lanuza (2021) produced two articles on the topic, one of which analyzed how disinformation is formed in different media systems across Southeast Asia, while the other categorized three general policy approaches to disinformation (Lanuza & Arguelles, 2022). Ultimately, they concluded that due to the differences in disinformation vulnerabilities in media systems, there is no uniform policy approach to accommodate all Southeast Asian countries. Understanding how communication practices are shaped in a given society is thus crucial for the future modeling of media policy.

These findings urge the necessity to explore and compare how nations belonging to different media systems safeguard (or risk) freedom of expression within their disinformation regulation. State actions against disinformation should support the right to express oneself, which calls attention to the types of responses the state may

utilize. This study adopts Bontcheva et al.'s (2020) taxonomy of responses to disinformation to explore how they differ across countries with different media systems. A key distinction of Bontcheva et al.'s typology is that it is organized around the aims and mechanisms of each policy rather than the institutional origin (e.g., state vs. digital platforms vs. civil society actors). While the original typology included 10 categories, Cipers et al. (2023a, 2023b) expanded it with the 11th category: (1) Factchecking and monitoring, (2) Investigative, (3) Countercampaigns, (4) Election specific, (5) Curational, (6) Technical and algorithmic, (7) Demonetisation, (8) Ethical and normative, (9) Educational, (10) Empowerment, and (11) COVID-19 specific responses. This thesis contributes to this body of literature by assessing the compliance of disinformation policies with human rights standards and comparing them both on a micro level (policy type) and the macro level (state and media system analysis).

To sum up, Hallin and Mancini's typology provides a foundational framework for analyzing the political and media environments of countries from a human rights perspective. Recent scholarship has emphasized the increasing hybridization within these categories due to digitalization (Hardy, 2021; Brüggemann et al., 2014). The media system model is widely criticized for its normative nature of looking for the way the world should be rather than seeing the world as it is. Instead of criticizing this aspect, this study embraces it: the normative assessment of government responses to online disinformation provides a fruitful ground "to stress test media systems" (Hardy, 2021, p. 5). This thesis examines whether countries cluster in similar ways when grouped by the indicators of risk to freedom of expression.
This thesis is guided by the main research question (RQ):

> **RQ:** How can government responses to disinformation be systematically assessed for their risk to freedom of expression?

To empirically test the conceptualized risk and use it in applied contexts, the following sub-questions are posed:

> **SQ1:** How well do media systems explain variation in countries' Risk to Freedom of Expression (RFE) scores compared to empirically derived clusters?
> **SQ2:** To what extent are particular types of responses, or the diversity of approaches used within a single policy, associated with the risk to freedom of expression?

3 Methodology

Having first identified what constitutes online disinformation and how disinformation governance is conceptualized, the previous chapters situated these issues within broader theoretical debates on freedom of expression. This chapter now turns to the methodological approach adopted to investigate how state responses to online disinformation align with international human rights standards. The first section introduces the dataset and justifies the country selection. The next section designs the assessment framework by justifying the purpose of the composite score measure developed to quantify the degree to which policies risk restricting free speech. The composite score consists of three operationalized concepts: clarity of definition, centralization of regulatory authority, and strictness of regulation. The third section details the usage of mixed methods in policy assessment and the design of the scoring system. Finally, the methodology chapter concludes with a ethical considerations that guided this research.

3.1 Data

This thesis employed secondary data analysis by building on Cipers et al. (2023a) global dataset that comprises public policy initiatives and laws set up by governments to respond to online disinformation. Secondary data analysis relies on the data collected by other researcher(s) and may be used for "replication, re-analysis, and re-interpretation" of the data (Johnston, 2014). Cipers et al. (2023a) themselves initially relied on the dataset by Bontcheva et al. (2020), which included 91 government initiatives. They expanded the database to 103 countries and 10 international organizations and updated it through the end of 2021. Both generations of research teams included legislative and non-legislative responses initiated by government or intergovernmental bodies (Bontcheva et al., 2020; Cipers et al., 2023a).

This study is a third-generation effort of this dataset application. I extracted a subset of 12 democratic countries from Cipers et al. 's (2023b) Excel-based dataset, enriching it in two ways. First, I updated the dataset with new policy initiatives through the end of 2024. Secondly, I enriched each policy entry with new variables the risk to freedom of expression. This approach enabled a novel comparative analysis not covered in the original study. My sample of countries, similar to

Humprecht et al. (2022), was based on the original typology outlined by Hallin & Mancini (2004) and enriched by countries of Central and Northeastern Europe and Oceania, thus comprising Finland, the Netherlands, Germany, the United States, the United Kingdom, Australia, France, Portugal, Spain, Hungary, Latvia, and Slovakia (Table 3.1).

The countries were selected to represent a diverse selection of media system types, following Hallin and Mancini's typology and supplemented with a cluster of Central and Northeastern Europe to capture transitional media environments shaped by shared legacies of state-socialist control and post-1990 reforms (Lauk, 2008; Dobek-Ostrowska, 2015; Boshnakova & Dankova, 2023). Table 3.1 exhibits the final sample and the key characteristics outlined by media systems literature.

Table 3.1: Country Sample and their Media Systems Characteristics

Media System	Key Characteristics	Countries
Democratic Corporatist	Strong public service broadcasting, press freedom, high political parallelism, strong journalistic professionalism[1]	Finland, Netherlands, Germany
Liberal	Market-driven media, low state intervention, high journalistic autonomy[1]	US, UK, Australia
Polarized Pluralist	Strong state involvement, elite-oriented press, weaker professionalism[1]	France, Portugal, Spain
Post-Communist	Historically state-controlled media, evolving regulation, varying levels of press freedom[2]	Hungary, Latvia, Slovakia

[1]Hallin & Mancini, 2004, p. 67; [2]:Maniou, 2023; Voltmer, 2011

In line with the best practices in secondary data research (Heaton, 2008; Johnston, 2014), the dataset was systematically updated and expanded to include new policy initiatives introduced between 2021 and 2024 using a two-level search approach: 1) primary analysis of official legislative documents, policy statements, press reports; 2) secondary analysis of policy reports from think tanks, national and regional legal research studies, legal commentaries, and academic articles. As Bowen (2009) claimed, such document analysis approach enriches the findings from original policy content with contextual and interpretative sources for triangulation, thus reducing bias and validating the analysis.

While secondary data analysis is seen as a valuable research approach "to advance knowledge across many disciplines through the use of quantitative, qualitative, or mixed methods data to answer new research questions" (Kelly et al., 2024; Polit & Beck, 2021), it presents its own limitations. As Greenhoot and Dowsett (2012) noted, one limitation is that researchers have no control over how the original data were sampled, defined, or measured, which might undermine the validity of subsequent research. To overcome this, I systematically reviewed every policy entry in the subset of 12 countries and excluded those that did not align with the analytical framework of this study using two filtering questions:

1. Is the initiative initiated at the level of national government?
2. Does the initiative contain directives that aim to fight disinformation?

After reviewing both primary and secondary sources on the policies, the ones that failed to answer positively to either of the two questions were excluded from the dataset. This reduced the number of policies from 63 to 50.

3.2 Building the Assessment Framework

To compare countries' disinformation responses, this study designed a composite score measure to evaluate the risk to freedom of expression within government responses to online disinformation by international human rights standards. By breaking down the content of the policy into three measurable concepts, we can assess the policy's risk to freedom of expression. Composite scores alike help assess the performance of countries in issues of interest and can be subsequently used for cross-country comparison (Nardo et al., 2008). This research method is well-established in public policy and international research. However, if poorly constructed, composite scores may "convey erroneous messages or interpretations" (Ibid, p. 5; OECD, 2008). To build a reliable composite indicator, the methodology of this framework was guided by the foundational work of Robert Adcock and David Collier, who provided a structured approach to the process of concept formation, operationalization, and scoring (Adcock & Collier, 2001).

A “systemized concept” refines the background concept(s) formulated by a scholar or group of scholars into a precise definition that can be consistently applied in measurement tasks (Adcock & Collier, 2001). The previous sections (2.2.1) outlined how the conceptualized dimensions were formulated and grounded in international human rights law standards. The second step of operationalization involved translating these systematized concepts into measurable indicators.

This study operationalized three dimensions into indicators: clarity of definition, centralization of regulatory authority, and strictness of regulation. The underlying construct that these indicators measure is the risk of limiting expression. Hence, the risk is a latent construct that is “dependent upon a constructivist, operationalist or instrumentalist interpretation by the scholar” (Borsboom et al., 2003). This means that risk to limit freedom of expression does not exist independently in the world but is theoretically constructed, conceptualized, and measured through its indicators, making it a formative construct. Lastly, the research methods and scoring process were transparently outlined and supplemented by the coding dictionary in the Appendices section.

Figure 3.1: Conceptual framework of the dimensions of Risk to Freedom of Expression (RFE).

Figure 3.1 illustrates the conceptual framework that foregrounds the construction of the RFE score. To assist the reader, this diagram shows how three operationalized dimensions outline the boundaries between illegal disinformation within the larger concept of general false expressions. Three dimensions of clarity, regulation, and strictness form a triangle containing illegal disinformation. The better these dimensions score, the clearer it is which content is illegal and punishable. On the

other side of the spectrum, we can witness that they fail to contain illegal disinformation when dimensions score worse. This mixes illegal disinformation with false but lawful expressions, hence the gray color and higher risk to freedom of expression. The next sections will describe the operationalization of the three key dimensions/sub-indicators of the composite score.

3.2.1 Operationalizing Clarity of Definition

Clarity is the first key determinant of precision and accessibility of a state initiative against disinformation, which is necessary for individuals to reasonably comprehend what is prohibited and act accordingly. This study operationalized the clarity concept through the following key concepts:

(1) whether the policy defines disinformation as provably false information,
(2) whether it includes intent to deceive,
(3) whether it narrowly identifies and justifies a harm or protected interest.

The coding was drawn on international human rights interpretations and the ICCPR Article 19(3) tripartite test, as discussed in the theoretical framework. After analyzing the primary policy document, each fulfilled criterion scored a point, resulting in a range of 0 (no clarity in definition) to 3 (full clarity). For further score construction, the clarity scores were reversed to represent the logic “the lower – the better” in line with two other dimensions and rescaled to a 0–12 scale.

3.2.2 Operationalizing Centralization of Regulatory Authority

This dimension captured who holds the power to define what constitutes disinformation. Higher centralization implies greater state control over defining truth, which may increase the risk of censorship.

Regulatory centralization was coded as follows:

- 3 = Statutory (state/government decide what is disinformation)
- 2 = Self-regulatory (platforms decide what is disinformation)
- 1 = Co-regulatory (independent regulatory bodies decide what is disinformation)

- 0 = No regulation (no one decides what is disinformation)

The values were derived from both primary policy documents and secondary literature surrounding the policy, ranging from 0 to 3 (the lower the score, the better). Similar to the previous dimension, they were subsequently rescaled to a 0–12 range.

3.2.3 Operationalizing Strictness of Regulation

Strictness assessed the severity of state intervention in controlling the spread of false information. It focused on how and to what extent lawful false content was targeted rather than penalties for illegal disinformation.

It was coded as:

- 0 = No punishment
- 1 = Administrative burdens (e.g., registration or transparency requirements)
- 2 = Content regulation (e.g., control, removals, or takedowns)
- 3 = Monetary fines
- 4 = Criminal sanctions (e.g., imprisonment)

Scores were assigned based on whether the official policy document enforced a regulatory mechanism. If no such mechanism exists, the score was 0, while the rest were scaled from 1 to 4 (the lower the score, the better). For future composite score construction, the strictness score was later rescaled to the 0-12 range.

3.3 Methods

3.3.1 Research Design

This thesis implemented a mixed-methods research design. Mixed-methods research has gained popularity due to the growing need for more complex research designs that deepen the insights from strictly quantitative or qualitative studies (Sandelowski, 2000). Health, behavioral, and social sciences often utilize this method to analyze qualitative data in an objective and quantifiable way, especially in comparative case analysis (Drozdova & Gaubatz, 2017). This data transformation inquiry is referred to as “quantizing” and involves “numerical translation or conversion of qualitative data, which enables quantitative precision in balance with narrative complexity”

(Sandelowski et al., 2009, p. 208). In the context of this study, as Landman (2003) notes, coding human rights into quantitative terms enables the assessment of formal legal commitments to the protection of freedom of expression. Furthermore, it is useful for drawing comparative inferences between different disinformation policy approaches. First, I performed quantitative policy content analysis. Then, I used the indicator scores as bases for my composite score construction. Lastly, I showed how the constructed score can be used for systematic comparison.

As Zeller et al. (1980) argued, while quantification transforms text into analyzable variables, such transformation entails a trade-off between specificity and comparability. To balance this conundrum and account for the subtleties of policy formulations, I leveraged quantitative content analysis, for which I manually coded disinformation policies, assigning scores for conceptually derived indicators: *clarity*, *regulation*, and *strictness*. There are two main types of quantitative content analysis: manifest, which focuses on counting explicit terms or categories (e.g., the number of times "disinformation" is mentioned), and latent analysis, which interprets the meaning "in, beneath and around the text" (Cardno, 2018, p. 633) to infer the latent content (Kohlbacher, 2006, p. 16). I utilized the latter, since the context was key to understanding abstract dimensions like quality of policy or, in this study, respect to freedom of expression encrypted within the policies.

Drawing from Paalman (1997) and Yanovitzky and Weber (2020), the analysis followed key steps in content policy research: defining objectives, sampling, constructing a coding scheme, conducting the coding, and analyzing patterns. Having defined objectives through building a theoretical framework in section 2.2, I built a preliminary coding rubric that guided the initial round of analysis. I first conducted a "piloting" or "skimming (superficial examination)" level of analysis in my Excel dataset to solidify the coding scheme (Bowen, 2009, p. 32), going policy by policy, assigning a score for each indicator, and providing sources from the policy text that substantiated the scoring decision.

Some policies in the original dataset by Cipers et al. (2023b) missed links for the policy document, so I filled the gaps as I went on. Since the content analysis is contextually dependent, I consulted secondary sources in the form of literature

surrounding the policy, like policy documents, legal commentaries, and academic research articles, which were also added to the dataset as a secondary source. This step provided context into how this policy was received locally and helped validate or challenge the surface-level interpretation of some of the vague policy's wording.

After finalizing the coding dictionary (see Appendix B), I did three rounds of iterations using primary and secondary sources. These revisions were necessary to ensure intra-coder reliability, which refers to the consistency in the scoring done by the same coder at different points in time (O'Connor et al., 2020). So, I returned to the dataset, re-did the coding bi-weekly, and measured the consistency rate. The final intra-rater agreement rate was 97.2%. Such technique is commended for its promotion of researcher reflexivity (Joffe & Yardley, 2003) and the overall improvement of transparency and systematicity of the coding process.

3.3.2 Scoring

The Risk to Freedom of Expression (RFE) composite score is based on the sum of indicators (*clarity* + *regulation* + *strictness*) and supplemented by two multiplicative weights that account for each policy's legal and temporal status. These weight indices were added to ensure the representation of policies' varying enforceability and operational status. Firstly, *legal status* was introduced under the notion that not all disinformation responses carry the same normative or regulatory weight; for example, law enforcement has more substantial legal implications than non-legislative government initiatives.

To account for this variation, *legal status* was operationalized based on a typology from the original dataset by Cipers et al. (2023b), which categorized each policy into five mutually exclusive types: (1) proposed legislation, (2) counter-narratives, (3) non-legislative initiatives, (4) adopted legislation, and (5) enforced law. These categories were assigned weights from 0.6 to 1.0 depending on their legal base: binding laws have greater normative force (Bayer & Bárd, 2020), while proposals and non-legislative instruments exert weaker power (Flew & Martin, 2022; Bradshaw & Lim, 2022). *Legal status* weight is conceptually different from *the regulation* dimension due to its focus on legal enforceability, not power over defining and controlling disinformation.

Secondly, this study assigned different weights to active and outdated policies to represent current disinformation governance. *Activity status* was assessed by examining whether each policy remains legally or practically in force at the time of analysis at the beginning of 2025, as procured from official government sources. Active policies received a full weight of 1.0, meaning they have a full effect on the resulting score. Outdated, repealed, or COVID-19-specific initiatives were weighted at 0.8, based on the observed average difference between active and inactive policies. Since this thesis collectively analyzed policies issued in the 2010-2024 timeframe, inactive policies were relevant as indicators of policy evolution and historical legislative environment despite no longer shaping current disinformation governance (Funke & Flamini, 2019). However, their effect on the risk score was discounted. The scoring rubric for the multiplicative weights is detailed in the Appendix B. Thus, each policy response was scored using a composite score formula based on the sum of three dimensions and multiplied by two additional weight variables. The formula was as follows:

$$RFE = (C + R + S) * L * A$$

Where:

- RFE = Risk to Freedom of Expression Composite Score
- C = Clarity score (0–12)
- R = Regulation score (0–12)
- S = Strictness score (0–12)
- L = Legal Status (0.6-1.0)
- A = Activity Status (0.8 or 1.0)

This formula synthesizes the policy content analysis, which is reflected by three indicators and scaled by legal enforcement context, reflected by legal and activity statuses. Clarity and Regulation were originally scored on a 0–3 ordinal scale, while Strictness was scored on a 0–4 scale. Guided by the OECD Handbook on Composite Indicators (Nardo et al., 2008), I linearly scaled all three of them to a 0–12 range while reversing the direction of the clarity score, since it was based on the principle "the higher, the better", while the other two were based on "the higher, the more restrictive". While these are ordinal variables, rescaling maintained the rank order and relative weight of each dimension while standardizing their equal contribution to the final score.

For further cross-national comparison, the RFE country score was aggregated by calculating the country's average composite score based on all RFE scores of the country's disinformation policies. This method facilitated a cross-country comparison while preserving the unit-level integrity of policy-level analysis, retaining transparency in how each policy contributes to national performance. The resulting country scores served as the basis for comparison across media systems, thus testing whether the media system typology would predict the grouping of countries according to their risk scores. The formula was as follows:

$$C_i = \frac{1}{n_i} \sum_{j=1}^{n_i} RFE_{ij}$$

Where:

- C_i = Average composite score (C) for country (i)
- H_{ij}= Risk to Freedom of Expression Composite Score (RFE) for policy (j) in country (i)
- n_i= Total number (n) of policies coded in country (i)

After the construction of the composite score, this study performed analysis using both policy-level and country-level RFE scores. This thesis employed descriptive statistical techniques (means and variances, distributions), validation tests (pairwise Variance Inflation Factors, and average inter-item correlation), and inferential statistical techniques (correlation and regression analyses). All of the aforementioned statistical analyses were conducted using RStudio (version 1.3.1093), an integrated development environment for R. I used various packages, including tidyr, dplyr, ggplot2, lme4, and car, to perform data cleaning, variable transformation, and model diagnostics.

In sum, Figure 3.2 provides a simplistic overview of how the policies (circles) issued by different countries (differing by color) were individually scored (triangles). Averaging policy-level scores resulted in country-level RFE scores (squares) that were subsequently used for cross-country analysis.

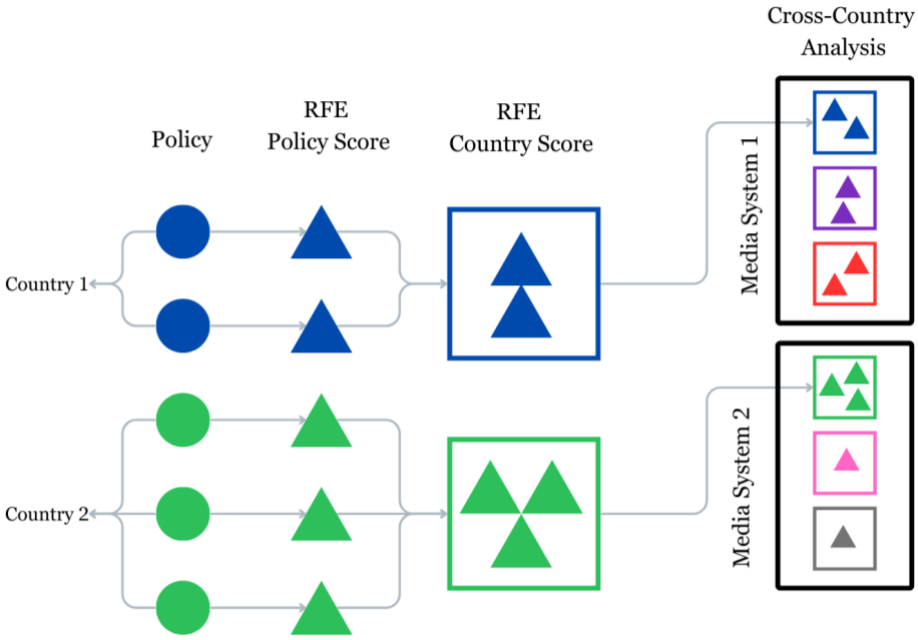

Figure 3.2: Conceptual framework for the RFE scoring process

3.3.3 Ethical Considerations

This thesis was written in observance of the principles of research ethics of the University of Helsinki (University of Helsinki, n.d). The analysis relied on the publicly available dataset by Cipers et al. (2023b) retrieved from the Harvard Dataverse repository, the source of which is cited accordingly in the References section. The original data did not contain personal sensitive data concerning human subjects, thus not requiring ethical review from the Research Ethics Committee. To update the dataset, I utilized existing publicly available policy documents and secondary data sources surrounding the policy, all of which were linked in the updated dataset. Striving for transparency and replicability, my updated version of the dataset and the R script report will be made publicly accessible. Methodological and analytical choices were justified and clearly outlined in Chapters 3 and 4, while the additional information about the coding dictionary, scoring rubrics, and assigned scores for each of the 50 policies is provided in the Appendices section.

4 Results and Analysis

In this part of the thesis, the results of the data analysis will be presented. The first section presents basic descriptive visual graphs to familiarize the reader with the data. It provides an initial descriptive overview of the policy dataset, including the timeline of the adoption of policies, the construction of the Risk to Freedom of Expression (RFE) score, and its validation through a three-step process: (1) convergent validity test, (2) indicator collinearity test, and (3) internal consistency test.

Subsequently, the RFE score was empirically tested using two sub-questions. First, I address the first sub-question by comparing RFE outcomes within traditionally established media systems and introduce the results of hierarchical cluster analysis as a more empirically grounded grouping alternative. The final part answers the second sub-question, analyzing how the type and diversity of disinformation responses relate to freedom of expression risks, concluding with regression results that identify which policy strategies most consistently align with rights protection. Exploration of these research questions ultimately proved that by conceptualizing three theoretically grounded indicators (clarity, regulation, and strictness) into a composite measure, we can meaningfully capture the risk of disinformation policies to limit expression.

4.1 Risk to Freedom of Expression

RQ: How can government responses to disinformation be systematically assessed for their risk to freedom of expression?

4.1.1 Construction of the RFE Composite Score

To explore this question, this thesis used a dataset comprising 50 government responses to disinformation across 12 countries issued between 2010 and 2024. Figure 4.1 illustrates the growth of governmental efforts to address the risks posed by disinformation throughout this period. Early adopters like Latvia (2010), Finland (2013), and France (2013) implemented foundational policies, with Latvia's Media Law Enforcement that focused on misleading content that threatens public health or could pose serious and grave risks of endangering it, while Finland and France's legislations set media literacy policies to be introduced into educational programs.

However, in 2016, the phenomenon of “fake news” became a part of the international agenda following the U.S. presidential election and Brexit referendum, which exposed vulnerabilities in digital information governance and spurred policy activity, marked by the first spike in 2019 (Cipers et al., 2023a). The number of policies skyrocketed with the COVID-19 pandemic due to many countries issuing legislation to tackle virus-related disinformation. We can witness a peak of activity in 2021-2022, with 10 out of 12 countries in the dataset issuing disinformation-related policies (Figure 4.1).

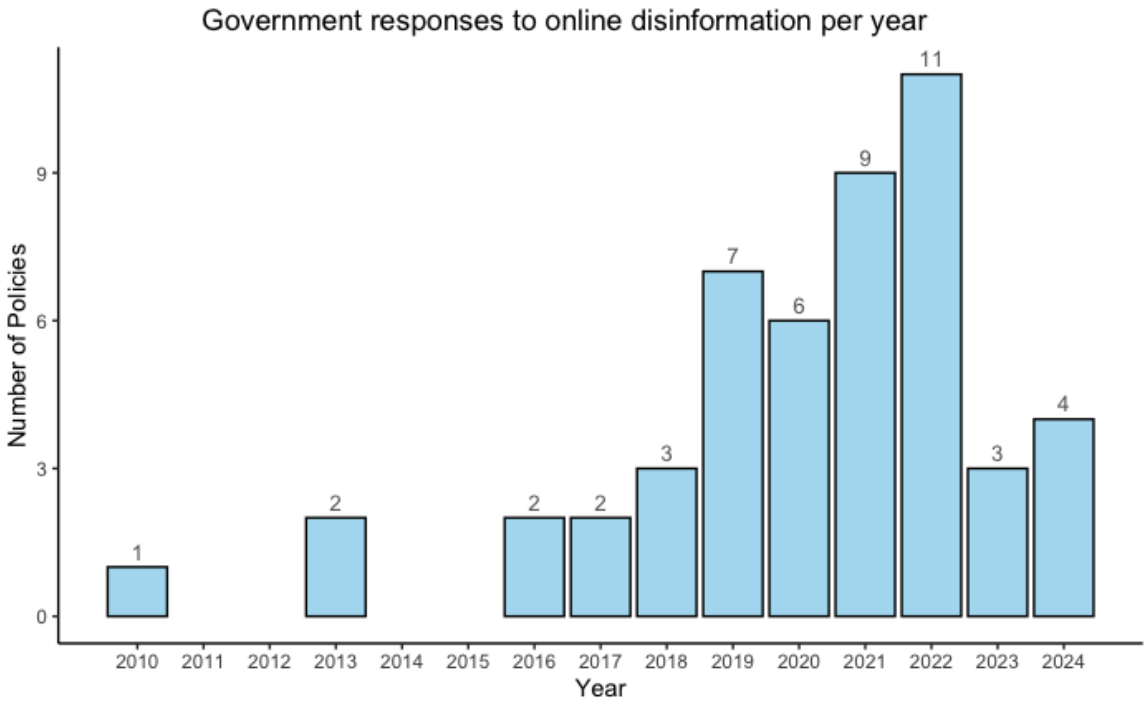

Figure 4.1: Government responses to online disinformation per year (2010-2024).

The Risk to Freedom of Expression (RFE) composite score was constructed to systematically evaluate the extent to which state responses to disinformation risk infringing on freedom of expression. Higher scores indicate a higher risk of rights-restrictive policy design. The score is based on three dimensions derived from international legal standards and prior research: clarity of definition (0–12, higher = more vague definition), centralization of regulatory authority (0–12, higher = more centralized government control), strictness of regulation (0–12, higher = harsher penalties).

Histograms (Figure 4.2) show the distribution of each indicator across all policies. Clarity skewed toward lower scores, with most policies providing full definition (n=22) or minimal definition (n=14), suggesting good overall compliance with the

tripartite definition of disinformation. Regulation scores were relatively balanced, leaning toward no regulation (n=16) and co-regulation (n=15) approaches. Strictness was the most skewed and showed a bimodal distribution, with most policies imposing no penalties (n=25) or content regulation measures (n=14).

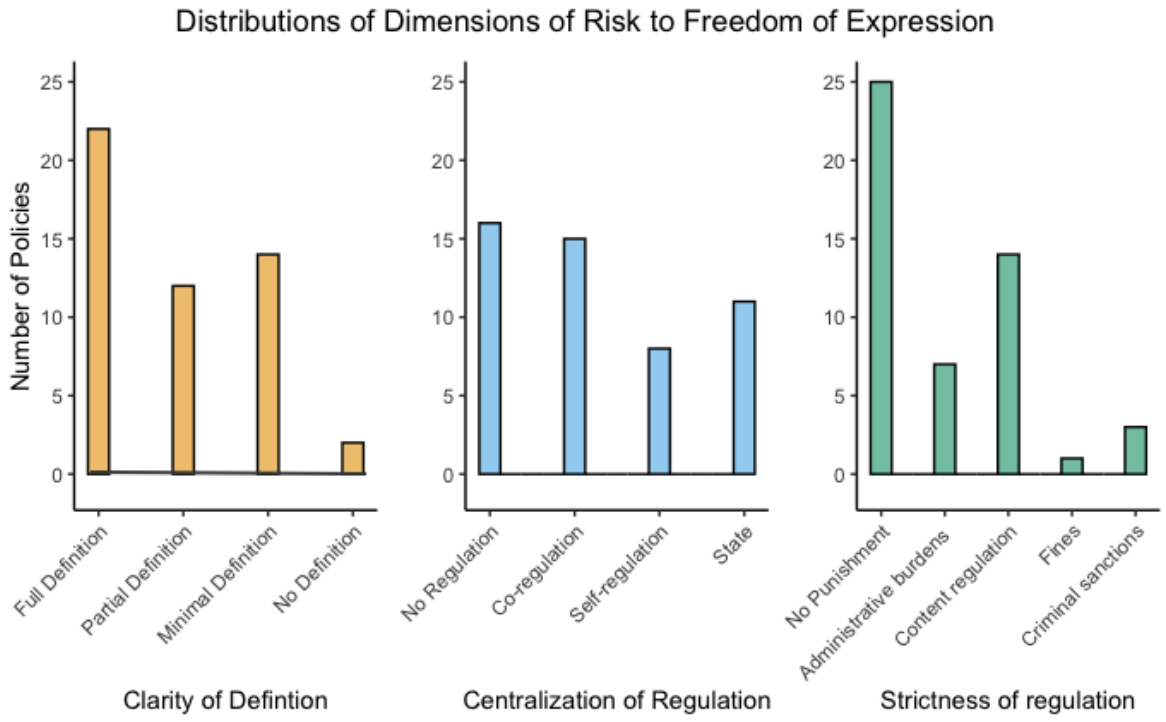

Figure 4.2: Distributions of government responses to disinformation across dimensions of Risk to Freedom of Expression (RFE).

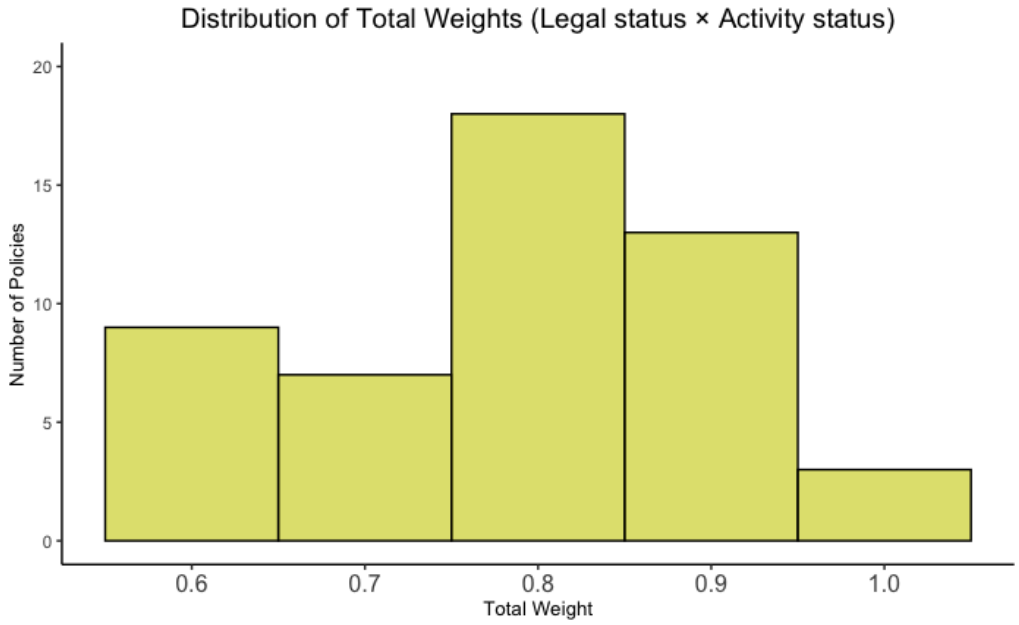

Figure 4.3: Distribution of total weights (legal status * activity status)

To adjust for varying enforceability and the current status of policies, each policy's raw score was multiplied by a product of two weights: *legal status* and *activity status*. The total weights ranged from 0.6 to 1.0. The distribution (Figure 4.3) is slightly left-skewed, with a moderate clustering around 0.8 and 0.9 values. This trend reflects that many policies were active and non-legislative in nature.

Thus, the sum of three dimensions results in a raw RFE score. The final scores were calculated by multiplying the raw RFE score by the weights described above (Figure 4.3). To illustrate the impact of the weighting of scores with legal and activity statuses, we can see the distribution of raw RFE scores and weighted RFE scores (Figure 4.4). Both distributions are right-skewed, displaying visible peaks at 0 (zero), indicating that nine policies carry no risk based on their policy content, comprising 18% of the sample. Weighting also centralized the distribution and compressed the range from a maximum raw RFE score of 32 to a maximum weighted RFE score of 25.6. See Appendix C for a detailed overview of each of the 50 coded policies, with a list of all government initiatives and scores for each of the three assessed dimensions and their respective weights.

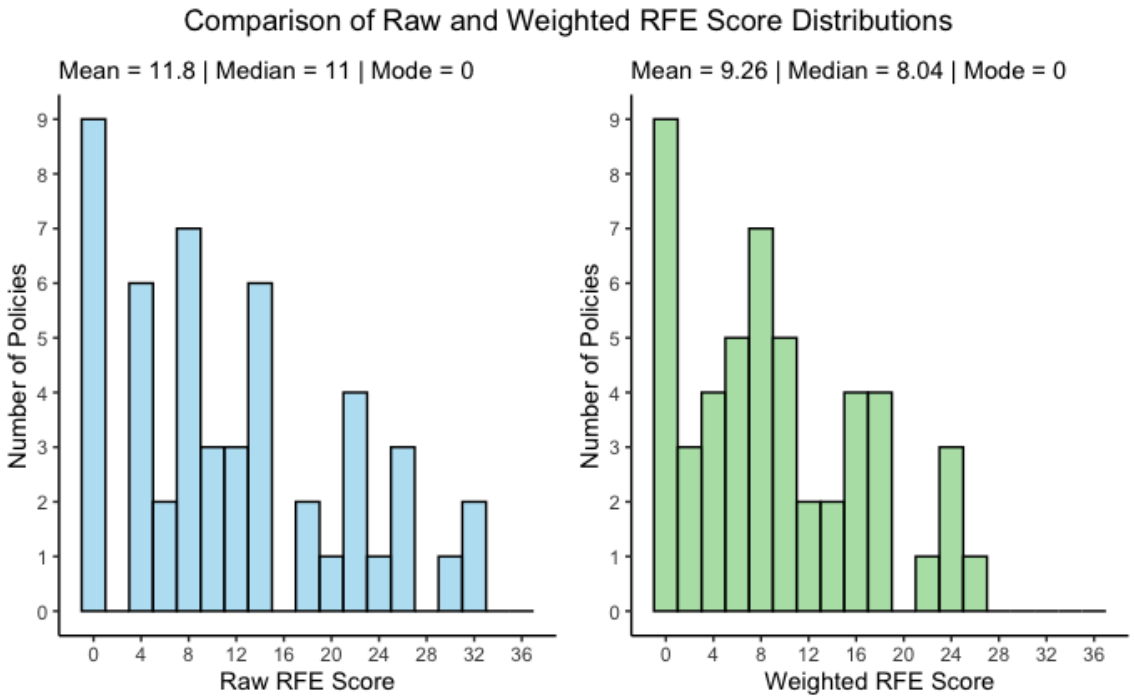

Figure 4.4: Comparison of raw and weighted RFE score distributions

The final step involved aggregating policy scores at the country level. Each country received an average raw RFE score and an average weighted RFE score, which would be used for subsequent statistical analysis. Figure 4.5 shows a scatterplot comparing the two. The diagonal reference dashed line indicates where raw and weighted scores would be equal. All countries fall below this line, indicating that weighted scores are consistently lower than raw scores. The color of the dots indicates how much the risk is reduced after weighting. This highlights how the application of legal and activity status substantially reduces the perceived risk posed by many policies, especially in countries like Hungary, where disinformation-related laws were repealed. Meanwhile, countries such as Latvia and France remained close to the reference line, indicating little change after weighting due to active and hard law status.

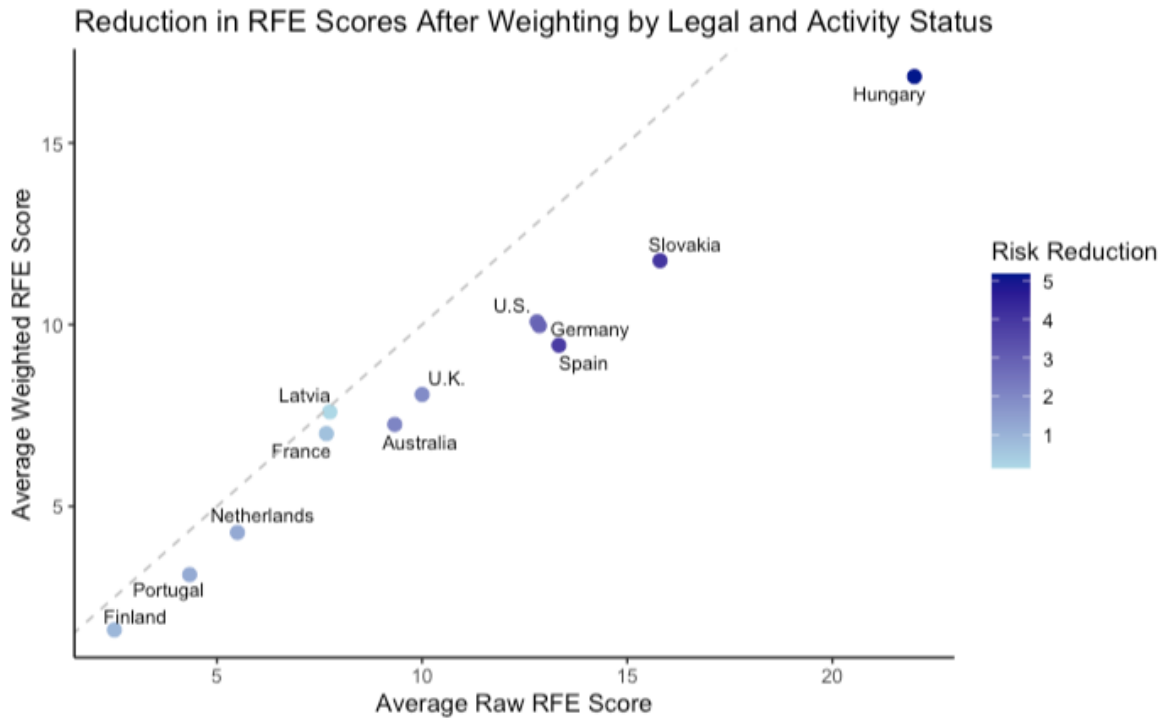

Figure 4.5: Scatterplot of the relationship between country average raw and weighted RFE scores. Countries below the diagonal dashed line exhibit a reduction in their RFE score after weighting by legal and activity statuses.

Note: The color gradient of country dots represents the magnitude of score reduction, with darker blue indicating greater reduction.

The mean weighted RFE country score among 12 countries was 9.09, while the scores ranged from 1.92 to 18.34. It is important to note that higher scores do not necessarily represent the entirety of the state's disinformation governance scene but

rather reflect the design features of the content of state policies that may conflict with international human rights standards. As shown in Figure 4.6, Hungary obtained the rather expected highest average RFE score of 18.34, indicating the highest risk to freedom of expression. Conversely, only three countries occupied the first quartile below five on the scale: Finland, Portugal, and the Netherlands amounted to the lowest scores of 1.92, 3.36, and 5.24, respectively, aligning with stronger safeguarding of freedom of expression.

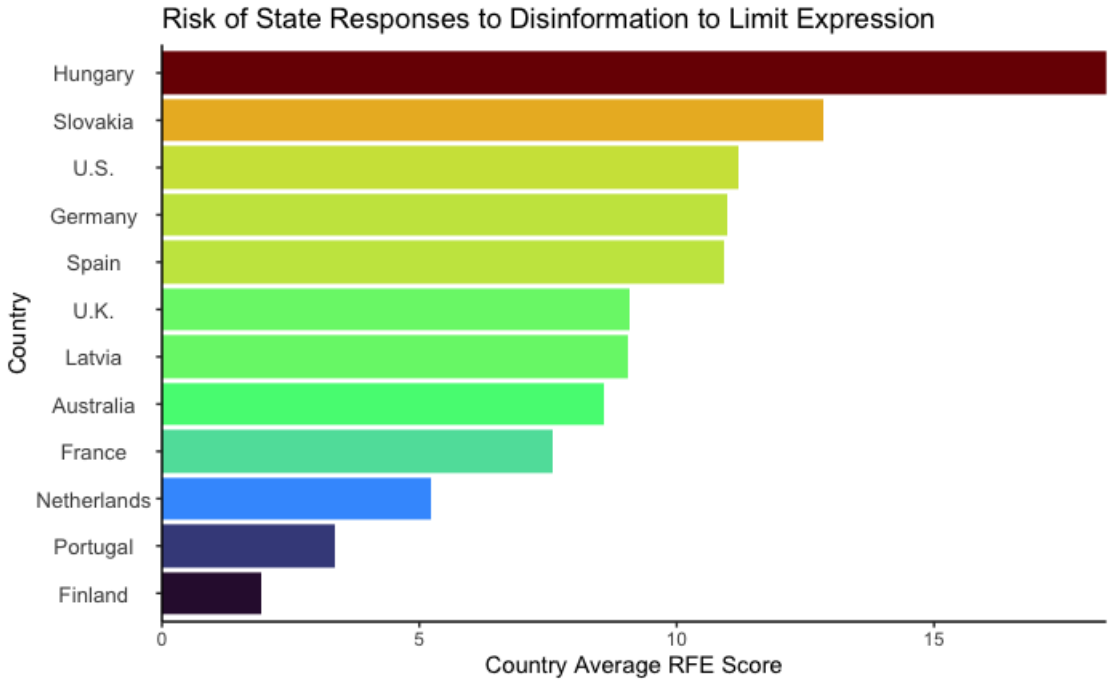

Figure 4.6: Risk of government responses to disinformation to freedom of expression (calculated from country average RFE scores)

4.1.2 Validation of the RFE composite score

A novel systematic assessment tool requires validation to confirm that the proposed conceptual dimensions meaningfully reflect the underlying construct of the risk to freedom of expression. To validate the RFE composite score, I followed a three-step process developed by Hair et al. (2021). Despite this framework being developed initially for partial least squares structural equation modeling (PLS-PM), a statistical technique for confirming measurement models for predictive purposes, this systematic approach is also applicable for validating my composite score. PLS-PM is a composite-based approach designed for formative measurement models, which my RFE composite score is. To assess the conceptual soundness of my constructed RFE

score, three tests are performed: 1) convergent validity, (2) indicator collinearity, and (3) internal consistency. Table 4.1 provides an overview of the results.

Table 4.1: Summary of the three-step validation process of RFE score

Step	Test/Procedure	Result	Interpretation
1. Convergent Validity	Spearman's correlation between RFE score and RSF Press Freedom Index	ρ = -0.39, *p* = .004	Significant negative correlation, supporting convergent validity.
2. Indicator Collinearity	Pairwise VIF tests between clarity, regulation, strictness	All VIFs < 1.58	No multicollinearity detected; dimensions contribute distinct information to the composite.
3. Internal Consistency	Inter-item correlations	$\bar{r} \approx 0.426$	All dimensions are moderately correlated and collectively represent a coherent underlying construct.

The first step of the convergent validity test verified whether the composite score measure correlates with theoretically related measures. If the RFE score truly reflects the risk to freedom of expression, it should negatively correlate with already validated measures, like press freedom. Reporters Without Borders Press Freedom Index 2024 (RSF) measures the degree of freedom available to journalists, which includes dimensions of legal, political, and economic freedom closely related to the protection of other human rights like the right to freedom of expression (Reporters Without Borders, 2024; UNESCO, 2023).

Due to the non-normal distribution of data and small sample size, I ran Spearman's correlation analysis (Hauke & Kossowski, 2011) to test the relationship between press freedom levels from 2024, similar to the original study by Cipers et al. (2023) that used the most recent country scores, and country-level RFE scores, comprised of data from 2010 to 2024. A Spearman's rank correlation showed a significant negative association between countries' average RFE scores and their 2024 RSF Press Freedom scores ($\rho = -0.3939$, $p = 0.0046$), suggesting that states with lower levels of press freedom have a history of disinformation policies with higher risks to limit speech (Figure 4.7).

The scatterplot (Figure 4.7) shows a clear downward trend: as average country-level RFE scores increase, RSF scores decrease, signifying diminished press freedom. The plot shows outliers like Germany, who scored higher RFE scores despite high press freedom, and the US, with lower press freedom and lower than expected RFE score. In total, the correlation test shows that my composite measure behaves as expected relative to an established measure.

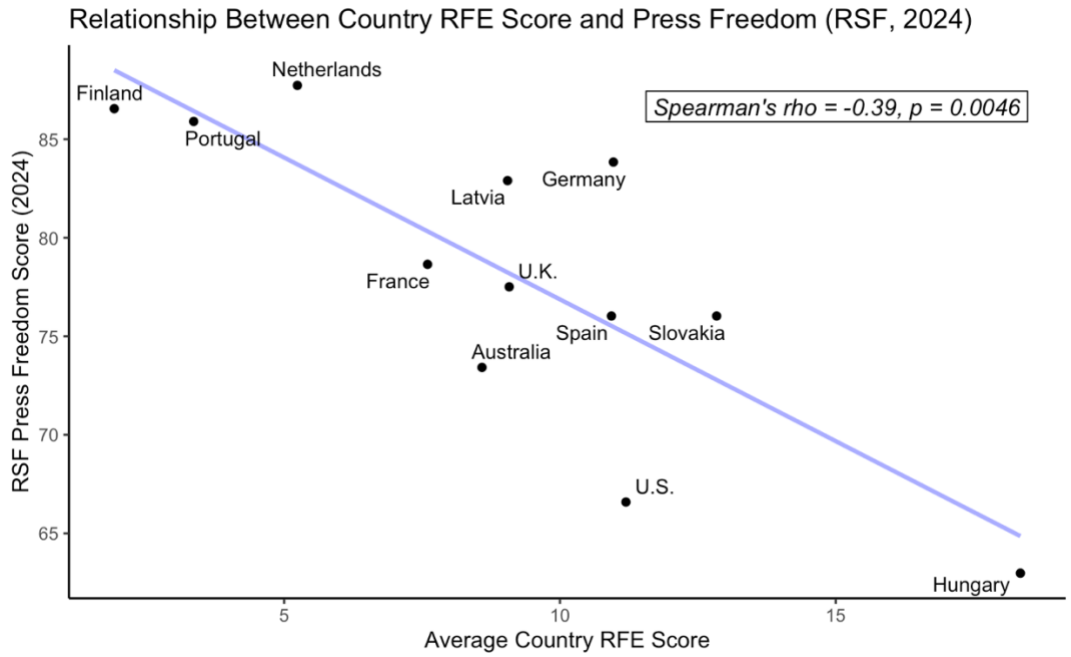

Figure 4.7: Scatter plot of the relationship between RSF Press Freedom Index (2024) and the Risk to Freedom of Expression (RFE) score.

The second step of indicator collinearity assessed the degree of correlation between the three dimensions (clarity, regulation, and strictness). Hair et al. (2021) stipulated that the high collinearity of two or more indicators indicates that they essentially depend on each other, increasing standard errors and making coefficients unstable and challenging to interpret. Variance Inflation Factors (VIFs) were used to examine indicator collinearity among the three dimensions. All VIF values were between 1.04 and 1.58, thus well below the critical threshold of 5, indicating that no serious multicollinearity threatens the interpretability of individual indicator weights. This result (Table 4.1) shows that the RFE composite score is conceptually balanced, and each indicator contributes independently, supporting the validity of my measurement model.

Lastly, to evaluate internal consistency, I tested the average inter-item correlation. Based on the results, the composite score is reliable due to the score of 0.426, which

falls within the recommended range of 0.15 – 0.50 (Hair et al., 2018). Cumulatively, these steps confirmed the conceptual coherence of the constructed RFE score, making it a valid tool for evaluating the risk to free speech disinformation policy design, thus answering the main research question. By combining these three theoretically grounded dimensions into a composite measure, we can effectively capture speech restrictiveness in disinformation policy.

4.2 Risk Across Media Systems

SQ1: How well do media systems explain variation in countries' Risk to Freedom of Expression (RFE) scores compared to empirically derived clusters?

4.2.1 Ideal Media Systems

To commence the empirical testing of the usability of the RFE score measure, this thesis inspected the variation in national levels of risk among countries of different media systems models. Media system typology originally formulated by Hallin & Mancini (2004) has been imperative to comparative media studies. The typology has been criticized and refined widely since then, and adapted to broader geographical contexts and disciplinary applications. Hallin and Mancini (2013) themselves acknowledged some of the criticism a decade later and reiterated that their work was a stepping stone to new explorations of similarities and differences in the development of media systems.

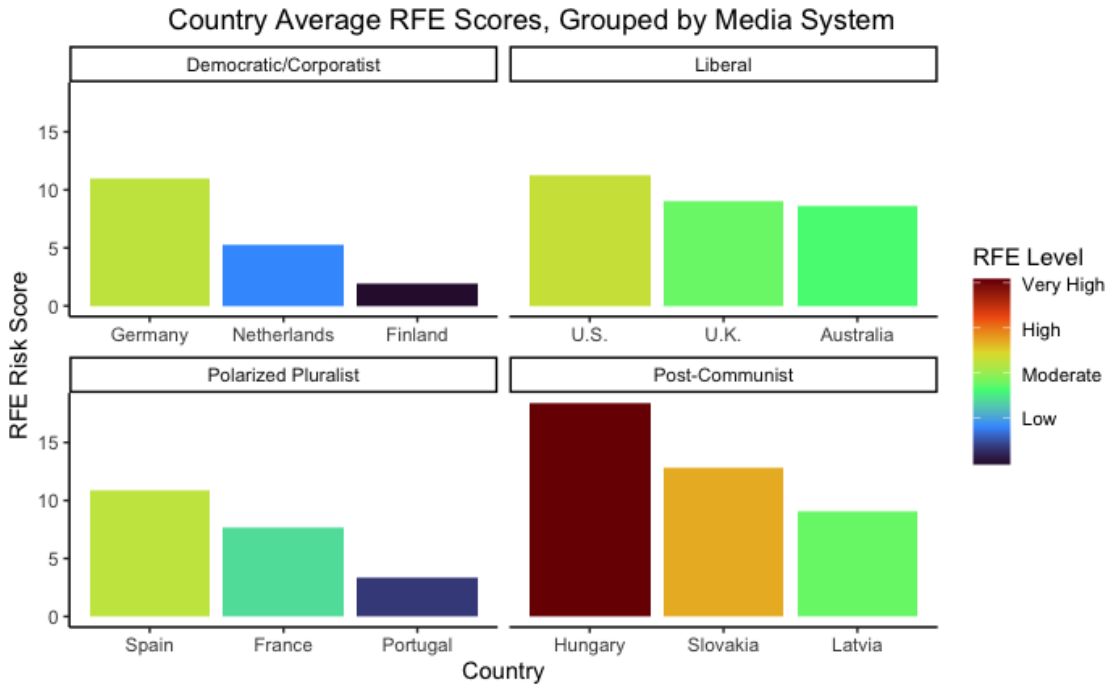

Figure 4.8: *RFE Scores by Country, Grouped by Media System*

This study harnessed an enriched media typology that includes the Post-Communist cluster for additional robustness. As a part of the SQ1 that focused on testing the media system typology in the content of government responses to disinformation, I have drawn a cross-country analysis to compare RFE scores among 12 countries of four media systems. In theory, countries with the same media system model would exhibit similar policy approaches due to the 'state intervention' dimension. Initial visual assessment (Figure 4.8) revealed inconsistencies in country scores within and between media systems.

The most discernable outlier was observed within the Democratic-Corporatist media system, with Germany scoring 10.97 compared to the average scores of Finland and Portugal, with 1.92 and 3.36, respectively. Germany's elevated score was primarily driven by the currently inactive Network Enforcement Act (NetzDG), which mandated social media platforms to swiftly remove 'obviously illegal' content or face substantial fines, thereby centralizing regulatory power at the hands of digital platforms (regulation score 2) to remove false content in fear of fines (strictness score 2).

Similarly, the Liberal media system exhibited an outlier, albeit less pronounced than in the Democratic Corporatist system (Figure 4.8). The U.K. and Australia's disinformation policy responses were on a moderate risk scale. Both countries' legislations tend to clearly outline what constitutes disinformation while mostly giving the decision-making power to digital platforms or independent regulatory bodies. The relatively high RFE risk score of the United States (11.19) stemmed from inconsistencies in the formulation of disinformation and a combination of policies involving statutory regulation, like the Global Engagement Center (2016) or the Honest Ads Act (2017).

Hungary's Criminal Code Amendment (2020) has been notorious for its criminalization of spreading "false or distorted facts" during emergencies, with penalties of up to five years in prison (Cox, 2020). This law scored 1 for clarity, 4 for strictness, and 4 for centralization of regulatory authority. Slovakia's moderate-to-high RFE score stemmed from a series of statutory laws, including the Cybersecurity Act (2018) and its 2022 update, as well as the Media Services Act (2022), which enable content blocking and regulate "serious misinformation" without clear legal

definitions, reflecting centralized statutory enforcement and content regulation. On the contrary, the Latvian scored visibly lower on the risk in comparison to the other two countries. Latvian recent proposal to amend the Criminal Code to include an article on "Influencing the Electoral Process Using Deep Fraud Technologies" (2024) was praised for its timely effort to tackle cyber fraud to secure electoral integrity (Dockrell, 2024).

The Polarized Pluralist media also contained an outlier, with RFE scores ranging from 3.36 in Portugal to 10.93 in Spain (Figure 4.8). Portugal's low score can be explained by its focus on soft governance measures, which are state-led but non-punitive. Both French and Portuguese legislation scored high on clarity on the definition of disinformation, while Spain's largely non-punitive approach lacked consistency in its formulations. Overall, the Polarized-Pluralist countries appear to have scored well in clarity, yet their high variance in centralization of regulatory authority dimension contributed to their medium scores on the risk scale.

Table 4.2: Descriptive statistical analysis of three dimensions of RFE score across media systems.

Note: The table displays the original scores distributions: clarity (0-3, higher = better), regulation (0-3, higher = worse), and strictness (0-4, higher = worse).

Media system	Ave. clarity	Var. clarity	Ave. regulation	Var. regulation	Ave. strictness	Var. strictness
Democratic/ Corporatist	2.26	0.79	1.06	0.63	0.66	0.80
Liberal	2.21	0.95	1.42	1.49	1.14	1.36
Polarized Pluralist	2.23	0.69	1.00	1.50	0.78	0.94
Post-Communist	1.58	0.99	1.58	1.90	1.41	2.62

Further basic descriptive statistical analysis of the individual three dimensions revealed that countries within media systems do not follow a uniform approach to disinformation governance, which contradicts the assumption that media system typology would predict patterns in policy approaches. For instance, Table 4.2 shows the Post-Communist media system group exhibiting a regulation score variance of

1.90 and a strictness variance of 2.62, meaning that countries classified under this category have vastly different levels of regulation of disinformation, ranging from direct state control to a complete absence of regulation. Similarly, Liberal and Polarized Pluralist media systems show overall regulation score variances of 1.49 and 1.50, respectively, indicating a lack of consistency. Such high within-group variance undermines the explanatory power of media system typologies in predicting policy approaches.

Finally, a series of statistical analyses solidified that media systems typology did not meaningfully explain how countries approach freedom of speech in disinformation governance. Firstly, I ran the Kruskal-Wallis rank sum test to check for the differences between media systems, namely whether at least one media system differed significantly. Kruskal-Wallis showed no statistically significant differences (p = 0.2913). Secondly, to check for variations within media systems themselves, I ran Levene's test that showed no media system was notably more internally consistent than others, i.e., RFE scores were not that similar within groups (p = 0.5821). Ultimately, when assessing how countries design policies to combat disinformation in terms of their compliance with international freedom of expression standards, the conventional media system typology did not predict the grouping of countries based on their risk to freedom of speech. These findings indicate that media system typology alone does not sufficiently explain how different countries formulate disinformation policies.

4.2.2 Empirically Derived Clusters

Given these results, this thesis employed hierarchical cluster analysis to extract more meaningful groupings than pre-existing media system models. This approach was encouraged by the original authors Hallin and Mancini to further explore and reconfigure the typology based on the research objectives: "A system is not an entity with a fixed set of characteristics, but a pattern of variation" (Hallin & Mancini, 2017, p.167). Hierarchical clustering was performed using the RFE score and score variables of three key dimensions. Due to the small sample size, I chose agglomerative clustering that operates on a bottom-up principle: it starts with individual points that merge into clusters (Weigand et al., 2021). Additionally, I chose

the Manhattan distance since it is particularly effective due to its robustness to outliers and its application in small datasets like mine (Murtagh & Legendre, 2014).

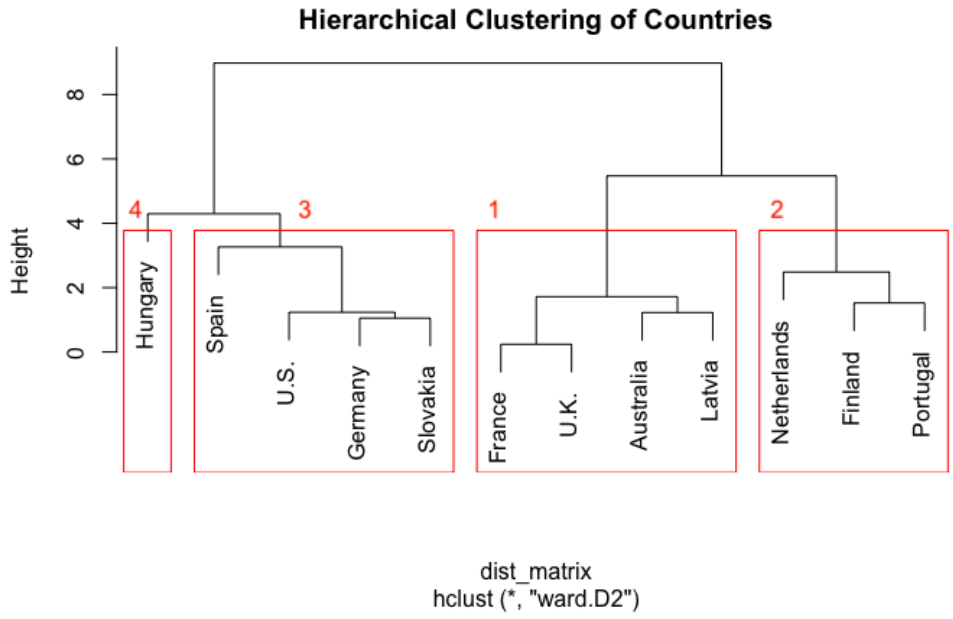

Figure 4.9: Hierarchical clustering dehdogram by RFE score dimensions.

Note: Cluster 1 (France, the U.K., Australia, Latvia); Cluster 2 (the Netherlands, Finland, Portugal); Cluster 3 (Spain, the U.S., Germany, Slovakia); Cluster 4 (Hungary).

The resulting clusters differed from the previously conceptualized media system typologies (Figure 4.9). The resulting empirical clustering suggests that the design of disinformation policy differs among countries with the same ideal media system. This hints at a growing divergence in regulatory responses to online disinformation, even within countries of similar media systems landscapes. I conducted a series of cluster validation statistics to evaluate hierarchical clustering groupings, proving they offer a more meaningful grouping than the original media system typology, which can be consulted in Appendix D.

Table 4.3: Descriptive statistical analysis of three dimensions of RFE score across empirical clusters.

Note: The table displays the original scores distributions: clarity (0-3, higher = better), regulation (0-3, higher = worse), and strictness (0-4, higher = worse).

Cluster	Average clarity	Mode clarity	Average regulation	Mode regulation	Average strictness	Mode strictness
1	2.56	3	1.31	1	1.06	0
2	2.72	3	0.73	1	0.27	0
3	1.50	1	1.45	0	1.10	2
4	1.00	1	2.00	3	2.67	4

Cluster 1

The first cluster consisted of Liberal (Australia and the U.K.), Polarized Pluralist (France), and Post-Communist (Latvia) media systems. The cluster was characterized by high levels of clarity (2.56/3.0), co-regulation (1.31/3.0)., and a moderate level of punishment for spreading false information (1.06/4.0) (Table 4.3). The high clarity score implies policies in this group tend to clearly outline the triad of falsity, intent, and harm(s). A deeper look into descriptive statistics reveals that the most frequent regulatory approach (mode) was 1, indicating many policies involve independent regulatory and judicial bodies to decide on acts of disinformation.

The disinformation approaches of this cluster show that only the policies focusing on foreign or electoral interference tend to have higher levels of regulation and stricter punishment mechanisms, suggesting a "securitization" of disinformation approaches (Casero-Ripollés et al., 2023). Casero-Ripollés et al. (2023) claimed the securitization approach applies hard power and frames disinformation as a cardinal threat to security and democracy. This approach was utilized in policies like the U.K.'s Online Safety Act (2023) and Foreign Interference Offence (2023) that target harmful and state-sponsored disinformation with obligations for platforms to act, imposing penalties of up to 14 years in prison. Similarly, the Latvian Parliament recently passed amendments to the Criminal Code that criminalize the creation and/or distribution of information using deepfake technologies to influence elections (2024).

The counter-disinformation measures that did not address foreign or election interference were far less punitive and more closely corresponded to international human rights standards. The risk to freedom of speech was lower among policies like Latvia's government-backed anti-disinformation platform (2023), Australia's COVID-19 Mythbusters (2021), and France's Media and Information Literacy (EMI) initiative (2013) (McGowan, 2021). Thus, the countries in this cluster employ a hybrid strategy: States assert stricter regulatory control in the name of security during elections or foreign interference campaigns while clearly outlining the cases of illegal disinformation. Yet, to differentiate illegal disinformation from false but lawful information, they rely on administrative reporting measures and content correction, which have the potential to censor speech.

Cluster 2

The next cluster consisted of two Democratic-Corporatist states, Finland and the Netherlands, and Portugal, originally belonging to the Polarized-Pluralist media system. The countries within this cluster held the lowest scores in RFE, as evidenced by their high clarity scores (average = 2.72; mode = 3) and a dominant regulatory approach of co-regulation (average =0.73; mode = 1) (Table 4.3). In contrast to the previous grouping, the approach to disinformation regulation is much softer and focuses on resilience-building, which aligns with the Nordic and Western European media regulation model (Karppinen, 2013). Governments maintain oversight rather than full control, while the decision-making power over disinformation is given to independent regulatory bodies and citizens themselves who report disinformation on digital platforms. The policymakers here do not punish false information, with a strictness average of 0.27 and mode of 0 (Table 4.3).

Government-wide strategies of this cluster primarily rely on educational, empowerment-based, and non-punitive approaches that retain the principle of protection of fundamental rights. Some countries, like Finland, largely practice no control in defining disinformation while prioritizing media and digital literacy through the Media Education Policy (2013) and its National Emergency Supply Agency's Knowledge Centre (2022) (Salomaa & Palsa, 2019; Sillanpää, 2021). Other countries, like The Netherlands, combine different approaches. In National Strategy Against Disinformation (2022), the Dutch government explicitly stated that the government alone cannot determine what reliable information is or is not. In its Updated Government Strategy to Combat Disinformation (2024), the Dutch government announced its plan to set up a "reporting facility" for citizens to report disinformation on digital platforms, which would be decided by the independent out-of-court dispute resolution bodies (Fathaigh, 2024).

Portugal introduced a similar policy in 2021, yet it faced significant criticism. The Charter of Human Rights in the Digital Age (2021) set out rights and duties of conduct in the digital environment, including a controversial Article 6 (the right to protection against disinformation) that would grant power to the Media Regulator to decide on the acts of mis-/disinformation. The policy sparked legal and public

criticism due to concerns over the vague definition of disinformation and its prospects to censor speech (Farinho, 2021; Soares, 2021). Thus, Article 6 was partially repealed in 2022, highlighting Portugal's responsiveness and reinforcing the cluster's overall pattern of freedom-preserving disinformation governance (Baptista & Morgado, 2022).

Cluster 3

Cluster 3 was the most diverse in its composition, comprising countries from all four media systems: Germany, the U.S., Spain, and Slovakia. While at first glance these countries seem politically and institutionally diverse, these countries share a similarity in their approaches to disinformation. Similar to Cluster 1, they often frame disinformation as a national security issue. The difference lies in its reactive and top-down approach, which is further undermined by the vagueness of definition and its focus on controlling flows of harmful content. The cluster's main feature is the vagueness and inconsistencies in policy formulation (average = 1.50, mode = 1) (Table 4.3). While the dominant regulatory approach (mode) is 0, or "no regulation", Cluster 3 had the highest average centralization of regulatory authority score of 1.45 among all groups. The difference between the mode and average revealed the presence of a couple of highly centralized statutory policies. Most of these policies inflict administrative burdens and/or content restrictions, as shown by moderate strictness scores (average = 1.10, mode = 2).

Vague formulation combined with centralized content regulation suggested inconsistencies in disinformation governance and the potential to undermine democratic processes (Napoli, 2018). In Germany, regulations like NetzDG outlined the types of unlawful disinformation while it was in force, yet it punished "underblocking" with fines while leaving "overblocking" unsanctioned, which qualified the policy for strictness of regulation of 2 (content regulation). In Slovakia, several initiatives, including the Cybersecurity Act and the Media Services Act (2022), lack consistent definitions for disinformation but authorize either administrative or judicial bodies to block or regulate content. Spain follows a similar path. The Procedure for Intervention Against Disinformation (2020) places decisions in the hands of national security bodies, highlighting the centralization of regulatory

authority. The U.S. policies pioneer a more self-regulatory approach to disinformation, thus essentially delegating the role to platforms, yet the definitions of disinformation often lack consistency. The efforts that did provide clear formulations on disinformation were heavily criticized and repealed, like the Disinformation Governance Board (2022), which had a life span of 3 weeks, or the Global Engagement Center (2016), which was recently dissolved amidst censorship allegations (Gedeon, 2025). Altogether, Cluster 3 combines content-level regulation and inconsistencies in legal clarity, producing a favorable landscape for self-censorship.

Cluster 4

Hungary has formed its own cluster due to its highly restrictive and vaguely formulated policies, with the highest average strictness score (2.67) and consistently low clarity scores (average and mode = 1) (Table 4.3). Both its COVID-19 misinformation-related laws from 2020 criminalized the dissemination of vaguely defined “false or distorted facts” during times of emergency with up to five years in prison. Hungary’s draconian disinformation laws, state centralization of power over falsehoods, and vague formulations suppress alternative voices in fear of imprisonment, all separating it from other countries.

By operationalizing the risk to freedom of expression (RFE) within disinformation policies, this thesis revealed novel groupings of countries sharing commonalities in their regulatory approaches. The countries grouped within Cluster 2 exhibited the lowest levels of risk to speech with their disinformation regulation. Members of Cluster 1 followed with low to moderate scores due to its co-regulatory model preference and high clarity of policies. What drove the scores in this cluster was the presence of policies with a securitization focus that tend to have a statutory form of control and punitive measures. Next, we have Cluster 3, with moderate to high levels of risk due to inconsistent levels of clarity and regulatory approaches. Lastly, Hungary, the sole occupant of Cluster 4, stood out for its highly punitive measures and lowest clarity in definitions, which grossly contravene international human rights standards on freedom of expression.

4.3 Risk Across Types of Policies

SQ2: To what extent are particular types of responses, or the diversity of approaches used within a single policy, associated with the risk to freedom of expression?

Drawing on Bontcheva et al.'s (2020) taxonomy of disinformation responses, I analyzed how the type and diversity of policy responses correlate with states' risk to freedom of expression (RFE composite score). See Appendix A for the full description of online disinformation government response types. Bontcheva et al. (2020) specifically outlined this typology's potential for analyzing the impact of disinformation policy responses on fundamental rights, particularly freedom of expression. For simplicity, this taxonomy is further referred to as *policy types*.

4.3.1 Diversity

First, I performed a series of country-level analyses among 12 countries. In Cipers et al.'s (2023a) study, the diversity of policies was assessed using a coverage-based approach that counted "the number of categories from the typology that were marked 'yes' for at least one initiative in the country." However, the Pearson's correlation result was not statistically significant (p = 0.908). The lack of correlation in my sample suggests that the broader policy coverage does not automatically predict better human rights compliance. For example, some of the countries with low diversity scores like Portugal (coverage = 3) and Hungary (coverage = 5) had drastically different RFE scores of 3.36 and 18.34 respectively (Table 4.4). This suggests that looking into the individual RFE responses might better explain the association of diversity scores with RFE score.

Focusing on the individual policy scores, I ran a Pearson's correlation test using unweighted RFE policy scores (without the multiplicative weights of legal status and activity status) to isolate the effects of three dimensions. This test revealed a statistically significant negative relationship between policy diversity and RFE score (cor = -0.2740, p = .05), indicating that, on average, more multifaceted policy interventions tend to safeguard freedom of expression better, regardless of the country in which they are implemented. To illustrate this effect, the scatterplot (Figure 4.10) shows the relationship between diversity and risk score. As we can see, the negative relationship is relatively modest on the policy level (-0.27).

Table 4.4: Diversity and average RFE scores by country.

Country	Average RFE	Diversity Score
Australia	8.58	11
U.K.	9.08	11
U.S.	11.19	10
Latvia	9.05	8
Finland	1.92	7
Netherlands	5.24	7
Spain	10.97	7
France	7.60	6
Germany	10.97	6
Slovakia	12.84	6
Hungary	18.34	5
Portugal	3.36	3

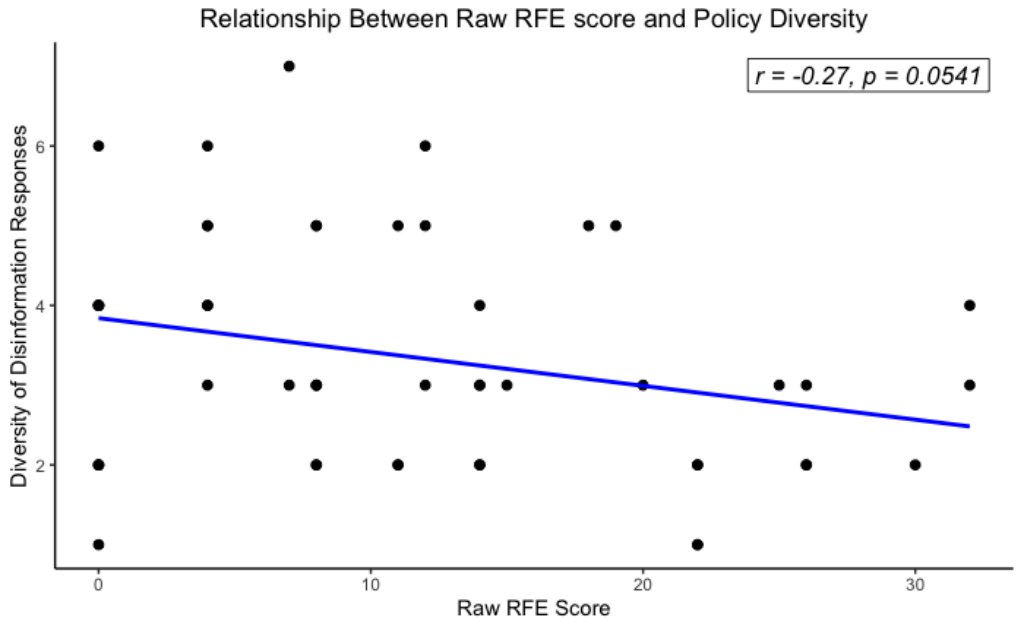

Figure 4.10*:* Scatterplot of the relationship between raw RFE score and policy diversity scores.

The small sample of countries possibly limited the applicability of diversity of policy approaches on a national level. This result reinforced the need to examine which specific types of policies were driving the risk up or down. The following section investigates this question by analyzing the relationship between each policy category (e.g., educational, investigative, curational) and the degree of rights protection it tends to afford.

4.3.2 Policy Types

To what extent are particular types of responses associated with the risk to restrict expression?

Finally, let's identify which types of government responses significantly affect the risk to freedom of speech. Before conducting statistical analysis, we need to look at the distribution of policy types. As one can see from Figure 4.11, the most common policy type is Ethical/normative, while policy categories like Economic/Demonetizing and COVID-19-specific regulations are sparse, with 5 and 4 policies, respectively. This comes as no surprise since ethical and normative responses involve public condemnation of acts of disinformation "designed to embed values and actions at the individual level that can help counter the spread of disinformation" (Bontcheva et al., 2020, p. 203). Any legislation on disinformation would contain some degree of condemnation of disinformation to underscore its legitimacy.

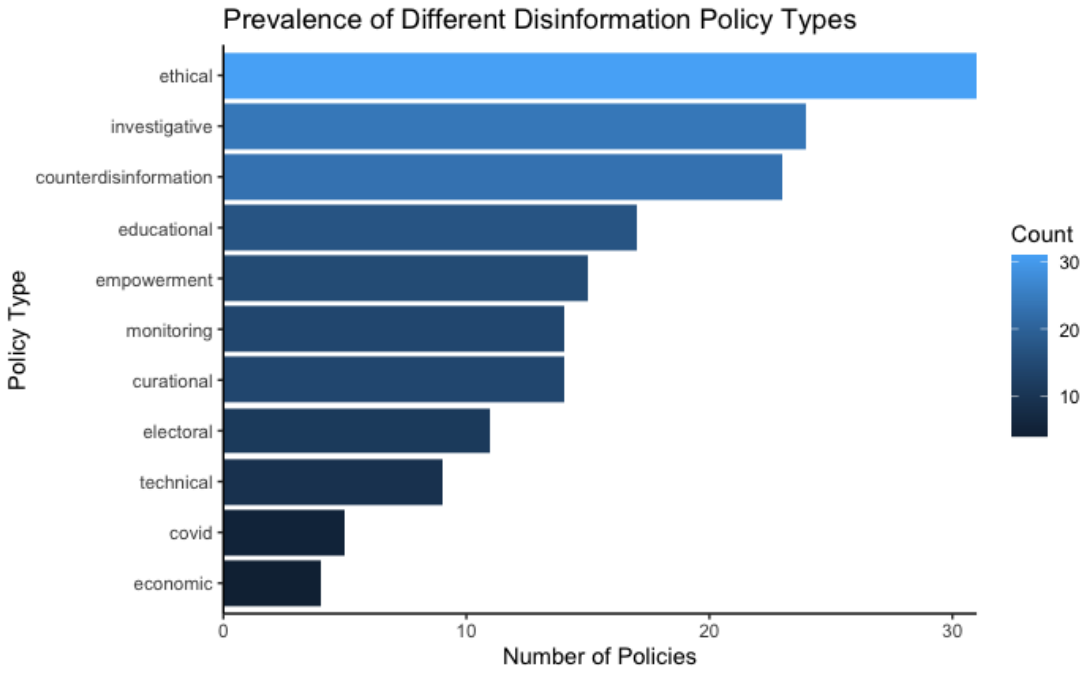

Figure 4.11: Prevalence of types of government responses to disinformation in the dataset.

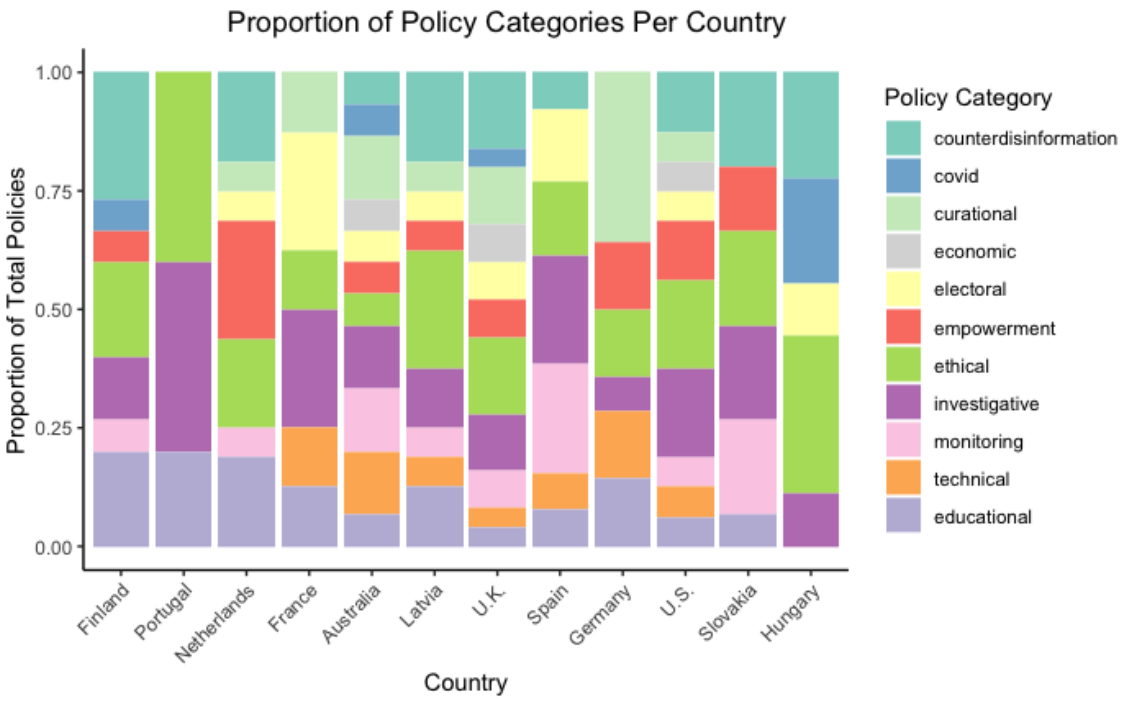

Figure 4.12: Bar charts of proportions of policy categories per country.

Note: Countries on the horizontal axis are ordered from the lowest (left) to highest (right) average RFE scores.

Figure 4.12 displays the proportions of policy categories within each country, with the horizontal line representing the growth in the RFE score. So, looking at the policy types shows us the exact "tools" the state fights disinformation with. The proportion of educational policies is shown at the bottom of the stacked bar chart, from which we can see a steady decline as the risk to freedom of expression grows. Finland has media literacy as a core component of its national curriculum, while The Netherlands and Portugal utilize educational platforms against disinformation, which shows how these states prioritize strengthening public resilience to disinformation.

French legislation emphasizes its strategic focus on protecting democratic elections, so it stood out for the high proportion in the electoral category. German legislation favors curational responses against disinformation, emphasizing content moderation practices and framing platforms as frontline gatekeepers. Notably, Australia, the U.K., and the U.S. had more balanced and diverse distributions of policy categories, which, as we have concluded from the previous section, do not necessarily correspond to lower risk on the state level. Overall, these visuals contextualize the next section of

the analysis, where I have used linear regression to assess whether certain policy categories are statistically associated with the RFE scores.

Table 4.5: Comparative summary of Full and Stepwise linear regression models predicting RFE score.

Note: B = unstandardized regression coefficient; s.e. = standard deviation of a coefficient; p = p-value indicating statistical significance (p < .05 = *; p < .01 = **; p < .001 = ***); AIC = Akaike Information Criterion, a measure of model fit that penalizes complexity (lower = better).

Predictor	Full Model B	s.e.	p	Stepwise Model B	s.e.	p
(Intercept)	10.613	2.588	0.000	10.661	1.038	0.000
Monitoring	-2.277	2.202	0.307	-2.960	1.725	0.102
Investigative	-1.002	2.047	0.627	–	–	–
Counterdisinformation	0.002	2.122	0.998	–	–	–
Electoral	-1.313	2.754	0.636	–	–	–
Curational	3.095	2.420	0.208	2.736	1.738	0.08
Technical	-0.189	2.495	0.940	–	–	–
Economic	1.581	4.369	0.719	–	–	–
Ethical	1.490	2.086	0.479	–	–	–
Educational	**-6.817**	2.139	**0.002****	**-7.281**	1.816	**0.000*****
Empowerment	-1.908	2.213	0.377	–	–	–
Covid specifc	0.461	3.194	0.885	–	–	–
Model Fit	**Full Model**			**Stepwise Model**		
R^2	0.407			0.375		
Adjusted R^2	0.234			0.334		
Residual Std. Error	5.838	df=38		5.439	df=46	
AIC	340.5			327.1		

To identify the most statistically relevant policy categories associated with the RFE scores, I ran a series of multiple linear regression models and solidified my final model choice with AIC comparison and Mann–Whitney U tests. The Full Model involved all 11 of the policy types as independent binary variables and the RFE score as a continuous dependent variable. As seen from Table *4.5*, the results of the Full Model did not reveal categories that drive up the risk but identified educational policies to be significantly negatively associated with the risk to speech. However,

many policy types (e.g., economic, covid), likely due to their low frequency in the dataset exhibited previously by Figure 4.11, do not reach significance. This is a common limitation when working with sparse categorical predictors in policy analysis (Gelman & Hill, 2007).

To reduce the model, I employed stepwise regression, which sequentially removes the least significant variables (Diekhoff, 1992). This procedure retained only three predictors, only one of which was significant: educational, with $p<0.001$ ($p = 0.0002$), while monitoring ($p = 0.092$) and curational ($p = 0.122$) which were not significant (Table 4.5). The final model demonstrated improved interpretability while maintaining comparable explanatory power, making it more parsimonious in comparison to the original one. However, as with all automated selection techniques, stepwise regression has limitations, such as the risk of capitalizing on sample-specific noise. Therefore, its results were interpreted in conjunction with theoretical expectations and robustness checks.

The reduced model, which retained only educational and empowerment variables, yielded a substantially lower AIC (327.1) compared to the Full Model (340.5), indicating improved fit relative to model complexity (Burnham & Anderson, 2002). To further validate the findings from the linear regression, I also conducted a Mann–Whitney U test (Wilcoxon rank-sum) for each policy category. This test confirmed that policies with educational ($p < 0.001$) and curational ($p < 0.05$) focus had significantly different HRS scores compared to non-educational and non-curational policies (Table 4.5).

Concluding this section, the results revealed a weak negative relationship between the diversity of policies and policy level risk, meaning the diversity is associated with a slight reduction of the risk to freedom of expression. Analysis of policy types confirmed that educational policies significantly lower the risk to freedom of speech. These findings are theoretically and empirically sound, as countries with the lowest RFE scores employed multifaceted responses aimed at improving access to reliable information and building societal resistance to disinformation with media literacy programs.

5 Discussion

The following chapter interprets the main findings of this thesis and contextualizes them in light of prior research on disinformation. It focuses on unpacking the findings achieved with the use of the Risk to Freedom of Expression (RFE) score, examining why particular countries scored higher or lower, and considering broader implications for its applicability to the discussions of the state's role in online disinformation governance. The chapter also connects these insights to existing debates about regulatory models and media systems before considering the value and limits of the analytical tools employed.

5.1 Applicability to prior research

This thesis evaluated the content of disinformation policies adopted between 2010 and 2024 in 12 countries. Specifically, it pursued two aims: first, to design a tool to systematically assess the risk to freedom of expression within state responses to disinformation; second, to empirically test this tool's explanatory potential. To pursue those goals, the study designed the RFE composite score measuring the risk to freedom of expression based on the international human rights standards, analyzed 50 national disinformation policies across 12 countries, and used statistical techniques to identify the influence of media systems and individual policy types on the composite score.

Designing the composite score index enabled this thesis to transform international human rights standards into measurable indicators and capture the latent risk of restricting freedom of speech embedded within the content of disinformation policies. This theoretical grounding enabled a robust metric to systematically assess and compare disinformation policy formulations across democratic states. By averaging the scores of individual policies to generate a national-level indicator, this thesis took a comprehensive approach that avoided reducing governance to a single model, ideology, or system. Many previous studies assessing governance models have acknowledged that regulation often exists on a spectrum of possibilities (Stasi & Parcu, 2021). So, this thesis provided a panoptic view of each country's actual policy outputs, allowing for testing the resilience of media systems as ideal types. As it turned out, these typologies could not meaningfully explain cross-national variation when confronted with policy-level data.

This thesis engaged with media systems theory as a heuristic device, not a deterministic framework. Hallin and Mancini's typology (2004), as discussed in section 2.1.2, was initially developed on traditional legacy media and was subjected to scrutiny and reformulations to adapt it to digitalization. However, this study's analysis showed divergence within and across typologies. For instance, while the Democratic Corporatist model (Finland, the Netherlands, Germany) was grouped due to shared "consensus political systems, strong welfare states, and pronounced democratic corporatism" (Hallin & Mancini, 2004) and "high resilience to online disinformation" (Humprecht et al., 2020), their policy formulations in respect to freedom of expression diverged significantly. Germany leaned more toward delegated content control, and Finland favored preemptive educational measures. Similarly, Portugal did not conform to the high-risk profile often associated with the Polarized Pluralist model, which is in line with previous studies claiming Portugal's deviation from this model (Brüggemann et al., 2014; Norris, 2009; Hallin & Mancini, 2013).

Ultimately, this study found that employing policy characteristics reveals new patterns across state disinformation regulation. The empirically derived clusters revealed novel country groupings that transcended classic typology boundaries. This aligns with critiques from Brüggemann et al. (2014), Voltmer (2011), and Hardy (2021), who called attention to the hybridization and fluidity of contemporary media systems, especially under globalizing and digitalizing pressures. While Hardy's (2021) "role of the state" dimension was useful as a conceptual entry point, this thesis supports recent efforts by scholars like Humprecht et al. (2022) and Maniou (2023) to rethink media systems through the lens of information governance, digital media transformation, and press freedom challenges. While the factors influencing state behavior in online disinformation governance should be explored further in future comparative research, the concept of "centralization of regulatory authority" might be useful as one of the indicators, as it reveals the regulation is dependent on the relationship with another stakeholder of media governance, digital platforms.

This is not to say the resulted groupings are definitive or conducive to the state information governance systems, nor that they should be. On the contrary, this thesis supports recent media system scholarship in divergence from prototyping media systems into traditionally defined boxes. More importantly, the operationalization of

dimensions comprising the composite score shows how policy formulation can be used to identify telltale signs of socio-cultural and political motives that impact state behavior.

Analysis of state policies and their cluster groupings revealed patterns in how state view their role in online disinformation governance, which in turn shows how well the states have adapted to globalized and digitalized forms of communication. Media and communication governance are weaved from interdependencies of a panoply of policy actors and processes. Joseph Krasner (1991) attributed the main feature of global communication to be 'multiplicity' due to various forms of media outlets being governed by ever-evolving media cultures and norms, decision-making principles, and procedures. The multiplicity aspect of global media communication governance still holds true even three decades later, making it impossible to centralize and challenging to systematically govern.

Moreover, there have been new power shifts with the emergence of technologies, which further complicate the ‘multiplicative’ and ‘ungovernable’ nature of global communication. Social media platforms gained new editorial and curational powers “through the use of algorithmic systems, in the dissemination of content produced by the media and by others, and thus have a huge impact on the way people perceive the world and are exposed to [new] information” (Council of Europe, 2022; Phiri, 2023 p. 196). Although national governments still occupy a major stakeholder role in media governance (Raboy & Padovani, 2010), power shifts in media governance prompt state actors to produce new policy approaches, especially regarding disinformation shared online.

To assess the risk of these policies to freedom of expression, three indicators were designed under the premise that no political actors nor social media actors should define what is a lie. The state has a duty to provide instruments that facilitate individuals to discover the truth for themselves. This duty includes dealing with digital platforms who govern digital spaces where citizens practice self-expression. Thus, the issue stems from the conundrum of accountability between states and platforms under new models of digital communication governance: if not well balanced, regulation steers either into direct “administrative” censorship by state, or

“delegated censorship”, under which social media platforms are given reigns to govern truth (and lies) (Phiri, 2023). The former is evident by the sole occupant of Cluster 4, Hungary, positioned on the very top end of the risk score range. Hungary displayed clear tendencies of administrative censorship, where governmental bodies assume the authority to dictate lawful expression. With no clarity on what constitutes “illegal” expression, citizens are forced into consensus out of fear of criminal liability.

The second type of censorship is gaining more prominence among the democratic states that seek ways to reduce state control over information yet inadvertently delegate this onus onto digital platforms. For example, Germany and the U.S. embrace democratic principles and refrain from unnecessary direct state control, instead delegating the task of regulating harmful content to digital platforms. The key insight here is the way the rules of governance are assigned may undermine free speech. For instance, while Germany undoubtedly has strong mechanisms ensuring the protection of speech, evidenced by its high ranking at 10th in the 2024 World Press Freedom Index by Reporters Without Borders (RSF, 2025), German’s history NetzDG that aimed to fight hate speech demonstrated signs of delegated censorship, where platforms are coerced into overblocking content under threat of heavy fines. Since 2022, they have adapted the legislation to the EU-wide DSA regulations through the Digital Services Act. Yet, the NetzDG measure had global consequences and influenced 25 legislations around the world, including flawed democracies and authoritarian regimes, who used nebulous terms like “anti-government propaganda”, “defamation of religions”, or “fake news” to curb dissent (Mchangama & Alkiviadou, 2020).

This is not to say Germany is liable for the draconian laws imposed in already compromised autocratic structures, nor to argue for decreased regulation of disinformation, hate speech, and extremism. This thesis argues that by delegating its duties, the state reaps what it sows. Social media censors have at times mislabeled parody or satire as extremist content and deleted it to avoid fines, illustrating how over-blocking censors legitimate critique and stifles democratic discourse (Hallam, 2018). Consequently, public dismay surrounding such censorship has been exploited by far-right groups like Alternative for Germany (AfD), who appeal to democracy

while framing themselves as victims of government suppression to gain political support (Hallam, 2018). Rather than simply protecting society from hate, poorly calibrated enforcement end up chilling legitimate expression and feeding into the very extremist narratives it seeks to curb.

The U.S. approach to disinformation regulation is shaped by the First Amendment jurisprudence's counterspeech doctrine, under which 'negative' false speech is best addressed through *'positive' true speech*, not state intervention (Estrella, 2023). Hence, the U.S.'s freedom of speech is a *negative* right (i.e., freedom from government interference), which contrasts with a *positive* approach adopted in the EU's Digital Services Act (i.e., state plays a role in safeguarding individuals) (Huang, 2022). Such a laissez-faire approach to disinformation leaves the task of controlling content to digital platforms, failing to acknowledge their asymmetrical power in relation to users (Ibid).

Meanwhile, state attempts to address disinformation have been heavily criticized and repealed. In 2016, The U.S. State Department established The Global Engagement Center to address foreign disinformation campaigns. Last year, it was dissolved amid controversy and accusations of enforcing censorship (Gedeon, 2025). Some initiatives have been repealed even before they became fully operational: The Disinformation Governance Board (DGB), created under the Department of Homeland Security in 2022, was dismantled within just three weeks of its launch (Gedeon, 2025). A few hours after the announcement, far-right influencer Jack Posobiec responded to it on X, blaming the Biden administration for creating a "Ministry of Truth" aimed at policing speech (Lorenz, 2022). In reality, the DGB did not possess any regulatory powers to monitor Americans; rather, it was designed to merely collect best practices and issue recommendations pertaining to the harms of disinformation.

While the U.S. rejects state regulation of speech under the premise of protecting democracy, it simultaneously tolerates or even promotes selective suppression of dissent, particularly when it challenges dominant ideologies or geopolitical narratives. Since the beginning of 2025, federal employees working at the Election Security and Resilience division to counter election-related disinformation have been purged under claims of censorship (Sakellariadis & Miller, 2025), while students

expressing support for Palestine face visa revocations and deportation threats due to "promoting antisemitic views online" (Drenon, 2025), revealing a selective enforcement of speech rights that punishes critique while shielding dominant narratives. In this sense, we can see how "public interests", "defamation of public beliefs", and "safety", as interpreted from ICCPR's Article 19 as legitimate reasons to suppress speech, are readily weaponized since they protect the interests of those in power.

This is precisely what Gunatilleke (2021) cautioned about. He claimed the vulnerability of ICCPR's Article 19(3) proportionality test is that "majoritarian interests can infiltrate limitation grounds such as national security, public order, public health, and public morals" and advance their personal agenda (Gunatilleke, 2021). Political philosophy scholars contend that the states have no rights comparable to those of a citizen, for they have the duty to secure and respect those rights (Phiri, 2023, p. 96). For this reason, the interests of those in power should not dictate the interests of the public. As this thesis has established, some states find ways to safeguard the public against disinformation while upholding democratic principles and ensuring the equal exercise of rights. Weaponizing "free speech" to silence opposition while surreptitiously repealing any regulation of harmful content is decidedly not one of those ways.

For this reason, states that embraced their role as co-regulators of disinformation had the lowest RFE scores in this sample. From the example of the Netherlands, we can see how their responses to disinformation emphasize multi-stakeholder collaboration between states, platforms, and, most crucially, citizens. The 2024 version of Government-Wide Strategies to Combat Disinformation declares the following: "Addressing disinformation in terms of content is primarily not a task for governments, but for journalism and science, whether or not in collaboration with internet services". This approach de-centralizes the power of the government while holding them accountable for the coordination and oversight of these measures. By empowering citizens as equal stakeholders in disinformation governance, the Dutch approach aligns with the recommendations of the European Commission's High-Level Expert Group (HLEG) on Fake News and Online Disinformation, which called for a multi-dimensional approach and fostering civic engagement. HLEG report

emphasizes that strengthening the news ecosystem rests on the pillars of transparency, media literacy, and user empowerment, the combination of which is essential to safeguarding freedom of expression in the digital era (European Commission, 2018).

This thesis also explored what differentiated policies with lower risk to freedom of expression scores by inspecting the tools these policies used. State responses to disinformation rarely involve a singular legal instrument but rather a combination of strategies that vary widely in scope, intent, and potential to restrict expression. Firstly, this thesis has found that well-rounded or multi-dimensional policy responses (i.e., policies employing a variety of tools used to combat disinformation) were associated with a lower risk to speech. While the results showed only a modest negative correlation ($r = -0.3$), this finding foregrounds promising areas for research, especially for analyzing which combinations of strategies have the most positive effect on speech. This finding also supports evidence-based recommendations from the supranational organizations and international human rights bodies that advocate for multi-pronged approaches to protect freedom of speech. However, on a state level, diversity of policy types was not associated with a lower average RFE score, which speaks to the importance of harmonization of approaches.

In the United Kingdom, policymakers employed a variation of strategies. Early initiatives such as the 2016 Foreign and Commonwealth Office's Counter Disinformation Programme and the 2019 House of Commons inquiry into 'Disinformation and Fake News' recommended a vast array of strategies, which eventually led to the legislation of the Online Safety Act that was passed in 2023. The U.K. was one of the few countries in the sample that covered all 11 policy categories, proving its "textbook" multifaceted approach. Although some of these recent legislative responses have drawn criticism for their potential regulatory overreach, the overall strategy reflects a deliberate attempt to balance monitoring, public education, and legislative control (Sabbagh, 2023). However, countries with significantly narrower policy coverage, like Finland, Portugal, and France, also scored low levels of risk. This suggests that while diversity can help, it is ultimately the legal clarity and structure of a policy — not just how many types of responses it covers — that makes the difference. In short, quality overrules quantity. A country needs not to

issue a dozen disinformation laws. It needs to craft the few it does have in ways that are clear, proportionate, and protective of rights.

Secondly, when looking at specific policy types, educational and media literacy policies had the most significant effect in lowering the risk to freedom of expression. This finding aligns with prior scholarship (e.g., Napoli, 2018; Marecos et al., 2023) emphasizing counterspeech over censorship, particularly in algorithmically mediated environments where filter bubbles distort truth-seeking. One stark example from this study is Finland, which is internationally celebrated as ‘winning the war on fake news” (Mackintosh, 2019). Despite recommendations to issue disinformation-related legislation (Moilanen et al., 2023), Finnish policymakers focused on media literacy and educational efforts that build informational resilience and, as this thesis has shown, positively safeguard freedom of expression. However, this is not to claim that certain policy types are guaranteed to protect freedom of speech, nor that there is a single policy that could ‘solve’ disinformation. As Jackson and Bateman (2024) aptly put it, there is no silver bullet or "best" option — most interventions' effectiveness is highly context-dependent. The success of fact-checking labels, counter-messaging, or content moderation often hinges on factors like formulation, platform design, and credibility of sources.

By scrutinizing the content of government responses to disinformation, this thesis developed a model for a composite score measuring the risk to freedom of expression. Democratic countries with different media systems succumbed to the same models of censorship due to their inability to account for platform power that further destabilizes information ecosystem. All in all, the fact that this thesis confirmed many of the recommended approaches to tackle disinformation do indeed safeguard speech attests to their usefulness and applicability for informing future public policymaking. The theoretically grounded and empirically tested concepts may thus be useful for future attempts to build measurable indicators capturing the risk to freedom of expression.

5.2 Limitations

This thesis has offered a novel framework for assessing the risk to freedom of expression embedded in disinformation governance. Yet, it also faces several limitations. Firstly, the generalizability of the findings was constrained by the scope of the collected data. The dataset comprised only 50 policies and was limited to only 12 countries within the "Global North", which is heavily over-represented in mis-/disinformation literature as it is (Sanfilippo et al., 2024). While this argument is sound, disinformation regulation is constantly evolving in countries of Europe and North America, which justified a content analysis of policy formulations to seek out the risks to freedom of speech (Nieminen, 2024; Puppis, 2010; Raboy & Padovani, 2010). This limitation notwithstanding, I encourage future researchers to build on the existing dataset with new countries and government responses.

While this thesis did not aim to assess the entirety of each of the state's disinformation governance landscapes, numerous other disinformation-related laws and speech protections may exist beyond the scope of this analysis, especially on local and municipal levels. This is not to mention earlier legislation that impacts current media governance. For instance, the U.S. Communications Decency Act (1996) is foundational for the state's technocratic stance on social media platforms immunity from liability from legal liability for unlawful content posted by users. In Europe, on the other hand, historical protections for speech have existed dating back to the 19th century and were not accounted for, though they have shaped the regulatory landscape today. In France and Germany, media laws granted a right of reply at least since the 1800s (Danziger, 1986).

Another limitation pertained to the data analysis process. The thesis involved a single coder for the entire process of content analysis. While this was understandable due to the scope of the master thesis, steps were taken to mitigate this limitation. Coding rubrics were developed systematically and grounded in international human rights law, which fostered their validity (Creswell, 2012). For intra-coder reliability, I first coded the data in December 2024, then did three iterations of coding with a 2-week break each. This method, as Mackey and Gass (2005) explained, allows for comparative re-coding of parts of data. Additionally, my coding process was validated

externally by my supervisor, which partially ensured an “external audit” of the inter-coder reliability check.

Ultimately, the construction of this study’s composite score is just one way to do it, while there are admittedly more possibilities to operationalize the dimensions or assign score values. For instance, each dimension can be further broken down so that it comprises its own composite score, while RFE becomes a composite index score. Furthermore, regulation can be operationalized based on other frameworks, for example, six dimensions proposed by Hardy (2021) or Ginosar’s (2013) governance models developed as an alternative to media systems. Future researchers should pay attention to those if they want to develop generalizable and empirically grounded index scores.

Translation of individual policies into quantifiable risk scores involves a degree of subjectivity. While the indicators were grounded in international human rights law, the scoring inevitably simplified legal and contextual complexities that may vary across national jurisdictions, as well as across the same policy types. The language barrier presented another challenge, particularly for legal documents in non-English-speaking countries of my sample. This limitation was addressed by triangulating my data collection process, namely by cross-referencing secondary legal analyses and expert commentary from native-speaking scholars, especially in identifying definitional ambiguities.

The scoring of individual policies could benefit from more in-depth policy research, which may be a fruitful area of exploration for future public policy scholars. Universal indices related to human rights are bound to face challenges because there is no universal understanding of freedom of expression, as this study has demonstrated with the comparative analysis between European and U.S. views on freedoms. Nevertheless, the concepts developed through this endeavor — *clarity of definition, centralization of regulatory authority*, and *strictness of regulation* — rely on the disinformation literature that scrutinized the conundrum between regulation and speech and thus can be applied as a heuristic tool for in-depth qualitative analyses on national levels.

6 Conclusion

Through a comparative and legally grounded framework, this research examined how different disinformation responses balance regulation and protection of freedom of expression. The study's key contribution is the development of a Risk to Freedom of Expression (RFE), a composite score measure based on three theoretically grounded indicators: clarity of legal definition, centralization of regulatory power, and strictness of regulation. By assessing 50 disinformation policies across 12 Western countries from 2010 to 2024, this framework quantified latent risks to free speech embedded in legal and policy responses to disinformation. While democracies of this sample commonly share commitments to free expression, the RFE scores of their disinformation responses revealed a different picture.

The first reason for these variations is that countries view their duty to protect their citizens from disinformation differently. The second reason stems from whether the regulatory approaches sufficiently estimate the power of online platforms in current media governance. The Netherlands, for example, acknowledges its duty to protect and provide its citizens with tools to tackle disinformation through non-punitive and co-regulatory approaches with platforms. The U.S., on the other hand, struggles to fulfill the duty to address disinformation due to the deep-rooted doctrine of counter-speech that is used to stifle any regulatory safeguards. By developing the dimensions of risk, this study has shown how to recognize ambiguous policy formulation and improper delegation of regulatory duties. The existence of policies in this sample receiving a score of zero (i.e., no risk to freedom of expression) may be used as exemplary case studies for state action against disinformation. Concern for freedom of speech and actions against disinformation should go hand in hand, both for the sake of the effectiveness of these measures against the issue at hand and for the sake of democracy. Freedom of expression should, in itself, be a tool to combat disinformation (Bontcheva et al., 2020).

There is a strong need to view the issue of disinformation holistically and aim for a general improvement of pluralism of information and opinions. One of the most significant findings of this thesis was identifying an inverse relationship between educational and multidimensional policies and reduced risk to freedom of speech.

The empirical findings generated with this data essentially confirmed the recommendations made by a rich body of disinformation governance literature, which prove the usability of the developed conceptualized dimensions for foregoing qualitative in-depth research case studies. Furthermore, this thesis supports the evolution of ideal media system typologies in the digital age. As governance becomes more hybrid and transnational, policy responses to disinformation become molded not only by historical media-state relationships but also by digital platforms and the transnational information ecosystem overall. The need to inspect these policies from a human rights angle will persist, just like the issue of disinformation itself. As Hardy (2021) recommended, the proliferation of mis-/disinformation governance as a phenomenon can be used to construct new media system typologies that would reflect the changes in state behavior in relation to politics and media.

The inherent value of this thesis is the development of a conceptually grounded tool for foregoing research and robust data on the policy content of government responses of 12 democratic states. The approach taken here furnishes mixed-method research with a statistical basis to guide in-depth case studies. This study, therefore, establishes the foundation for scholars who wish to test the applicability of this tool to seek out factors that exacerbate said risk. Alternatively, the same dataset can be used as a base for a more comprehensive composite score. I would recommend performing a sensitivity analysis by altering the conditions of the composite score, for example, by changing the scoring of dimensions to examine how that affects the model. Finally, the individual dimensions can be applied as concepts guiding qualitative analysis of legislation, supplemented by relevant legal and national level expertise, to call out disproportionate and chilling legislation.

Reaching our terminus, this thesis concludes by recommending policymakers resist the temptation to abdicate their responsibilities by over-relying on private platforms. As the UN Special Rapporteur David Kaye (2018) and the ECtHR (2022) have both emphasized, states cannot absolve themselves of human rights obligations by delegating regulatory power to corporate actors. Doing so risks empowering corporate logic over democratic values and users over systems designed to protect them.

7 References

Act to Improve Enforcement of the Law in Social Networks. (2017). *NetzDG*. Federal Ministry of Justice and Consumer Protection. https://www.bmj.de/SharedDocs/Downloads/DE/Gesetzgebung/RefE/NetzDG_engl.pdf?__blob=publicationFile&

Act No. 69/2018 Coll. on Cybersecurity and on Amendments and Supplements to certain Acts. (2018, January 30). https://www.nbu.gov.sk/act-no-692018-coll-on-cybersecurity/

Act No. 264/2022 Coll. on Media Services and on Amendments and Supplements to Certain Acts. (2022, June 22). https://www.culture.gov.sk/wp-content/uploads/2019/12/Act-No.-264_2022-Coll.-on-media-services-and-amending-certain-acts-Media-Services-Act-1.pdf

Adcock, Robert & Collier, David. (2001). Measurement Validity: A Shared Standard For Qualitative and Quantitative Research. *American Political Science Review*. 95. 10.1017/S0003055401003100.

Alex Rochefort (2020) Regulating Social Media Platforms: A Comparative Policy Analysis, Communication Law and Policy, 25:2, 225-260, DOI:10.1080/10811680.2020.1735194

Allegri, Maria. (2024). The Impact of Disinformation on the Functioning of the Rule of Law and Democratic Processes in the Eu. *Interdisciplinary Journal of Research and Development*. 11. 98. 10.56345/ijrdv11n1s116.

Amnesty International. (2022). *A human rights approach to tackle disinformation: submission to the Office of the High Commissioner for Human Rights*. (Index Number: IOR 40/5486/2022) 14 April 2022. https://www.amnesty.org/en/documents/ior40/5486/2022/en/

Amendments to the Criminal Law: Article 90 on Influencing the Electoral Process Using Deep Fraud Technologies, 2024. gada 9. maija likums "Grozījumi Krimināllikumā", Latvijas Vēstnesis, 97, 21.05.2024. https://likumi.lv/ta/id/352098

Appelman, N., Dreyer, S., Bidare, P. M., & Potthast, K. C. (2022, May 16). Truth, intention and harm: Conceptual challenges for disinformation-targeted governance. *Internet Policy Review*. https://policyreview.info/articles/news/truth-intention-and-harm-conceptual-challenges-disinformation-targeted-governance/1668

Arayankalam, Jithesh, Prakriti Soral, Anupriya Khan, Satish Krishnan, Indranil Bose. (2024). Does centralization of online content regulation affect political hate speech in a country? A public choice perspective, *Information & Management*, Volume 61, Issue 2, 2024,103919, ISSN 0378-7206, https://doi.org/10.1016/j.im.2024.103919 (https://www.sciencedirect.com/science/article/pii/S0378720624000016)

Arguelles, Cleve V. and Jose Mari H. Lanuza. (2021). Media System Approach to Disinformation Vulnerability: Developing Disinformation Resilience in Southeast Asia. In *The Next Digital Decade: Case Studies from Asia,* Volume 1 Traces and Divides. Digital Asia Hub and Konrad-Adenauer-Stiftung, Regional Programme Political Dialogue Asia.

Banchio, Pablo Rafael. (June 24, 2024). *Legal Framework to Combat Disinformation and Hate Speech on Digital Platforms* Available at SSRN: https://ssrn.com/abstract=4879162 or http://dx.doi.org/10.2139/ssrn.4879162

Baptista, C., & Morgado, M. M. (2022, September 2). *Digital era and protection against disinformation*. Caiado Guerreiro. https://www.caiadoguerreiro.com/en/digital-era-and-protection-against-disinformation/

Bateman, J., & Jackson, D. (2024, January 31). *Countering disinformation effectively: An evidence-based policy guide | Carnegie Endowment for International peace*. Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2024/01/countering-disinformation-effectively-an-evidence-based-policy-guide?lang=en

Baldwin, Robert, Cave, Martin and Lodge, Martin (2011) *Understanding regulation: theory, strategy, and practice*. Business & management. (2nd). Oxford University Press, Oxford, UK. ISBN 9780199576098

Balkin, Jack M. (2004). Digital Speech and Democratic Culture: a Theory of Freedom of Expression for the Information Society. *New York University Law Review*, Vol. 79, No. 1, 2004, Yale

Law School, Public Law Working Paper No. 63, Available at SSRN: https://ssrn.com/abstract=470842 or http://dx.doi.org/10.2139/ssrn.470842

Bayer, J., & Bárd, P. (2020*). Hate speech and hate crime in the EU and the evaluation of online content regulation approaches*. European Union Policy Department for Citizens' Rights and Constitutional Affairs.

Bayer, J. (2024). The EU policy on disinformation: aims and legal basis. *Journal of Media Law*, 16(1), 18–27. https://doi.org/10.1080/17577632.2024.2362478

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, 33(2), 122-139. https://doi.org/10.1177/0267323118760317 (Original work published 2018)

Berg, N., & Kim, J. Y. (2018). Free expression and defamation. Law, Probability and Risk, 17(3), 201–223.

Bevir, M. (2013). *A Theory of Governance*. 55-71. Retrieved from https://escholarship.org/uc/item/2qs2w3rb

Blauth, T. F., Gstrein, O. J., & Zwitter, A. (2022). Artificial intelli- gence crime: An overview of malicious use and abuse of AI. IEEE Access, 10, 77110–77122.

Bleyer-Simon, Konrad, Reviglio Della Venaria, Urbano. (2024). Defining disinformation across EU and VLOP policies. *EUI, STG, European Digital Media Observatory (EDMO), Policy Report, 2024* - https://hdl.handle.net/1814/77435

Bontcheva, K., Posetti, J., Teyssou, D., Meyer, T., Gregory, S., Hanot, C., & Maynard, D. (2020). Balancing Act: Countering Digital Disinformation while respecting Freedom of Expression. *UNESCO*.

Borsboom, D., Mellenbergh, G.J., Heerden J.V. (2003). The theoretical status of latent variables. *Psychological Review*. 2003; 110(2): 203-219

Boshnakova, D., Dankova, D. (2023). The Media in Eastern Europe. In: Papathanassopoulos, S., Miconi, A. (eds) The Media Systems in Europe. Springer Studies in Media and Political Communication. Springer, Cham. https://doi.org/10.1007/978-3-031-32216-7_7

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, 9(2), 27-40.

Bowman, A. O. M., & Krause, G. A. (2003). Power Shift: Measuring Policy Centralization in U.S. Intergovernmental Relations, 1947-1998. American Politics Research, 31(3), 301-325. https://doi.org/10.1177/1532673X03251381 (Original work published 2003)

Boyd, D., & Crawford, K. (2012). Critical questions for Big Data. Information, Communication & Society, 15(5), 662–679. doi:10.1080/1369118X.2012.678878

Braman, S. (2004). Where Has Media Policy Gone? Defining The Field In The Twenty-First Century. *Communication Law and Policy*, 9(2), 153–182. https://doi.org/10.1207/s15326926clp0902_1

Brüggemann, M., Engesser, S., Büchel, F., Humprecht, E. and Castro, L. (2014), Hallin and Mancini Revisited: Four Empirical Types of Western Media Systems. *Journal of Communication*, 64: 1037-1065. https://doi.org/10.1111/jcom.12127

Burnham KP, Anderson DR. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*, 2nd edn. Springer, New York

Bychawska-Siniarska, D. (2017). *Protecting the right to freedom of expression under the European Convention on Human Rights - A Handbook for Legal Practitioners. Council of Europe Publishing*. https://edoc.coe.int/en/fundamental-freedoms/7425-protecting-the-right-to-freedom-of-expression-under-the-european-convention-on-human-rights-a-handbook-for-legal-practitioners.html

Caled D, Silva M.J. (2022). Digital media and misinformation: An outlook on multidisciplinary strategies against manipulation. *Journal of Computer Social Science*. 2022;5(1):123-159. doi: 10.1007/s42001-021-00118-8. Epub 2021 May 27. PMID: 34075349; PMCID: PMC8156576.

Callamard, Agnes. (2008). Expert Meeting of the Links Between Articles 19 and 20 of the ICCPR: Freedom of Expression and Advocacy of Religious Hatred that Constitutes Incitement to Discrimination, Hostility or Violence. *UN HCHR*, October 2-3, 2008, Geneva. (Article 19). URL: http://www.article19.org/data/files/pdfs/conferences/iccpr-links-between-articles-19-and-20.pdf.

Calo, Antonella & Longo, Antonella & Zappatore, Marco. (2023). Comparative Analysis of Disinformation Regulations: A Preliminary Analysis. 162-171. 10.1007/978-3-031-47112-4_15.
Cardno, C. (2018). Policy document analysis: A practical educational leadership tool and a qualitative research method. *Kuram ve Uygulamada Eğitim Yönetimi*, 24(4), 623-640. doi: 10.14527/kuey.2018.016
Casero-Ripollés, A., Tuñón, J. & Bouza-García, L. (2023). The European approach to online disinformation: geopolitical and regulatory dissonance. *Humanities Soc Sciences Communication* 10, 657 (2023). https://doi.org/10.1057/s41599-023-02179-8
Cavaliere, P. (2022). The truth in fake news: How disinformation laws are reframing the concepts of truth and accuracy on digital platforms. *European Convention on Human Rights Law Review*, 3(4), 481-523. https://doi.org/10.1163/26663236-bja10044
Chadwick, A. (2013). *The Hybrid Media System: Politics and Power*, 1st edn, Oxford Studies in Digital Politics (2013; online edn, Oxford Academic, 26 Sept. 2013), https://doi.org/10.1093/acprof:oso/9780199759477.001.0001, accessed 13 March 2025.
Chadwick A., Stanyer J. (2022). Deception as a bridging concept in the study of disinformation, misinformation, and misperceptions: Toward a holistic framework. *Communication Theory*, 32(1), 1–24. https://doi.org/10.1093/ct/qtab019
Charrad, M., Ghazzali, . N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, *61*(6), 1–36. https://doi.org/10.18637/jss.v061.i06
Chokshi, Niraj. (2017). That Wasn't Mark Twain: How a Misquotation Is Born. *New York Times*, 27 Apr. 2017, p. NA(L). *Gale Academic OneFile*, link.gale.com/apps/doc/A490676996/AONE?u=anon~79c59d14&sid=sitemap&xid=b2d57975. Accessed 30 Mar. 2025.
Cipers, S., Meyer, T., & Lefevere, J. (2023a). Government responses to online disinformation unpacked. *Internet Policy Review*, 12(4). https://doi.org/10.14763/2023.4.1736
Cipers, Samuel; Meyer, Trisha; Lefevere, Jonas. (2023b). Government Responses to online disinformation unpacked [Dataset], https://doi.org/10.7910/DVN/ZGIKLS, *Harvard Dataverse*, V2
Colomina C., Héctor Sánchez Margalef, Richard Youngs (2021). The impact of disinformation on democratic processes and human rights in the world. European Parliament, *Policy Department for External Relations*, April 2021. Retrieved from: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO_STU(2021)653635_EN.pdf
Cosponsors - S.314 - 104th Congress (1995-1996): Communications Decency Act of 1995. (1995, February 1). https://www.congress.gov/bill/104th-congress/senate-bill/314/cosponsors
Couldry, Nick and Hepp, Andreas (2012*) Media cultures in a global age: a transcultural approach to an expanded spectrum*. In: Volkmer, Ingrid, (ed.) The Handbook of Global Media Research. Wiley-Blackwell, Chichester, pp. 92-109. ISBN 9781405198707
Council of Europe, Committee of Ministers. (2022). Recommendation CM/Rec(2022)11 of the Committee of Ministers to member States on principles for media and communication governance. https://search.coe.int/cm/#{%22CoEIdentifier%22:[%220900001680a61712%22],%22sort%22:[%22CoEValidationDate%20Descending%22]}
Cox, M. (2020). States of Emergency and Human Rights During a Pandemic: A Hungarian Case Study, *Human Rights Brief* , 24(1), pp. 32-41.
Crepaz, M., & Chari, R. (2018). Assessing the validity and reliability of measurements when evaluating public policy. Journal of Public Policy, 38(3), 275–304. doi:10.1017/S0143814X16000271
Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4 ed.). Boston, MA: Pearson Education, Limited.)
Dan V., Paris B., Donovan J., Hameleers M., Roozenbeek J., van der Linden S., von Sikorski C. (2021). Visual mis- and disinformation, social media, and democracy. Journalism & Mass Communication Quarterly, 98(3), 641–664. https://doi.org/10.1177/10776990211035395
Danziger, Charles. (1986). The Right of Reply in the United States and Europe. 19 *New*

York University Journal of International Law and Politics. 171.
De Blasio, E., & Selva, D. (2021). Who Is Responsible for Disinformation? European Approaches to Social Platforms' Accountability in the Post-Truth Era. American Behavioral Scientist, 65(6), 825-846. https://doi.org/10.1177/0002764221989784 (Original work published 2021)
de la Mata, D., Guede, J., & Sebastián, M. (2024). Hallin and Mancini: Two Decades of Influence in Politics and Communications. *Media and Communication, 12*, Article 7695. https://doi.org/10.17645/mac.7695
Diekhoff, G. (1992). *Statistics for the Social and Behavioral Sciences: Univariated, Bivariate, Multivariate*. Dubuque: William C Brown Publishers.
Dobek-Ostrowska, Boguslawa. (2015). 25 years after communism: four models of media and politics in Central and Eastern Europe. In *Democracy and media in Central and Eastern Europe 25 years on,* Bogusława Dobek-Ostrowska / Michał Głowacki. (eds.), pp. 11-45
Dockrell , V. (2024, May 22). Global crackdown on election deepfakes: Latvia proposes new criminal law, Australia calls for stronger regulations - national security news. *National Security News*. https://nationalsecuritynews.com/2024/05/global-crackdown-on-election-deepfakes-latvia-proposes-new-criminal-law-australia-calls-for-stronger-regulations/
Dov Bachmann, S.-D., Putter, D. & Duczynski, G. (2023) Hybrid warfare and disinformation: A Ukraine war perspective. *Global Policy*, 14, 858–869. Available from: https://doi.org/10.1111/1758-5899.13257
Dowling, M. E. (2022). Foreign interference and digital democracy: Is digital era governance putting Australia at risk? Australian Journal of Political Science, 57(2), 113–128.
Drozdova, K., & Gaubatz, K. T. (2017). *Quantifying the qualitative: Information theory for comparative case analysis*. Sage Publications.
Drenon, B. (2025, April 9). *Why has Trump revoked hundreds of international student visas?*. BBC News. https://www.bbc.com/news/articles/cg411rrnkkko
Electronic Mass Media Law, *Latvijas Vēstnesis* No. 82 (2010). https://wipolex-res.wipo.int/edocs/lexdocs/laws/en/lv/lv075en.html
Estrella, Eliza J. (2023). False Speech and the First Amendment: The Problem with Free Speech in a Fake News Crisis. *Brooklyn Law Review*. 1313 (2023). https://brooklynworks.brooklaw.edu/blr/vol88/iss4/6
European Commission: Directorate-General for Communications Networks, Content and Technology. (2018). A multi-dimensional approach to disinformation : report of the independent High level Group on fake news and online disinformation. Publications Office. https://data.europa.eu/doi/10.2759/739290.
Evangelista, R., & Bruno, F. (2019). WhatsApp and political insta- bility in Brazil: Targeted messages and political radicalisation. Internet Policy Review, 8(4), 1–23.
Farinho, Domingos. (2021). The Portuguese Charter of Human Rights in the Digital Age: A legal appraisal. *Revista Española de la Transparencia*. 2. 85-105. 10.51915/ret.191.
Fathaigh, R. Ó. (2024). [NL] new government measures to tackle disinformation . *IRIS Legal Observations of the European Audiovisual Observatory*. https://merlin.obs.coe.int/article/10117
Finck, Michèle. (2017). Digital Co-Regulation: Designing a Supranational Legal Framework for the Platform Economy (June 20, 2017). *European Law Review* (2018 Forthcoming), LSE Legal Studies Working Paper No. 15/2017, Available at SSRN: https://ssrn.com/abstract=2990043 or http://dx.doi.org/10.2139/ssrn.2990043
Flew, T., & Martin, F. (2022). *Regulating Platforms and Content Moderation in the Disinformation Age*. Journal of Digital Media & Policy.
Foreign interference: National Security Bill factsheet. GOV.UK. 13 July 2023. Retrieved 19 February 2025.
Fox, J., Welch, D. (2012). Justifying War: Propaganda, Politics and The Modern Age. In: Welch, D., Fox, J. (eds) *Justifying War*. Palgrave Macmillan, London. https://doi.org/10.1057/9780230393295_1
Freelon D., Wells C. (2020). Disinformation as political communication. *Political Communication,* 37(2), 145–156. https://doi.org/10.1080/10584609.2020
Funke, D., & Flamini, D. (2019). *A Guide to Anti-Misinformation Actions Around the World.* Poynter Institute Report.

Gedeon, J. (2025, April 16). *Trump administration shutters US Office countering foreign disinformation*. The Guardian. https://www.theguardian.com/us-news/2025/apr/16/trump-state-department-foreign-disinformation
Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.
Giansiracusa N (2021). *How Algorithms Create and Prevent Fake News*. Berkeley, CA: Apress.
Gillespie, A., Glăveanu, V., de Saint-Laurent, C., Zittoun, T., & Bernal Marcos, M. J. (2024). Multi-Resolution Design: Using Qualitative and Quantitative Analyses to Recursively Zoom in and out of the Same Dataset. Journal of Mixed Methods Research, 0(0). https://doi.org/10.1177/15586898241284696
Ginosar, A. (2013), Media Governance: A Conceptual Framework or Merely a Buzz Word?. *Communication Theory*, 23: 356-374. https://doi.org/10.1111/comt.12026
Global Partners Digital. (2023). A framework for analysing government responses to disinformation from a human rights perspective. GPD. Retrieved from: https://www.gp-digital.org/wp-content/uploads/2023/08/A-framework-for-analysing-disinformation-laws-and-policies-from-a-human-rights-perspective-1.pdf
Goldman, Eric. (2024). The United States' Approach to 'Platform' Regulation (January 01, 2024). *Santa Clara Univ. Legal Studies Research Paper* No. 4404374, Available at SSRN: https://ssrn.com/abstract=4404374 or http://dx.doi.org/10.2139/ssrn.4404374
Government of the Netherlands. (2022, December 23). *Government-wide strategy for effectively tackling disinformation*. https://www.government.nl/documents/parliamentary-documents/2022/12/23/government-wide-strategy-for-effectively-tackling-disinformation
Gradoń, K. T., Hołyst, J. A., Moy, W. R., Sienkiewicz, J., & Suchecki, K. (2021). Countering misinformation: A multidisciplinary approach. *Big Data & Society*, 8(1). https://doi.org/10.1177/20539517211013848 (Original work published 2021)
Greene, S., et al., 2021. Mapping fake news and disinformation in the Western Balkans and identifying ways to effectively counter them. The European Parliament's Committee on Foreign Affairs, EP/EXPO/AFET/FWC/2019-01/Lot1/R/01.
Greenhoot, A. F., & Dowsett, C. J. (2012). Secondary data analysis: An important tool for addressing developmental questions. *Journal of Cognition and Development*, 13(1), 2–18. https://doi.org/10.1080/15248372.2012.646613
Gunatilleke, G. (2021). Justifying Limitations on the Freedom of Expression. *Human Rights Review* 22, 91–108 (2021). https://doi.org/10.1007/s12142-020-00608-8
Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2018). *Multivariate Data Analysis* (8th ed.).United Kingdom: Cengage Learning.
Hair, J.F., Hult, G.T.M., Ringle, C.M., Sarstedt, M., Danks, N.P., Ray, S. (2021). *Evaluation of Formative Measurement Models. In: Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R*. Classroom Companion: Business. Springer, Cham. https://doi.org/10.1007/978-3-030-80519-7_5
Hallam, M. (2018, January 3). *AFD seeks to profit from "censorship" online – DW*. dw.com. https://www.dw.com/en/germanys-populist-afd-seeks-to-turn-online-censorship-to-its-advantage/a-42004730
Hallin, D.C. and Mancini, P. (2004). *Comparing Media Systems: Three Models of Media and Politics*. Cambridge University Press, Cambridge.
https://doi.org/10.1017/CBO9780511790867
Hallin, D. C., & Mancini, P. (2013). Comparing media systems: A response to critics. In F. Esser & T. Hanitzsch (Eds.), Handbook of comparative communication research (pp. 207–220). Routledge
Hallin, D. C., & Mancini, P. (2017). Ten Years After Comparing Media Systems: What Have We Learned? Political Communication, 34(2), 155–171. https://doi.org/10.1080/10584609.2016.1233158
Hameleers, Michael. (2023). Disinformation as a context-bound phenomenon: toward a conceptual clarification integrating actors, intentions and techniques of creation and dissemination. *Communication Theory,* Volume 33, Issue 1, February 2023, Pages 1–10, https://doi.org/10.1093/ct/qtac021

Hancock J. T., Bailenson J. N. (2021). The social impact of deepfakes. *Cyberpsychology, Behavior, and Social Networking*, 24(3), 149–152. http://doi.org/10.1089/cyber.2021.29208.jth
Hardy, J. (2012). Comparing Media Systems. In *Handbook of Comparative Communication Research*, edited by F. Esser, and T. Hanitzch, 185–206. London: Routledge.
Hardy, Jonathan. (2021). Media systems and misinformation. In*: Tumber H and Waisbord S (eds) The Routledge Companion to Media Disinformation and Populism. Routledge media and Cultural Studies Companions. Abingdon, Oxon* . New York: Routledge, 59–70
Hatano, A. (2023). Regulating Online Hate Speech through the Prism of Human Rights Law: The Potential of Localised Content Moderation. *The Australian Year Book of International Law Online*, 41(1), 127-156. https://doi.org/10.1163/26660229-04101017
Hauke, J. and Kossowski, T. (2011). Comparison of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. *Quaestiones Geographicae*, 30, 87-93.
Heaton, J. (2008). Secondary Analysis of Qualitative Data: An Overview. *Historical Social Research / Historische Sozialforschung*, 33(3 (125)), 33–45. http://www.jstor.org/stable/20762299
Helm, Rebecca K ,Hitoshi Nasu. (2021). Regulatory Responses to 'Fake News' and Freedom of Expression: Normative and Empirical Evaluation. *Human Rights Law Review*, Volume 21, Issue 2, June 2021, Pages 302–328, https://doi.org/10.1093/hrlr/ngaa060
Herzstein, R (1978). The most Infamous Propaganda Campaign in History, GP Putnam & Sons (NY) p.492
Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, 1, 104–121. http://dx.doi.org/10.1177/109442819800100106
Hinton, P. R. (2014). Statistics explained (3rd ed.). New York, NY: Routledge.
Hofmann, Jeanette, Christian Katzenbach & Kirsten Gollatz, Between Coordination and Regulation: Finding the Governance in Internet Governance, 19 *NEW MEDIA & SOC'Y* 1406 (2016). Retrieved from https://www.econstor.eu/bitstream/10419/171970/1/f-19856-full-text-Hofmann-et_al-Between%20coordination-v3.pdf
Hsu, A. (2016.) Environmental Performance Index, Technical Report. Yale University, New Haven, CT. https://www.bl.uk/world-war-one/articles/patriotism-and-nationalism
Huang, Tzu- Chiang. (2022). Private Censorship, Disinformation and the First Amendment: Rethinking Online Platforms Regulation in the Era of a Global Pandemic, 29 *Michigan Technology Law Review*. 137 (2022). Available at: https://repository.law.umich.edu/mtlr/vol29/iss1/5
Humprecht, E., Esser, F., & Van Aelst, P. (2020). Resilience to Online Disinformation: A Framework for Cross-National Comparative Research. *The International Journal of Press/Politics*, 25(3), 493-516. https://doi.org/10.1177/1940161219900126 (Original work published 2020)
Humprecht, Edda & Herrero, Laia & Blassnig, Sina & Brüggemann, Michael & Engesser, Sven. (2022). Correction to: Media Systems in the Digital Age: An Empirical Comparison of 30 Countries. *Journal of Communication*. 72. 10.1093/joc/jqac010.
Humprecht, E. (2023). The Role of Trust and Attitudes toward Democracy in the Dissemination of Disinformation—a Comparative Analysis of Six Democracies. *Digital Journalism*, 1–18. https://doi.org/10.1080/21670811.2023.2200196
H.R.4514 - 118th Congress (2023-2024): Disinformation Governance Board Prohibition Act. (2023, July 10). https://www.congress.gov/bill/118th-congress/house-bill/4514/all-info
International Covenant on Civil and Political Rights, opened for signature Dec. 16, 1966, arts. 19, 20, S. Exec. Doc. E, 95-2 (1978), 999 U.N.T.S. 171, 178 (entered into force Mar. 23, 1976)
Jahan, S. (2015). Human Development Report, Technical Report. United Nations Development Programme, New York.
Jacobs, Leslie Gielow. (2022). Freedom of Speech and Regulation of Fake News. *The American Journal of Comparative Law*, Volume 70, Issue Supplement_1, October 2022, Pages i278–i311, https://doi.org/10.1093/ajcl/avac010
Jakubowicz K. (2010). Media systems research: An overview. In: Dobek-Ostrowska B, Glawacki M,Jakubowicz K, et al (eds*) Comparative Media Systems: European and Global Perspectives*.Budapest and New York, NY: Central European University Press, pp.1–22
Janjić, Stefan & Jelena Kleut. (2022). Understanding Disinformation Through the Media Systems Perspective. In *Digitalne medijske tehnologije i društveno-obrazovne promene 10*. Retrieved

from: https://digitalna.ff.uns.ac.rs/sites/default/files/db/books/978-86-6065-752-9-.pdf#page=90
Joffe H., Yardley L. (2003). Content and thematic analysis. In Marks D. F., Yardley L. (Eds.), *Research methods for clinical and health psychology* (pp. 56–68). Sage.
Johnston, Melissa. (2014). Secondary Data Analysis: A Method of Which the Time has Come. *Qualitative and Quantitative Methods in Libraries*. 3. 619-626. https://www.qqml-journal.net/index.php/qqml/article/view/169/170
Kaminska, I. (2017). A lesson in fake news from the info-wars of ancient Rome. *Financial Times*.
Kari, M. J., & Hellgren, R. (2021). Case study: Finland as a Target of Russian Information Influence. In K. Ilmonen, & P. Moilanen (Eds.), *The Political Analyst's Field Guide to Finland* (pp. 155-172). Jyväskylän yliopisto. JYU Reports, 10. http://urn.fi/URN:ISBN:978-951-39-8931-6
Karppinen, Kari. (2013). Uses of Pluralism in Contemporary Media Policy. In *Rethinking Media Pluralism* (pp. 125–178). Fordham University Press. http://www.jstor.org/stable/j.ctt13wzz1r.12
Kaye, David. (2018). *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression* (A/HRC/38/35). United Nations. https://documents.un.org/doc/undoc/gen/g18/096/72/pdf/g1809672.pdf
Kaye, David. (2020). *Disease pandemics and the freedom of opinion and expression : report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*. United Nations, Geneva. https://digitallibrary.un.org/record/3862160/files/A_HRC_44_49-AR.pdf
Katsirea, I. (2018). "Fake news": reconsidering the value of untruthful expression in the face of regulatory uncertainty. *Journal of Media Law*, 10(2), 159–188. https://doi.org/10.1080/17577632.2019.1573569Khan, Irene. (2021). UN Special Rapporteur on Freedom of Opinion and Expression, *Report on Disinformation and Freedom of Expression*, United Nations, Geneva. Doc. A/HRC/47/25 (2021) https://digitallibrary.un.org/record/3925306/files/A_HRC_47_25-AR.pdf
Kelly, Michelle M., Tasha Martin-Peters, Jessica Strohm Farber. (2024). Secondary Data Analysis: Using existing data to answer new questions. *Journal of Pediatric Health Care*, Volume 38, Issue 4, 2024, pp. 615-618, ISSN 0891-5245, https://doi.org/10.1016/j.pedhc.2024.03.005.
Kooiman, J. (2003). *Governing as governance*. SAGE Publications Ltd, https://doi.org/10.4135/9781446215012
Kohlbacher, F. (2006). The use of qualitative content analysis in case study research. *Forum: Qualitative Social Research*, 7(1), 1-23.
Koltay, A. (2025). Freedom of Expression and the Regulation of Disinformation in the European Union. In R. J. Krotoszynski, Jr., A. Koltay, & C. Garden (Eds.), *Disinformation, Misinformation, and Democracy: Legal Approaches in Comparative Context* (pp. 133–160). chapter, Cambridge: Cambridge University Press.
Krasner, S. D. (1991). Global communications and national power: Life on the Pareto frontier. *World Politics*, 43(3), 336–366. https://www.jstor.org/stable/2010398
Krippendorff, K. (2004). Content Analysis: An introduction to its methodology (2nd ed.). Thousand Oaks, CA: Sage
Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, **47**, 583–621.
Kuiper, F. K. and Fisher, L. (1975). A Monte Carlo comparison of six clustering procedures. Biometrics, 31, 777–783.
Landman, Todd. (2003). The Scope of Human Rights: From Background Concepts to Indicators. *Studying Human Rights*. 1st ed. Routledge. https://doi.org/10.4324/9780203358504
Lanuza, J.M.H. and Arguelles, C.V. (2022). Media System Incentives for Disinformation. In *Disinformation in the Global South* (eds H. Wasserman and D. Madrid-Morales). https://doi.org/10.1002/9781119714491.ch9
Lauk, Epp. (2008). How will it all unfold? media systems and journalism cultures in post-communist countries. In *Finding the Right Place on the Map: Central and Eastern European Media Change in a Global Perspective*. Ed. by Karol Jakubowicz & Miklós Sükösd. Bristo, UK/Chicago, USA: Intellect, pp. 193-212

Law No. 2013-595 of July 8, 2013 on orientation and programming for the reestablishment of the school of the Republic [Loi n° 2013-595 du 8 juillet 2013 d'orientation et de programmation pour la refondation de l'école de la République] (France). (2013, July 8). https://www.legifrance.gouv.fr/eli/loi/2013/7/8/2013-595/jo/texte

Lehtonen, Markku. (2017). Operationalizing information: measures and indicators in policy formulation. In *Handbook of Policy Formulation || Operationalizing information: measures and indicators in policy formulation*. , 10.4337/9781784719326(), 161–181. doi:10.4337/9781784719326.00017

LEXOTA (Interactive Tool), available at www.lexota.org (accessed 20/01/2025).

Lidsky, Lyrissa Barnett, Where's the Harm?: Free Speech and the Regulation of Lies, 65 *Wash. & Lee Law Review*. 1091 (2008).

Lightfoot, Geoffrey & Tomasz Piotr Wisniewski. (2014). Information asymmetry and power in a surveillance society. *Information and Organization*, Volume 24, Issue 4, 2014, Pages 214-235, ISSN 1471-7727, https://doi.org/10.1016/j.infoandorg.2014.09.001.

Lim, G., & Bradshaw, S. (2023, July). *Chilling legislation: Tracking the impact of "fake news" laws on press freedom internationally*. Center for International Media Assistance. https://www.skeyesmedia.org/documents/bo_filemanager/CIMA-Chilling-Legislation_web_150ppi.pdf

Lorenz, Taylor (May 18, 2022). How the Biden administration let right-wing attacks derail its disinformation efforts. *The Washington Post*. https://www.washingtonpost.com/technology/2022/05/18/disinformation-board-dhs-nina-jankowicz/

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mackintosh, E. (2019). *Finland is winning the war on fake news. other nations want the blueprint*. CNN. https://edition.cnn.com/interactive/2019/05/europe/finland-fake-news-intl/

Maniou, T. A. (2023). The dynamics of influence on press freedom in different media systems: A comparative study. *Journalism Practice*. Informa UK Limited. https://doi.org/10.1080/17512786.2022.2030246

Mansell, R., Durach, F., Kettemann, M., Lenoir, T., Procter, R., Tripathi, G., and Tucker, E. (2025). *Information Ecosystem and Troubled Democracy: A Global Synthesis of the State of Knowledge on New Media, AI and Data Governance*. International Observatory on Information and Democracy. Paris.

Marecos, J., Shattock, E., Bartlett, O., Goiana-da-Silva, F., Maheswaran, H., Ashrafian, H., & Darzi, A. (2023). Health misinformation and freedom of expression: considerations for policymakers. *Health Economics, Policy and Law*, 18(2), 204–217. doi:10.1017/S1744133122000263

Maati, A., Edel, M., Saglam, K., Schlumberger, O., & Sirikupt, C. (2023). Information, doubt, and democracy: how digitization spurs democratic decay. *Democratization*, 31(5), 922–942. https://doi.org/10.1080/13510347.2023.2234831

Mattoni, Alice & Ceccobelli, Diego. (2018). Comparing hybrid media systems in the digital age: A theoretical framework for analysis. *European Journal of Communication*. 33. 026732311878483. 10.1177/0267323118784831.

McGowan, M. (2021, February 28). Australian "myth-busting" unit established to take on Covid misinformation. *The Guardian*. https://www.theguardian.com/world/2021/mar/01/australian-myth-busting-unit-established-to-take-on-covid-misinformation

Mchangama, J. and Alkiviadou, N. (2020). The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship - Act Two. Justitia. Retrieved May 10 2021, from https://justitia-int.org/wp-content/uploads/2020/09/Analyse_Cross-fertilizing-Online-Censorship-The-Global-Impact-of-Germanys-Network-Enforcement-Act-Part-two_Final-1.pdf

Miconi, A., Papathanassopoulos, S. (2023). On Western and Eastern Media Systems: Continuities and Discontinuities. In: Papathanassopoulos, S., Miconi, A. (eds) *The Media Systems in Europe*. Springer Studies in Media and Political Communication. Springer, Cham. https://doi.org/10.1007/978-3-031-32216-7_2

Minister van Binnenlandse Zaken en Koninkrijksrelatie, *Kabinet pakt desinformatie aan*, 23 december 2022 https://www.rijksoverheid.nl/regering/bewindspersonen/hanke-bruins-slot/nieuws/2022/12/23/kabinet-pakt-desinformatie-aan

Minister for Internal Affairs and Kingdom Relations, Cabinet tackles disinformation, 23 December 2022

Melns uz balta. (2023). *Par mums* [About us]. https://melnsuzbalta.lv/par-mums/

Moilanen P, Hautala M, Saari. D (2023). *The information landscape in Finland*. Available via EU Disinfo Lab. https://www.disinfo.eu/publications/disinformation-landscape-in-finland/

Monshizadeh, Mehrnoosh; Vikramajeet Khatri, Raimo Kantola, Zheng Yan. (2022). A deep density based and self-determining clustering approach to label unknown traffic, Journal of Network and Computer Applications, Volume 207, 2022, 103513, ISSN 1084-8045, https://doi.org/10.1016/j.jnca.2022.103513.

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *Journal of Classification, 31*(3), 274–295.

Napoli, P. M. (2018). What if more speech is no longer the solution: First Amendment theory meets fake news and the filter bubble. *Federal Communications Law Journal*, 70(1), 55–104. Retrieved from https://heinonline.org/HOL/Page?handle=hein.journals/fedcom70&id=67&collection=journals&ind ex=journals/fedcom

Nardo, Michela & Saisana, Michaela & Saltelli, Andrea & Tarantola, Stefano & Hoffman, Anders & Giovannini, Enrico. (2008). *Handbook on Constructing Composite Indicators and User Guide*. 10.1787/533411815016.

Nieminen, H. (2024). Why Does Disinformation Spread in Liberal Democracies? The Relationship between Disinformation, Inequality, and the Media. *Javnost - The Public*, 31(1), 123–140. https://doi.org/10.1080/13183222.2024.2311019

Norris, P. (2009). Comparative political communications: Common frameworks or Babelian confusion? *Government and Opposition*, 44(3), 321–340. DOI: http://dx.doi.org/10.1111/j.1477-7053.2009.01290.x

Ogden, C.K., & Richards, I.A. (1923). *The Meaning of Meaning*. Harcourt, Brace.

Online Safety Act 2023, c. 50. (UK). https://www.legislation.gov.uk/ukpga/2023/50/contents/enacted

Order PCM/1030/2020, of October 30, [Orden PCM/1030/2020, de 30 de octubre, por la que se publica el Procedimiento de actuación contra la desinformación aprobado por el Consejo de Seguridad Nacional] (Spain). (2020, November 5). Boletín Oficial del Estado (BOE), No. 292. https://www.boe.es/eli/es/o/2020/10/30/pcm1030/con

O'Connor, C., & Joffe, H. (2020). Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines. *International Journal of Qualitative Methods*, 19. https://doi.org/10.1177/1609406919899220 (Original work published 2020)

Paalman, Maria. (1997). How to do (or not to do)...: Media Analysis for Policy Making. *Health Policy and Planning*, Volume 12, Issue 1, 1997, Pages 86–91, https://doi.org/10.1093/heapol/12.1.86

Papathanassopoulos, S. (2007) The Mediterranean or Polarized Pluralist Model Countries. In: Terzis G (ed) *European Media Governance: National and Regional Dimensions*. Bristol: Intellect Book, pp. 191-200.

Perlman, A. (2021). Philip M. Napoli, *Social Media and the Public Interest: Media Regulation in the Disinformation Age*. *International Journal Of Communication, 15*, 4. Retrieved from https://ijoc.org/index.php/ijoc/article/view/17841

Phiri, C. (2023, November 30). *Political disinformation and freedom of expression: Demystifying the Net Conundrum*. [Doctoral dissertation, University of Turku]. UTUPub. https://www.utupub.fi/handle/10024/175986?show=full

Polit, D. F., & Beck, C. T. (2021). *Nursing research: Generating and assessing evidence for nursing practice,* (11th ed.). Wolters Kluwer.

Porter, D. O., & Olsen, E. A. (1976). Some Critical Issues in Government Centralization and Decentralization. *Public Administration Review*, *36*(1), 72–84. https://doi.org/10.2307/974743

Posetti , J., & Matthews, A. (2018). A short guide to the history of “fake news” and disinformation: A new ICFJ learning module. *International Center for Journalists*. https://www.icfj.org/news/short-guide-history-fake-news-and-disinformation-new-icfj-learning-module

Puppis, Manuel. (2010). Media Governance: A New Concept for the Analysis of Media Policy and Regulation. *Communication, Culture and Critique*, Volume 3, Issue 2, June 2010, Pages 134–149, https://doi.org/10.1111/j.1753-9137.2010.01063.x

Raboy, M. and Padovani, C. (2010), Mapping Global Media Policy: Concepts, Frameworks, Methods. *Communication, Culture & Critique*, 3: 150-169. https://doi.org/10.1111/j.1753-9137.2010.01064.x

Recommendation CM/Rec(2022)11 of the Committee of Ministers to Member States on Principles for Media and Communication Governance’ (adopted at the 1431st meeting of the Ministers’ Deputies, 6 April 2022)

Recommendation CM/Rec(2022)4 of the Committee of Ministers to member States on promoting a favourable environment for quality journalism in the digital age (Adopted by the Committee of Ministers on 17 March 2022 at the 1429 meeting of the Ministers’ Deputies), Preamble.

Reporters Without Borders. (2024). 2024 World Press Freedom Index. https://rsf.org/en/2024-world-press-freedom-index-journalism-under-political-pressure

Research integrity (RI). University of Helsinki. (n.d.). https://www.helsinki.fi/en/research/research-integrity/research-ethics/responsible-conduct-research

Rochefort, A. (2020). Regulating Social Media Platforms: A Comparative Policy Analysis. *Communication Law and Policy*, 25(2), 225–260. https://doi.org/10.1080/10811680.2020.1735194

Romanova, T., Sokolov, N., & Kolotaev, Y. (2020). Disinformation (Fake News, Propaganda) as a Threat to Resilience: Appoaches Used in the EU and its Member State Lithuania. *Political Regional Studies*, 53—67. https://balticregion.kantiana.ru/upload/iblock/5f0/4-Romanova_53-67.pdf

Rucinska, Silvia, Miroslav Fecko, and Ondrej Mital. (2023). Trust in public institutions in the age of disinformation. In *Proceedings of the Central and Eastern European eDem and eGov Days 2023* (CEEeGov '23). Association for Computing Machinery, New York, NY, USA, 111–117. https://doi.org/10.1145/3603304.3604075

Ruiz, Diaz C. (2023). Disinformation on digital media platforms: A market-shaping approach. *New Media & Society*, *27*(4), 2188-2211. https://doi.org/10.1177/14614448231207644

S.1989 - 115th Congress (2017-2018): A bill to enhance transparency and accountability for online political advertisements by requiring those who purchase and publish such ads to disclose information about the advertisements to the public, and for other purposes. (2018, June 26). https://www.congress.gov/bill/115th-congress/senate-bill/1989

Sabbagh, D. (2023, January 11). *National security bill may have “chilling effect” on investigative journalism in UK*. The Guardian. https://www.theguardian.com/media/2023/jan/11/national-security-bill-may-have-chilling-effect-on-investigative-journalism-in-uk#:~:text=foreign%20intelligence%20service%E2%80%9D.-,The%20Conservative%20peer%20highlighted%20a%20string%20of%20investigations%20that%20he,assist%E2%80%9D%20such%20a%20spy%20agency.

Salomaa, S., & Palsa, L. (2019, December 16). *Media literacy in Finland: National media education policy* (Publication No. 10024/162065). Ministry of Education and Culture, Finland. http://julkaisut.valtioneuvosto.fi/handle/10024/162065

Sakellariadis , J., & Miller, M. (2025). Trump continues federal purge, gutting cyber workers who combat disinformation - politico. *Politico*. https://www.politico.com/news/2025/02/07/trump-guts-cyber-workers-00203087

Sandelowski, M. (2000), Combining Qualitative and Quantitative Sampling, Data Collection, and Analysis Techniques in Mixed-Method Studies. Res. Nurs. Health, 23: 246-255. https://doi.org/10.1002/1098-240X(200006)23:3<246::AID-NUR9>3.0.CO;2-H

Sanfilippo, M. R., Zhu, X. A., & Yang, S. (2025). Sociotechnical governance of misinformation: An Annual Review of Information Science and Technology (ARIST) paper. *Journal of the Association for Information Science and Technology*, 76(1), 289–325. https://doi.org/10.1002/asi.24953

Sato, Y., & Wiebrecht, F. (2024). Disinformation and Regime Survival. *Political Research Quarterly*, *77*(3), 1010-1025. https://doi.org/10.1177/10659129241252811

Shahin, Saif. (2016). *When Scale Meets Depth: Integrating Natural Language Processing and Textual Analysis for Studying Digital Corpora. Communication Methods and Measures, 10(1), 28–50*. doi:10.1080/19312458.2015.1118

Shaping Europe's digital future. (2025, February 13). *The 2022 Code of Practice on Disinformation*. https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation

Shepherd, Amy. (2017). Extremism, Free Speech and the Rule of Law: Evaluating the Compliance of Legislation Restricting Extremist Expressions with Article 19 ICCPR. *Utrecht Journal of International and European Law* 33 (2017): 62-83. http://doi.org/10.5334/ujiel.405.

Shattock, E. (2023). Lies, liability, and lawful content: critiquing the approaches to online disinformation in the EU. *Common Market Law Review*, 60(5). https://kluwerlawonline.com/journalarticle/Common+Market+Law+Review/60.5/COLA2023094

Sillanpää, A. (2021). *The National Emergency Supply Agency builds the ability to counter hostile information influencing*. National Emergency Supply Agency (Huoltovarmuuskeskus). https://www.huoltovarmuuskeskus.fi/en/news/the-national-emergency-supply-agency-builds-the-ability-to-counter-hostile-information-influencing

Smarandache, Florentin & Vlăduţescu, Ştefan. (2014). Towards a practical communication intervention. *Revista de Cercetare si Interventie Sociala*. 46. 10.5281/zenodo.49167.

Soares, M. (2021, June 1). Verdade e Mentira Por Alvará. *PÚBLICO*. https://www.publico.pt/2021/06/02/opiniao/opiniao/verdade-mentira-alvara-1964925

Sperry, B. (2024, April 12). Knowledge and decisions in the information age: The law & economics of regulating misinformation on social-media platforms. International Center for Law & Economics. https://laweconcenter.org/resources/knowledge-and-decisions-in-the-information-age-the-law-economics-of-regulating-misinformation-on-social-media-platforms/

Spina, N. (2014). Decentralisation and political participation: An empirical analysis in Western and Eastern Europe. *International Political Science Review / Revue Internationale de Science Politique*, *35*(4), 448–462. http://www.jstor.org/stable/24573451

Stahl, Bernd Carsten. (2006). On the difference or equality of information, misinformation, and disinformation: A critical researchperspective. Informing Science 9: 83 *(PDF) Public Evaluations of Misinformation and Motives for Sharing It*. DOI: 10.28945/473

Standard Player Monthly. (1919). Talks with Tuners by One of Them, Volume 4, Number 2, Quote Page 9, Standard Pneumatic Action Company, New York City. (Google Books Full View) link

Stasi, M. L., & Parcu, P. L. (2021). "Chapter 21: Disinformation and misinformation: the EU response". In *Research Handbook on EU Media Law and Policy*. Cheltenham, UK: Edward Elgar Publishing. Retrieved February 21, 2025, from https://doi.org/10.4337/9781786439338.00030

Strauss T, von Maltitz MJ (2017). Generalising Ward's Method for Use withManhattan Distances. PLoS ONE 12(1): e0168288.doi:10.1371/journal.pone.0168288

Tambini, Damian. (2017). *Fake news: public policy responses*. LSE Media Policy Project Series, Tambini, Damian and Goodman, Emma (eds.) (Media Policy Brief 20). London School of Economics and Political Science, London, UK. https://eprints.lse.ac.uk/73015/

Tenove, C. (2020). Protecting Democracy from Disinformation: Normative Threats and Policy Responses. *The International Journal of Press/Politics*, 25(3), 517-537. https://doi.org/10.1177/1940161220918740 (Original work published 2020)

Tworek, H., & Leerssen, P. (2019). *An Analysis of Germany's NetzDG Law*. Transatlantic Working Group. https://pure.uva.nl/ws/files/40293503/NetzDG_Tworek_Leerssen_April_2019.pdf

UNESCO. (2023). *Higher levels of freedom of expression have a strong relationship with the protection of other human rights – UNESCO*. https://www.unesco.org/en/articles/higher-levels-freedom-expression-have-strong-relationship-protection-other-human-rights-unesco

United Nations (General Assembly). (1948). Universal Declaration of Human Rights (UDHR). *New York: United Nations General Assembly, 1948*. https://www.un.org/en/about-us/universal-declaration-of-human-rights.
United Nations (General Assembly). (1966). International Covenant on Civil and Political Rights. Treaty Series, 999, 171.
United States Department of State. (2023, October 5). *About us – Global Engagement Center* [Archived]. https://2021-2025.state.gov/about-us-global-engagement-center-2/
UNHRC, 'General Comment No 34' (12 September 2011) UN Doc CCPR/C/GC/34, 2, 36.
V-Dem Institute. (2025). "Democracy Report 2025: 25 Years of Autocratization – Democracy Trumped?". Varieties of Democracy Institute. https://www.v-dem.net/documents/60/V-dem-dr__2025_lowres.pdf
Vese, D. (2022). Governing Fake News: The Regulation of Social Media and the Right to Freedom of Expression in the Era of Emergency. *European Journal of Risk Regulation*, 13(3), 477–513. doi:10.1017/err.2021.48
Voltmer, K. (2011). How Far Can Media Systems Travel?: Applying Hallin and Mancini's Comparative Framework outside the Western World. In D. C. Hallin & P. Mancini (Eds.), *Comparing Media Systems Beyond the Western World* (pp. 224–245). chapter, Cambridge: Cambridge University Press.
Wardle, C., & Derakhshan, H. (2017). *Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking*. Council of Europe. https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html#
https://rm.coe.int/information-disorder-report-november-2017/1680764666
Weigand, Anna Christina; Lange, Daniel; Rauschenberger, Maria. (2021): How can Small Data Sets be Clustered?.*Mensch und Computer 2021 - Workshopband*. DOI: 10.18420/muc2021-mci-ws02-284. Bonn: Gesellschaft für Informatik e.V.. MCI-WS02: UCAI 2021: Workshop on User-Centered Artificial Intelligence. Ingolstadt. 5.-8. September 2021
Welch, D. (2014). Propaganda for patriotism and nationalism. British Library: Accessed: 28/0318
Yanovitzky, I., & Weber, M. (2020). Analysing use of evidence in public policymaking processes: a theory-grounded content analysis methodology. *Evidence & Policy*, 16(1), 65-82. Retrieved Apr 12, 2025, from https://doi.org/10.1332/174426418X15378680726175
Zeller R. A., Zeller R. A., Carmines E. G. (1980). *Measurement in the social sciences: The link between theory and data*. Cambridge University Press.
Zhu, Xiaohua & Yang, Shengnan. (2023). Toward a Sociotechnical Framework for Misinformation Policy Analysis. *In The Usage and Impact of ICTs during the Covid-19 Pandemic*. DOI: 10.4324/9781003231769-3

8 Appendices

Appendix A

Bontcheva et al. (2020) and Cipers et al. (2023a) taxonomy of disinformation policies

CATEGORY	DEFINITION
Monitoring and Factchecking responses	Usually carried out by news organisations, internet communications companies, academia, civil society organisations, and independent fact-checking organisations, as well as (where these exist) partnerships between several such organisations. These responses share a focus on informing the public on the authenticity and credibility of news stories, online chain messages and popular online narratives.
Investigative responses	Go beyond whether a given message/content is (partially) false, to provide insights into disinformation campaigns, including the originating actors, degree of spread, and affected communities.
Countercampaigns	Tend to focus on the construction of counter narratives by governments with the explicit aim of countering false or falsely deemed discourses.
Election specific responses	Developed specifically to detect, track, and counter disinformation that is spread during elections. This category of responses, due to its very nature, typically involves a combination of monitoring and fact-checking, legal, curatorial, technical, and other responses, which will be cross-referenced as appropriate.
Curational responses	Address primarily editorial and content policy (removing, flagging, … of posts, journalistic articles, forums, blogs, …) and 'community standards', although some can also have a technological dimension, which will be cross-referenced accordingly.
Technical and algorithmic responses	Use algorithms and/or Artificial Intelligence (AI) in order to detect and limit the spread of disinformation or provide context or additional information on individual items and posts. These can be implemented by the social platforms, video-sharing, and search engines themselves, but can also be third party tools (e.g., browser plug-ins) or experimental methods from academic research. This category also includes more basic technology-based responses such as limited or complete internet 'shutdowns'.
Demonetising and economic responses	Designed to stop monetisation and profit from disinformation and thus discourage the creation of clickbait, counterfeit news sites, and other kinds of for-profit disinformation. Rather than punitive measures such as fines, this category entails responses that focus on stopping disinformation practices from generating income and profits.
Ethical and normative	Public condemnation of acts of disinformation or recommendations and resolutions aimed at thwarting these acts. When carried out by a government actor, this "condemnation" is often executed as a result of a perceived security threat to the state and her population which legitimises government action.
Educational	Aims at promoting citizens' media and information literacy, critical thinking, and verification in the context of online information consumption, as well as journalist training.
Empowerment	Designing content verification tools and web content indicators, which are practical aids that can empower citizens and journalists to avoid falling prey to online disinformation. These efforts may also be intended to influence curation in terms of prominence and amplification of certain content – these are included under curatorial responses above.
COVID-19 specific responses	Developed specifically to detect, track, and counter disinformation concerning the Covid-19 pandemic. In analogy with electoral-specific responses, these responses typically encompass different types of responses which are cross-referenced in the data. While other responses are used to curb disinformation surrounding the coronavirus, initiatives are Covid-19 specific when they are introduced as a response, or when law(s) (proposals) are clearly amended, to disinformation that targets facts surrounding COVID-19, and government initiatives to curb the spread of the COVID-19 virus.

Back to text

Appendix B

Coding dictionary

Dimensions

Clarity of Definition of Disinformation (0–3): Assesses how precisely the policy defines disinformation and related concepts.

Score	Description	Example
0	No guiding definition or ambiguous reference to disinformation	"False or misleading content" without further elaboration and
1	Mentions disinformation but without specifying intent or harm	"Spreading of disinformation online"
2	Mentions 2 out of 3 clarity conditions: falsity, harm, intent	"Purposeful dissemination of false information during elections", “narrative that is demonstratably false and likely to cause harm”
3	Clearly defines disinformation using falsity, intent, and harm	"Disinformation refers to intentionally false content disseminated to cause public harm or deceive voters”

Centralization of Regulatory Authority (0–3): Measures how much power the state has in deciding what constitutes disinformation and how content is moderated.

Score	Description	Example
0	No regulatory authority	No mention of enforcement or oversight actors
1	Co-regulatory: Independent regulators, judicial bodies involved	Dispute handled by judiciary or independent regulatory body, Media Regulator
2	Delegated to private platforms with government oversight	NetzDG: platforms make decisions under threat of state fines
3	Statutory: Government or ministries decide directly	Law authorizes government ministry to define or remove content (e.g., Hungary COVID-19 law)

Strictness of Regulation (0–4): Captures the severity of penalties or consequences imposed for spreading disinformation.

Score	Description	Example
0	No penalties or purely educational measures	Media literacy or Information campaign
1	Administrative burdens	Licensing regimes, data localization, transparency requirements, or mandated press or media councils
2	Content regulation	Social media must remove content within 24h (Germany NetzDG)
3	Civil penalties or lawsuits	Possibility of civil litigation or large fines
4	Criminal penalties (e.g., imprisonment)	Hungary: up to 5 years in prison for COVID-19 disinfo

Multiplicative weights:

Legal status(0.6-1.0): Represents normative or regulatory weight.

Weight	Legal Status Type	Description (Cipers et al., 2023)	Justification of weight
0.7	Proposed legislation	Law that aims to curb the threat of disinformation has been proposed but is still in the process of being approved.	Lack immediate enforceability (Flew & Martin, 2022). (Funke & Flamini, 2019).
0.8	Counter narratives	Pre-legislative initiatives with the explicit aim of countering false or falsely deemed discourses	Explicitly designed to intervene in public debates,but they do not impose penalties or formal restrictions on speech (Bradshaw & Lim, 2023).
0.8	Non-legislative initiatives	Government initiatives without legal basis and enforcement (soft law)	Soft law, relying on voluntary compliance rather than binding regulations (Wardle & Derakhshan, 2017)
0.9	Adopted legislation	Initiatives adopted into the national law against disinformation-related offences (hard law)	Binding nature and legal authority influencing both media platforms and individual speech (Bayer & Bárd, 2020)
1.0	Law enforcement	Law is enforced against disinformation-related offences	Laws have the most direct impact on press freedom and media regulation, as they involve tangible legal consequences (Rediker, 2021)

Activity Status (0.5-1.0): Whether policy remains legally or practically in force. Inactive policies on average received higher RFE score of 14.47, while active policies on average scored 8.83. Based on the observed average difference between active and inactive policies, inactive policies were weighted at 0.8 to reflect their reduced impact in current governance.

Weight	Activity Status Type	Description
0.8	Inactive	Outdated, repealed, no longer updated, no longer has legal weight, or COVID-19 specific
1.0	Active	Active, available to the public, has legal weight

Back to text

Appendix C

Policy response	Legal Status	Year	Country	Clarity (0-3)	Regulation (0-3)	Strictness (0-4)	Activity (0/1)
Australia's Electoral Assurance Taskforce	Non Leg. Initiative	2019	Australia	3	3	2	1
Australia's Code of Practice on Disinformation and Misinformation	Non Leg. Initiative	2021	Australia	3	2	1	1
Australia's COVID-19 Mythbusters	Counter Narratives	2021	Australia	2	0	0	0
Finland's Media Education Policy	Adopted Leg.	2013	Finland	3	0	0	1
Finland's Media Education Policy Update	Adopted Leg.	2019	Finland	3	0	0	1
Finland's social media influencers campaign	Counter Narratives	2020	Finland	2	1	0	0
Finland's National Emergency Supply Agency (NESA)'s Knowledge Centre on Information Resilience.	Counter Narratives	2022	Finland	3	1	0	0
France's National Agency for the Fight Against Manipulations of Information	Non Leg. Initiative	2021	France	2	1	1	1
France's Fight against Manipulation of Information Law	Law Enforcement	2018	France	2	1	2	1
France's Media and information literacy (or EMI – Education aux médias et à l'information)	Adopted Leg.	2013	France	3	0	0	1
Germany's federal task force on Hybrid Threats	Non Leg. Initiative	2022	Germany	3	0	0	1
Germany's Act to Improve Enforcement of the Law in Social Networks	Adopted Leg.	2017	Germany	1	2	2	0
Germany's Network Enforcement Act update	Law Enforcement	2021	Germany	1	2	2	0
Germany: Implementation of increased reporting duties for platforms in amendment to NetzDG	Adopted Leg.	2022	Germany	1	2	2	0
Germany's Live Democracy! Project's Cooperation Network Against Online Hate and Disinformation	Non Leg. Initiative	2022	Germany	3	0	0	1

Germany's Interstate Media Treaty Amendment	Adopted Leg.	2021	Germany	2	1	1	1
Germany's Network Enforcement Act	Law Enforcement	2018	Germany	1	2	2	0
Hungary's covid-19 misinformation law	Adopted Leg.	2020	Hungary	1	3	4	0
Hungary's covid-19 misinformation law enforcement	Law Enforcement	2020	Hungary	1	3	4	0
Hungary's “White Paper” of the Digital Freedom Committee	Non Leg. Initiative	2021	Hungary	1	0	0	1
Latvia's Media Law Enforcement	Law Enforcement	2010	Latvia	2	3	3	1
Latvia's government-backed anti-disinformation platform	Non Leg. Initiative	2023	Latvia	3	0	0	1
Latvia's Amendment to Criminal Code: Influencing the electoral process using deep fraud technologies	Law Enforcement	2024	Latvia	3	1	0	1
Latvia's National Strategy for the Development of the Electronic Media Sector for 2023–2027	Adopted Leg.	2022	Latvia	3	2	0	1
Netherlands' Code of conduct on transparency of online political ads	Non Leg. Initiative	2021	Netherlands	2	2	1	1
Netherlands’ disinformation awareness campaigns	Non Leg. Initiative	2019	Netherlands	3	1	0	1
Netherlands' national strategy against disinformation	Non Leg. Initiative	2022	Netherlands	3	1	0	0
Dutch government’s strategy to combat disinformation	Non Leg. Initiative	2024	Netherlands	3	1	0	1
Portugal's "Defence against fake news" website	Non Leg. Initiative	2021	Portugal	3	0	0	1
Portugal's Portuguese Charter of Human Rights in the Digital Age.	Adopted Leg.	2021	Portugal	2	1	2	0
Simplified Portuguese Charter of Human Rights in the Digital Age.	Adopted Leg.	2022	Portugal	3	0	0	1
Slovakia's police Facebook page	Non Leg. Initiative	2020	Slovakia	0	0	0	1
The Act on Cyber Security (69/2018 Coll.)	Adopted Leg.	2018	Slovakia	1	3	2	1
Slovakia's Sybersecurity Act Update	Leg Proposals	2022	Slovakia	1	3	2	1

Concept Strategic Communication							
Concept of the Slovak Republic Strategic Communication	Leg Proposals	2024	Slovakia	1	0	2	1
Slovakia's The Media Services Act (264/2022 Coll.)	Adopted Leg.	2022	Slovakia	2	1	0	1
Spain's Procedure for Intervention against Disinformation	Adopted Leg.	2020	Spain	3	3	0	1
Spain's government hybrid threats unit	Non Leg. Initiative	2019	Spain	1	0	0	1
Spain's REGULATING THE RIGHT OF RECTIFICATION	Leg Proposals	2024	Spain	1	3	2	1
UK Foreign and Commonwealth Office Counter Disinformation and Media Development programme	Counter Narratives	2016	U.K.	1	0	0	1
UK's Counter Disinformation Unit/Covid Rapid Response unit	Non Leg. Initiative	2020	U.K.	3	3	0	0
UK's House of Commons (Digital, Culture, Media and Sport Committee) inquiry into 'Disinformation and 'Fake News"	Non Leg. Initiative	2019	U.K.	3	1	1	1
U.K.'s Online Harms White Paper	Leg Proposals	2019	U.K.	3	1	1	1
U.K.'s Online Safety Act 2023	Adopted Leg.	2023	U.K.	3	2	2	1
The Act Foreign Interference Offence	Adopted Leg.	2023	U.K.	2	1	4	1
U.S.' Global Engagement Centre	Adopted Leg.	2016	U.S.	2	1	2	0
U.S.' Devumi Case	Adopted Leg.	2019	U.S.	0	3	2	0
U.S.'s Educating Against Misinformation and Disinformation Act	Adopted Leg.	2022	U.S.	1	0	0	1
Honest Ads Act	Adopted Leg.	2017	U.S.	2	3	1	0
U.S.' Disinformation Governance Board	Adopted Leg.	2022	U.S.	3	0	0	0

Back to text

Appendix D

External cluster validation

I conducted a series of cluster validation statistics to evaluate hierarchical clustering groupings and prove it offers a more meaningful grouping than the original media system typology. By comparing my data-driven results to existing media systems grouping, I am performing what Charrad et al. (2014) referred to as "external cluster validation", which involves the comparison of clustering results to an externally validated results, which in my case are ideal media systems. First, I compared within-group variances, which measure how tightly grouped the clusters are. The variances within the media system clusters confirmed the visual incosistencies, with Liberal system showing low variance (1.32), while the Democratic Corporatist, Post-Communist, and Polarized-Pluarist clusters had average RFE variances of 16.13, 13.48, and 10.81 respectively. In contrast, the variances within hierarchical clusters show tighter groupings, with within-cluster variances ranging between 0.00 and 2.21. Secondly, I ran Kruskal-Wallis test to compare whether there are significant differences between cluster medians (Kruskal & Wallis, 1952). HCA clusters showed a statistically distinct improvement ($p = 0.00000$), suggesting better explanatory potential, comparing to media systems; result ($p = 0.001$), Lastly, I ran silhouette analysis for my HCA results to measure how similar a country is to its assigned cluster compared to other clusters (Monshizadeh et al., 2022). However, the results indicated only weak-to-moderate cohesion (mean silhouette width ≈ 0.38), raising concerns about cluster stability. Small sample size of countries is often attributed to inconsistencies in results, especially if data clusters vary in size ($n = 12$) (Ibid.). The stability of the cluster groupings would have to be validated with an expanded sample of countries. Nevertheless, by clustering countries based on the risks scores of actual policy characteristics, rather than relying on ideal typologies, this study shows how countries of varied media systems may share patterns within their disinformation regulation.

Back to text