Rebuilding Public Trust in Social Media Platforms? A Case Study of Meta's Oversight Board

Rumeng Cao

A thesis submitted to fulfil requirements for the degree of

Master of Philosophy

School of Art, Communication, and English

Faculty of Arts and Social Science

The University of Sydney

2026

Statement of Original Authorship

This is to certify that, to the best of my knowledge, the content of this thesis is my own work.

This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Generative AI statement

No generative AI tools were used as part of the research project or to assist with writing.

Acknowledgement of FASS funding

This research was supported by the Faculty of Arts and Social Sciences Postgraduate Research Scholarship (SC3602).

Acknowledgements

This thesis is dedicated to my supervisor, Professor Terry Flew. Without my supervisor's continuous guidance and support, I would never have had the opportunity to pursue my studies abroad. I will carry this kindness with me for the rest of my life.

I am also deeply grateful to my co-supervisor, Dr Xiang Ren, whose support and encouragement accompanied me throughout this journey. My sincere thanks extend to all the participants who generously shared their time and insights, providing invaluable perspectives that made this research possible.

I also want to express my gratitude to the Faculty of Arts and Social Sciences for providing me with a scholarship that covers my tuition and living expenses, enabling me to continue my research.

I remain profoundly thankful for every act of kindness and help I received along the way. I would also like to express my gratitude to my friends in both China and Australia. In my own way, I love you all.

Finally, I owe my deepest gratitude to my grandmother for her lifelong support and care. Thank you for looking after me since my earliest days. Now, it is my turn to take care of you.

This journey has been filled with both tears and joy. Thank you for taking the time to read this thesis. Although we may never meet, reading allows us to connect in a meaningful way. I hope you enjoy this work.

Thanks to Dr. David Harris from Writewell Australia for proofreading the thesis.

Abstract

Social media platforms have become an integral part of everyday life. Power within the digital environment has become concentrated in a few dominant platforms because of user behaviour and platform design. This concentration allows these platforms to exert a strong influence on civic discourse and the public sphere. However, the growing prevalence of misinformation has transformed how public discussions unfold online.

Companies, such as Meta, have faced ongoing criticism for permitting harmful content to spread and for their misuse of user data. These companies' ineffective responses to these issues have further undermined public trust. Trust plays a crucial role in the digital era. It determines whether users remain active and engaged on social media platforms. To restore public trust, Meta established the Oversight Board as an independent body to review user appeals for removing their content and to enhance transparency and accountability in content moderation.

This thesis investigates the effectiveness of Meta's Oversight Board and examines its decision-making processes. Following the platform regulation triangle theory, this study employs the methodology of document analysis and semi-structured interviews. This study analyses Facebook's regulatory history from 2010 to 2020 and evaluates Meta's current strategies for addressing misinformation. This study's findings contribute to broader debates on platform regulation and offer practical insights for other social media platforms seeking to improve content governance.

Keywords

Meta, Oversight Board, content moderation, platform regulation, trust

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Context

The development of Internet platforms typically goes through three stages. The first is the initial open internet stage; the second is the social media platform and platformisation stage; the last stage is the platform regulation stage. With the wide popularity of social media, like Facebook and Instagram, the number of social media users dramatically increased. Consequently, the user-generated content has grown rapidly; therefore, content moderation has become a critical issue. The rising volume of user-generated content has undoubtedly presented unprecedented challenges to content review. Hundreds of millions of daily content updates cannot be reviewed by relying on manual review models alone, as was once the case. Platforms have begun to adopt algorithmic and artificial intelligence (AI) moderation mechanisms to identify non-compliant or harmful content so as to maintain a healthy online environment. However, as pointed out by Noble (2018), AI moderation tends to display biased outcomes and erroneous judgments because the social biases and cultural assumptions of AI developers are reflected in algorithm design and training. On the other hand, manual moderation finds it difficult to address complex cases in a timely way because of insufficient staffing and delayed responses by human moderators to moderate content. As the volume of content increases and the complexity of the review scale grows, this task becomes increasingly challenging for human moderators. The combination of technical bias and the lag in human response has eroded the trust of users in platform

regulation.

The regulation of content on social media platforms is currently a major challenge faced by all countries all over the world. People's perception of their rights on the Internet has changed; however, the platform regulation is insufficient, with lagging development. The conventional self-regulation model is unable to satisfy the demands of society for accountability and transparency. The governance of content on social media platforms should be carried out in a differentiated and focused manner. Meta's Oversight Board, as an 'independent third-party mechanism', was established to provide users with an appeal channel and review platform content decisions. In this way, governance transparency can also be enhanced. Nevertheless, it remains unclear whether this governance mechanism is independent and can effectively monitor platform power. As noted by Helberger (2020), Meta's Oversight Board was established by Meta for the purpose of reviewing content moderation decisions; in this vein, the Board can be regarded as a way of self-correcting measures for platform regulation. This study explores whether Meta's Oversight Board could work as the 'supreme court' of Meta, conducting content moderation within the legal framework and forming an effective mechanism for rebuilding trust.

1.1.1 Open Internet, Digital Platforms and Platformisation

With the rapid development of the internet, dramatic changes have happened in people's daily lives. From the initial information era to the current digital era, the internet industry has gone through a revolutionary transformation. In the information era, the Internet was used to collect, store and transmit information. People can obtain

the information they want by searching on the Internet. However, as the number of Internet users sharply rises and massive amounts of information appear, the Internet industry is prompted to move towards a new era.

The transformation of the Internet from ‘open networks’ to ‘digital platforms’ has been a significant turning point in the development process of technological architecture and information communication. This transformation, which has encompassed technological advancement, has reshaped power distribution, modes of information circulation, as well as forms of social relationships. The early Internet, which was open, decentralised and liberalised, was primarily featured with interconnection and information sharing. The open Internet enables users to enter the Internet without obstacles, for users to receive and post information. Without information barriers and central control nodes, people were able to participate in the creation and dissemination of knowledge.

The open architecture of the Internet has combined programmable interfaces with intermediary mechanisms, which technically and structurally underpin ‘platformisation’. As stated by Helmond (2015), open application programming interfaces (APIs) have played an essential role in catalysing the transformation of the Internet from social networks to social media platforms. They are endowed with ‘programmability’ to empower platforms to absorb, integrate and regulate external resources. In addition, with APIs, developers and third-party websites are enabled to interact with platforms. ‘Facebook’ developed as a social media platform originates from its supply of open APIs and its role as a social space. Facebook is no longer just a

‘website’. Instead, it is an ecosystem: the boundaries of Facebook are expanded via technology; the functions of the platform are enriched by other websites. Facebook is the operator of the social media platform, using information technology to connect the interacting people, institutions and resources in the ecosystem, creating unexpected value and conducting value exchange.

Facebook, officially established in 2004, is a social networking service and social media website based in the United States, with its extensive user base and strong social influence. That year (2007) marked its transition from a social network to a platform enterprise (Helmond, 2015). A multi-faceted market structure was developed by connecting users, developers, advertisers, and external websites. The platform’s openness involved the technical sharing of APIs, the social signification of power reconstruction, and the economic amplification of network effects. It is hoped that the platform can achieve interconnectivity. The activities of users and third parties rely on the interfaces and algorithmic logic of Facebook. In this vein, social media platforms, such as Facebook, have become the infrastructure for the digital economy and information dissemination.

According to Gillespie (2010), platforms possess the capability of technical programmability or code execution and provide opportunities for communication, interaction, and transaction. The term ‘platform’ refers to a ‘neutral basis’ and a ‘power lever’. It is essentially an infrastructure that provides the necessary resources and environment, enabling users to carry out various activities such as information exchange, resource sharing, and collaborative cooperation. Social media platforms

provide an online space for users to express their ideas. However, it also provides structural advantages in terms of supervision and control for platform operators. Large-scale internet platforms integrate advanced technologies, vast capital and efficient organisations, and have become outstanding representatives of innovation and pioneering spirit. A vigorous digital economy, especially the development of social media platforms, has profoundly shaped people's daily lives. The growth process of platformisation constitutes both an evolution of technical architecture and a reconfiguration of socio-political-economic power.

As pointed out by Castells (1997), the network society was not produced by information technology itself. Without the transformation of information technology, the network society would not be able to form a comprehensive and widespread social form. The network society is the outcome of the reorganisation and restructuring of all aspects of society brought about by information technology. Unlike the traditional definition of network society in sociology, it is based on modern information technology.

The network society could form the basic structure of contemporary social organisations by permitting networks to connect a variety of human activities. Such infrastructural logic promoted the development of the early Internet, which realised free information flow through technological interconnection. Platformisation is one of the typical characteristics of the digital economy. A platform serves as an intermediary that brings together different users and acts as the infrastructure for user activities. The subsequent process of platformisation has been built upon new control mechanisms, including algorithms, interfaces, data and policies.

The rise of the Internet has driven the expansion of global communication, commerce, and social infrastructure (Flew, 2023). The Internet has facilitated the reorganisation of globalisation by allowing users to communicate in real time and share knowledge and cultural resources. As a universal technology, it has become the core of the global information infrastructure. It uses a combined ground-air information high-speed channel as the transmission medium and increasingly popular multimedia computers as the sending and receiving tools. It is a highly efficient, large-capacity and highly open form of communication medium. Nevertheless, due to the platform-centred development model of the Internet, this global interconnection is no longer 'symmetrical'. The flow of information now relies on the algorithmic distribution logic of platforms, and the visibility of content is determined by the policy orientation and commercial interests of those platforms. Platforms have shifted from 'neutral conduits' to 'active gatekeepers' and from 'connectors' to 'filters'.

Since the early 21st century, the Internet has undergone several major transformations. 'Platformisation', as the greatest structural impact (Flew, 2021a), describes the evolutionary trajectory of Internet infrastructure and reveals the logic underlying the reconfiguration of the digital economy and social relationships. 'Platformisation' means that interpersonal communication, information circulation, and mass media dissemination are dominated, filtered, and regulated by social media platforms and large technology companies (Flew, 2023, p. 16). The platform serves as a structured means of connecting individuals and society, which constantly constructs new social relationships. Furthermore, the platform increasingly permeates into the daily lives of

individuals and continuously becomes a fundamental driving force for constructing society. Platforms shift from ‘intermediaries’ to social infrastructure, and they wield structural power: they determine what information is seen, which interactions are amplified, and which voices are marginalised.

It is worth noting that large platform enterprises have gradually established their dominant market position and systemic influence. Scholars have observed a new wave of ‘antitrust populism’ catalysed by the concentration of corporate power (Shapiro, 2018). Data indicate that the combined market capitalisation of technology giants, such as Amazon, Google, Apple, Meta and Microsoft, represented two-thirds of the value of the Standard & Poor’s (S&P) 500 index from 2015 to 2019, which suggests their dominance in the global economy. Moreover, this landscape was further consolidated by the COVID-19 pandemic. With ‘working from home’, ‘distance education’ and ‘online social contact’ becoming prevalent, platform-based enterprises experienced a dramatic increase in the demand for digital services, and these enterprises were even more highly valued by capital markets (Godding, 2020). With the continuous development of technologies such as big data, cloud computing and artificial intelligence, consumption scenarios incorporating AI elements have gradually permeated into all aspects of people’s daily consumption behaviours. Furthermore, the significant development of AI has enhanced the dominance of these high-tech icons.

Although platformisation has brought about economic prosperity and technological innovation, it has been accompanied by profound social contradictions. Large-scale platforms are those internet technology companies with massive active users. They

have become an important information infrastructure, possess dominant positions in the industry and undertake the responsibilities of daily network supervision and content moderation. Large-scale platforms have become the de facto gatekeepers of global online content with the power to determine what information is made available, and which speech is amplified or suppressed. As pointed out by Flew (2024), such power exerts direct effects on public discourse, cultural expression, and political participation. A novel form of informational power is constituted through algorithmic mediation, with neutral technological governance to exercise de facto social control. With the continuous development of the network society and the continuous improvement of technological levels, data security and other risks have also been increasing. Information transparency, personal privacy and democratic accountability are faced with the unprecedented challenges of intensifying data concentration by multiple digital platforms and increasing user dependence on platforms.

As claimed by Flew (2019), Internet platformisation should be studied through a broader socio-economic and political governance framework. In addition, the concept of 'platform capitalism' has been proposed (Srnicek, 2017a), which has occurred with the rise of platform enterprises (Parker et al., 2016). Platforms represent complicated power networks on the basis of multilateral market logic, data-driven models, as well as algorithmic governance systems. Traditional regulatory means are difficult to address the cross-sector and cross-border governance problems faced by platforms, as the operation of social media platforms is global, while traditional regulation is based on the national level. Unbounded social media communication and bounded national

governance are difficult to reconcile. Social media in the digital age not only profoundly alters the international landscape but also brings new challenges to international internet governance.

The 'platformisation' of the Internet represents a fundamental reorganisation of traditional institutions and an evolutionary consequence of technologies. Novel challenges for regulation and governance have arisen due to the changes in economic production modes, information dissemination logics, and public power architectures. When platforms serve as both economic foundations and cultural intermediaries for controlling data and algorithms, the societal impacts of these changes go far beyond the scope of 'tech companies'. Digital regulation is a new type of regulation that emerges as digital technologies are increasingly widely applied in economic, social and political life. Digital technology is used as a tool or means within the existing regulation system, with the aim of enhancing regulation efficiency. The core issue of digital regulation in the 21st century is that a new regulatory framework should be established in the context of globalisation and digitalisation to safeguard competition, fairness, and public interest.

The regulatory shifts involve a discursive shift regarding the Internet and the digital environment. According to Bowers and Zittrain, the first wave of thinking about the Internet, which covered the 1990s and early 2000s (Schlesinger, 2020), was closely related to rights discourses and positive digital frameworks. The second wave of thinking, from the late 2000s to the late 2010s, was concerned with harms, risks, and public health. A much stronger focus is on the negative consequences of strongly linking populations by means of seemingly abstract algorithmic processes (Bowers &

Zittrain, 2020). Bowers and Zittrain argued that the third wave of thinking can be characterised as the digital governance era. A main question of this third wave thinking is whether accountable orientation can be achieved within the structure of existing platforms. If not, responsibility for important aspects of content governance must be assumed, at least partially, outside of platforms, in organisations that are independent of the platforms' business interests.

1.1.2 Platform Regulation

With the advancement of communication technology, the Internet has integrated into people's lives on an unprecedented scale. Social media has become the primary channel for people to exchange information and obtain news. Platform regulation has become extremely important to maintain a healthy environment on the Internet. Platform regulation, particularly content moderation exemplified by Facebook, has exposed the unprecedented governance challenges encountered by technology companies when they are confronted with vast volumes of user-generated content. As social media develops and proliferates, the landscape of information dissemination has experienced a fundamental shift from the centralised dissemination model of traditional media to a user-driven, decentralised ecosystem. Ordinary users are no longer mere information recipients; they have turned into content creators and communicators. Since Facebook entered the era of user-generated content, platforms have generated immense amounts of texts, images, videos, and comments daily. Rather than merely 'social tools' for connecting with their friends and acquiring information about their lives, social media platforms have gradually become important spaces for public discourse, political

expression, as well as social mobilisation.

Nevertheless, the massive production and circulation of information through platforms has also raised unprecedented regulatory challenges. One of the most critical challenges is the large-scale control of content. Facebook sees new content produced every second, as hundreds of millions of users are active every day. Conventional manual moderation has become unrealistic in the case of such a huge data stream. It is difficult for thousands of moderators to quickly and accurately determine which user contributions violate community standards and which belong to marginal or ambiguous ‘grey zones’. In the current digital age, the security issues of social platforms have drawn increasing attention. In light of this, Facebook introduced machine learning algorithms and automated moderation systems to identify misinformation, hate speech, violent imagery, and other harmful content. Through deep learning and natural language processing technologies, AI can automatically identify and analyse inappropriate remarks in text, including hate speech, malicious attacks, false information, etc. However, platform regulation faces challenges associated with misjudgement, opacity, and algorithm bias placed in the tension between ‘social justice’ and ‘technological rationality’, because the algorithms designed by the platform are not for the public interest but for commercial interests.

In the global operation of Facebook, cross-cultural and cross-contextual complexities should be navigated through content moderation. The same image or phrase can have completely different connotations across languages, nations, and political landscapes. When formulating global uniform community standards, the platform must strike a

balance between platform values and freedom of speech to accommodate diverse cultural contexts. As a result of these challenges and complexities, content policies have been enforced contentiously and inconsistently. Facebook has been criticised for 'inconsistent enforcement': under-enforcement by proliferating harmful information, and over-enforcement by removing legitimate expression.

The continuous increase in external regulatory pressure has further exacerbated the challenges in governance. Governments and the public expect platforms to take on greater social responsibilities and to take proactive measures against misinformation, hate speech, and extremist content. Nevertheless, platforms are inclined to face difficult trade-offs between 'free speech', 'corporate interests' and 'social responsibility'. Due to the significant differences in legal systems and political expectations across countries, the balance between compliance and openness creates an ongoing challenge.

The content moderation on Facebook constitutes not only a technical and managerial issue, but also a political and ethical concern. With unprecedented power over algorithmic control and information dissemination, platforms lack democratic accountability and institutional restriction. The complex interactions between information circulation, technological control, and social trust are demonstrated by Facebook's dilemmas relating to large-scale content moderation. The challenges faced by Facebook also provide a real-world case study for the management relationship between platform power, governance legitimacy, and user trust.

1.1.3 Rebuilding of User Trust

Apart from the challenges of platform regulation, the decline in users' trust in platforms has further diminished the social, economic and political influences of digital media platforms. Trust forms the bedrock upon which any information ecosystem operates. Once trust is eroded, the legitimacy and public value of platforms are subject to severe scrutiny. As asserted by Coleman (2012), 'trust can be defined as to meet social expectations…the more that such expectations are disappointed, the greater the risk to relationships of trust will be (p. 37).' Nevertheless, the expectations of people for Meta have remained unmet for an extended period. Trust is the foundation of interpersonal relationships, and social media occupies a large proportion in our contemporary interpersonal and social relationships. Trust exerts effects on what specific information is disseminated on the Internet between members of a given group and how that information is disseminated (Kacperska et al., 2022). Since the Cambridge Analytica scandal in 2018, Facebook (now Meta) has faced a global public outcry over its misuse of user data. From this incident, the public learned for the first time how political forces have exploited the vast datasets held by platforms to manipulate public discourse and influence elections. As a fracture point of trust, the incident marked a shift in the public perception of digital platforms from technological intermediaries to powerful institutions. At the global level, the digital society is a place where multiple institutions are no longer trusted. As mentioned by Liu and Mehta (2024), the omnipresence of untrustworthy information poses constant challenges to the information ecosystem, since false information distributed on social media platforms poses challenges to

moderating content.

The proliferation of misinformation has weakened user trust in the quality of information and the effective regulation of platforms. A well-designed platform system is more inclined to push content that can encourage user participation and prolong their stay on the platform, rather than genuine and rational public discussions. This emphasis on user engagement and time on the platform may lead to information pollution and social polarisation through misleading information dissemination, extreme viewpoints and even hate speech. In such circumstances, users become increasingly sceptical of the information they obtain on platforms. It is believed that social media is an opinion arena dominated by commercial interests and manipulated by algorithms, and not a neutral space for public discourse.

Social media platforms have attempted to combine algorithmic identification and manual review to strengthen content governance. However, since designed platform tools cannot guarantee the public interest, such as amplifying misinformation on social media platforms, a new crisis of trust has been triggered by the mechanism per se. Firstly, users have found it difficult to comprehend the criteria of moderating content by algorithms, resulting in the deletion of content or a decline in quality due to the lack of transparency in algorithmic moderation. Secondly, the credibility of the social media platform has been undermined by cultural bias and inconsistent enforcement arising from manual moderation. In fact, the platform has sole authority over the ‘life and death of information’. Hence, a classic ‘closed self-regulation system’ is created by setting rules and enforcing rulings. User trust in the platform has been steadily eroded by an

opaque decision-making process and the platform's highly centralised power structure.

Meta's Oversight Board, as an independent third-party oversight mechanism, was established in 2020 in response to the crisis of trust of users and external pressure from the public and the government. The Oversight Board's purpose was to decentralise content moderation authority, improve transparency, and rebuild public trust through 'institutionalised oversight'. Positioned as an 'independent quasi-judicial body', Meta's Oversight Board was given the authority to review content decisions on certain platforms and launch binding rulings alongside policy recommendations. This mechanism provides users with a symbolic platform to express their opinions, enabling them to question the decision-making logic within the platform. Furthermore, the Oversight Board was intended to provide the self-regulatory framework of Meta with a measure of external legitimacy.

Despite the Oversight Board's organisational and financial separation from Meta, it remained dependent on Meta. Although Meta's Oversight Board functioned as an external supervisory body, it was still an extension of Meta's governance system. The Board was designed not to restructure the power dynamic of Meta but to mend fractures in user trust via procedural transparency and accountability mechanisms, closer to a form of self-correcting self-regulation. In other words, Meta's Oversight Board can be regarded as enhancing self-regulation by the social media platform itself.

Questions arise, such as could user trust be genuinely rebuilt by Meta's Oversight Board, or would it merely moderate content within a legal framework through a symbolic form

of ‘independent oversight’? This study primarily aims to examine the positioning of Meta’s Oversight Board in the platform governance system, the actual effect of this ‘quasi-independent mechanism’ in the power structure of digital platforms, and its role and limitations in rebuilding user trust. By analysing the governance transformation of Meta after its trust crisis, this study seeks to explore how digital platforms can strike a balance between societal pressure and institutional innovation and perhaps find a new way to moderate the content in contemporary platform governance.

1.2 Research Questions

In this era, trust faces a crisis as the actions of social media platforms fail to convince users to continue trusting them. Given the significant role that social media platforms such as Meta play in society, it is crucial to rebuild public trust to attract users for commercial interests and regain public interest for users. In addition, the declining trust towards social media platforms has given rise to social and economic problems, such as the drop in trust towards traditional institutions or the spread of misinformation among users. The declining trust towards social media platforms has also led to user churn.

In this vein, Meta’s Oversight Board was established in 2020 and operates independently of Meta. By decentralising decision-making powers, Meta intends to enhance accountability and transparency by involving multiple stakeholders, thereby rebuilding public trust. The Board’s members provide recommendations about contentious issues. Additionally, the Board helps to advance regulatory transparency by submitting an open report tri-monthly. When choosing a case study for reviewing, the

Board posts requests for public comment on its website and social networking sites. These published contents can be regarded as a way for civil society to participate in the decision-making process of the relevant case studies.

In addition, rebuilding public trust in digital platforms also requires understanding and adapting to the aspirations of the public, fully appreciating their significance and difficulty, recognising how they use these platforms, and empowering them to control occasionally unavoidable obstacles (Gillespie et al., 2020). Meta's Oversight Board stands out as a novel model of governance and a content moderation approach.

I will argue in this thesis that Meta's Oversight Board has not achieved the goal of rebuilding public trust in Meta. Establishing channels for receiving users' appeals, publishing case decisions and collaborating with non-governmental organisations are only part of the way to restore users' trust. More importantly, Meta's Oversight Board has limited influence among users. Apart from banning Trump's account, it was not a decision that led to much of a stir among users. The right of Meta's Oversight Board has limited access to the information of Meta, which means that its decisions on content moderation on Meta are not timely and effective enough. In light of this, it is urgent and necessary for governments of various countries to introduce relevant laws when dealing with false information on social media platforms. In other words, the establishment of Meta's Oversight Board is a good start in moderating content by tech companies themselves; however, the lack of a genuine and effective penalty system makes it difficult to achieve real change.

Trust is in crisis. According to Flew and Jiang (2021), in contrast to sociology, philosophy, economics and other fields, there have been relatively few attempts to systematise diverse contributions of communication scholars to propose a distinctive perspective on trust. The research questions addressed in this study are concerned with how Facebook loses user trust, what role Meta's Oversight Board plays in rebuilding public trust, and the implications of Meta's Oversight Board for moderating misinformation. These research questions are:

1) What is the background to the establishment of Meta's Oversight Board? What are the operations of Meta's Oversight Board?

2) In regard to misinformation, how do social media platforms moderate content and who should be responsible for regulating social media platforms?

Due to the failure of self-regulation by the platform, these issues have become particularly important in today's era. Regulating platforms at a global level raises questions relating to geopolitical considerations. In this vein, it is important to study the background of establishing the Oversight Board and put it in the political, economic, and social context in this age of misinformation. Meta's Oversight Board was established by a global operating social media platform company to resolve the geopolitical considerations in content moderation on a global scale.

1.3 Outline of Chapters

This chapter is the first of the five chapters that make up this thesis, which also includes the conclusion. Next, Chapter 2 constructs a theoretical framework covering digital

platforms and media governance. It begins with the term ‘digital platform’, which is categorised into two types: digital platforms and social media platforms. In addition, the features of digital platforms, such as data, algorithms, and platformisation, are also examined for the fundamental infrastructure of the digital era. In this digital age, media power has not disappeared but continues to exist and exert its influence on digital development. The abuse of media power necessitates regulation because of the unequal power dynamics between digital platforms and users. Due to the concentration of platform power, media governance has focused on platform governance by various stakeholders, which is caused by the abuse of media power in digital platforms and the data-driven economic models of digital platforms. This section starts from governance to regulation, and then moves to the platform governance triangle, which includes external, co-, and self-regulation. By moderating content on its platform as a third party. Meta’s Oversight Board fills the gap between traditional self-regulation and co-regulation**.**

Chapter 3 primarily addresses the regulatory journey of Facebook from 2010 to 2020, and covers the company’s streamlined and strengthened self-regulation regimes beyond its proposal to create Meta’s Oversight Board. Chapter 3 also examines governmental participation in Facebook regulation over the past decade of ineffective self-regulation in order to discuss institutions, power dynamics, discourse, policies, as well as economic frameworks. In terms of the economic background, this chapter presents the market value of Facebook from 2010 to 2020. In addition, the Australian Competition and Consumer Commission’s (ACCC) report and digital platforms is also examined

(ACCC, 2019). The ACCC report is used as the basis for the discussion about the external regulation over the past decade. In Chapter 3, a conclusion with the introduction of Meta's Oversight Board is drawn, aiming to address the decline in Facebook's brand value and the lack of legitimacy of Facebook's content regulation, in order to cope with the regulatory challenges of this social media giant. Chapter 3 demonstrates that self-regulation is ineffective in modern times and that government involvement is necessary for regulating content from social media sites. The chapter also provides a strong framework for Chapter 4, which argues that the government should participate in regulating false information.

Chapter 4 discusses how to regulate misinformation in the current era and who should be responsible for defining and moderating misinformation. In the first part of this chapter, the term 'misinformation' is defined. The discussion around the regulation of misinformation on a large scale, and reflects on and critiques the current regulation of misinformation on a large scale, is then made. In addition, future instructions for social media content moderation are also outlined. Chapter 4 argues that contemporary governments should be responsible for defining and moderating misinformation within the scope of the legal system. Accordingly, it is necessary to involve the national governments to rebuild trust on social media platforms in the age of misinformation. Since the freedom of expression among users may be hindered by one-size-fits-all regulation, it is essential for the national states to strike a balance between human rights and free speech.

Chapter 5 concludes this thesis and offers solutions to the questions raised in Chapter

1. This conclusion highlights that content moderation is a dynamic process rather than a static one, and that multiple stakeholders are needed to rebuild public trust and appropriately regulate content on social media platforms. A model such as Meta's Oversight Board is an excellent starting point for technology companies when it comes to increasing transparency. The government should also participate in the moderation of misinformation, and the public and non-governmental organisations are also an important part in the moderation of misinformation on social media platforms.

This thesis combines multidisciplinary approaches, including critical political economy, so as to examine platform regulation in rebuilding public trust in social media platforms due to the concentration of platform power. Semi-structured interviews and document analysis are used for data collection and analysis. Through the implementation of Meta's Oversight Board, which is a novel approach to content moderation on social media platforms, this project bridges the gap in the aspect of rebuilding public trust. Following a discussion of inadequate self-regulation from 2010 to 2020, this thesis explores the urgent challenges currently faced in content moderation, such as misinformation.

Chapter 2: Literature Review

2.1 Introduction

This chapter comprises two sections: digital platforms and platform regulation. This chapter first addresses the definition and characteristics of digital platforms, such as the rise of platformisation, as platformisation has given rise to multiple stakeholders on social media platforms. Different stakeholders exert their own power on social media platforms and stakeholders can be divided into two groups: the government; and other stakeholders, such as users, technology companies and non-governmental organisations. Governments can exert effects on social media platforms by being persuaded by technology companies, for instance, lobbying the government to introduce unveiling policies favourable to themselves, thereby influencing election results or triggering populism. In this vein, regulating social media platforms is about legitimacy. Other stakeholders can also exert effects on social media platforms through various means, including datafication and online business by designed social media platforms, such as issues triggered by platform-designed algorithms, like the speed of misinformation dissemination and issues related to human-machine algorithms. From this perspective, moderating content on social media platforms is about the scale, technological problems and socio-economic transformation under advanced technologies.

After the discussion of issues on social media platforms, this section addresses the framework of platform regulation. Questions about ‘who governs’ are raised due to the fact that multiple stakeholders can exert power over social media platforms. To

address these questions, the framework of platform governance is divided into two parties: the government, and other stakeholders. External governance relates to the government, and self-regulation, co-regulation and algorithmic regulation relate to other stakeholders. Although various regulatory methods for digital platforms may be partly effective, the content moderation on social media platforms differs somewhat from that of traditional media because it is difficult to regulate social media platforms by introducing national laws. More importantly, the content on social media platforms is massive, which also poses challenges for human reviewers or algorithms to moderate content. In this vein, rebuilding public trust on social media platforms requires addressing related issues within a legal framework and effectively moderating content on a large scale. Finally, this chapter identifies the research gap relating to the regulation of digital platforms like Meta's Oversight Board. The implementation of Meta's Oversight Board deals with issues within the legal framework by employing lawyers in their team and using technological tools to cope with issues on a large scale that arise from technological development. Studying Meta's Oversight Board provides significant insights relating to the regulation of platforms and the rebuilding of public trust in this digital age.

2.2 Digital Platforms

Digital platforms have been at the heart of business and management discourse for at least a decade (Evans et al., 2006; Parker et al., 2016; Andersson Schwarz, 2017). They have been studied from through the lenses of political economy (Flew, 2021a; Srnicek, 2017b), digital society (including public value and common good (van Dijck

et al., 2018)), content moderation (Gillespie, 2018), surveillance capitalism (Zuboff, 2019), and infrastructure for digital platforms (Helmond, 2015; Gerlitz & Helmond, 2013). This study focuses on content moderation, since platforms cannot function effectively if they do not moderate content (Gillespie, 2019).

This section undertakes a review of the literature on platforms, digital platforms, and social media platforms, and provides an overview of digital platforms, the expansion of digital platforms to platformisation, and the emerging digital society and economy arising from platformisation. This review highlights several issues relating to digital platforms that have resulted in the loss of user trust and in calls for improvements to the governance of social media platforms.

2.2.1 Platforms, Digital Platforms, and Social Media Platforms

Platforms

This sub-section focuses on research relating to digital technology platforms. The term 'platform' has a variety of definitions, including the perspectives offered by Gillespie, van Dijck, Andersson Schwarz, and Helmond, as shown in table 1 below:

Table 1: Perspectives on Platforms

Perspectives on Platforms		
From the perspective of content	The concept of 'platforms' transcends computational and architectural boundaries. It includes social, political, and cultural dimensions. That is, platforms enable communication, interaction	Gillespie, 2010

	and commerce over and above providing a space for writing or running codes.	
From a social perspective	Platforms organise social activities into standardised protocols or display these processes through user-friendly interfaces by combining hardware, software or services, including their design, codes and algorithms. The purpose of the platform design is to direct the attention of users in a particular manner. Fundamentally, a platform is a digital infrastructure based on software or hardware. It enables users to operate computer codes in the traditional sense (e.g. retrieving data or running applications (apps)) or to undertake a range of human uses (defined, formalised, and patterned by the design of the platform).	van Dijck, 2013 Andersson Schwarz, 2017
From the perspective of infrastructure	A platform is a system that users can reprogram and develop to fulfil different demands and niches. Nevertheless, the original developers might not have considered these evolving needs.	Helmond, 2015

These perspectives on platforms explicitly show that they facilitate user communication and the exchange of ideas and products. However, user preferences will determine how a platform evolves or develops, and some developments may not remain under the control of the original developers. The question arises: what benefits

does the design or unique feature of the platform offer to digital platforms in this regard? In the following section, 'Digital platforms and their classification', how to define the concept of digital platforms is addressed.

Digital platforms and their classification

Diverse interpretations of platforms might reveal the interest of academics in the subject matter and their differing points of view. The term 'digital platforms' is also interpreted in different ways. From a social perspective, van Dijck, Poell and de Waal explained that digital platforms are driven by data, organised and automated via interfaces and algorithms, formalised via ownership relations driven by a business model, and governed through user agreements (van Dijck et al., 2018). From an economic perspective, Srnicek argued that digital platforms are a new type of company. They stand out by facilitating the connection between a variety of user groups through their infrastructure, exhibiting monopoly tendencies influenced by network effects, utilising cross-subsidisation to attract different user groups, and possessing a predetermined core architecture that dictates interaction possibilities (Srnicek, 2017b).

In regard to content moderation, Gillespie (2018) suggested that it is required on digital platforms. Gillespie argues that, curated, organised, archived, and moderated content are necessary to ensure that all the information on social media platforms is true, thereby supporting a better user experience (Gillespie, 2018). This argument is based upon the need to respond to digital platforms issues that deprive users of their trust in platforms. In these respects, the public and academics have begun to recognise

the value of content moderation along with high-quality platforms.

Scholars also use the concepts of business models (Staub et al., 2021) and platform capitalism (Srnicek, 2017a) to classify digital platforms. Regarding the business model concept, Langley and Leyshon (2017) are followed based on the view of Flew. Table 2 summarises these different types of digital platforms.

Table 2: Different Types of Digital Platforms (A)

Types of digital platforms	Concept	Example
Online exchange markets	A marketplace where products and services can be purchased and sold via physical distribution, downloads and streaming. These products and services are purchased and sold typically at a reduced price compared to the prices quoted in traditional documents. Two-sided vendor platforms with open APIs, and multisided platforms, are incorporated to foster developer innovation. Multisided platforms are commonly closed APIs.	Amazon
Social media and user-generated content	Provide a place for user communities to publish content. Open and multifaceted APIs foster developer innovation.	Facebook
Sharing economy	Typically at a discount to those charged by traditional incumbents. Provides a marketplace for the rental of underutilised or unrecognised assets and services. A multisided market with closed APIs.	Uber
Crowdsourcing	A marketplace for the exchange of expertise, freelance and informal labour, and transactional and contractual work. A multisided market with open APIs.	Amazon Mechanical
Crowdfunding and peer-to-peer (P2P) lending	A marketplace where individuals can donate, pledge, lend, or invest funds at interest rates normally higher than those offered by conventional financial service providers.	Lending Club

	A multisided market with closed APIs.	

Source: Adapted from Langley and Leyshon, 2017, p.16

In contrast, and based on the argument that data extraction and use have taken centre stage in modern capitalism in the digital age, Srnicek (2017a) classified digital platforms into five categories: advertising; cloud; industrial, product; and lean. Table 3 summarises these definitions of digital platforms and provides examples.

Table 3: Different Types of Digital Platforms (B)

Types of digital platforms	**Concept**	**Example**
Advertising platforms	Through data analysis, these digital platforms collect user information, which is then used to produce products sold for advertising space.	Facebook
Cloud platforms	These cloud platforms rent out their own software and hardware for businesses who depend on digital technologies. They make money by renting out the basic infrastructure of the digital economy instead of selling data to advertisers. As a result, these cloud platforms gather data for their own internal use.	AWS
Industrial platforms	These industrial platforms offer the software and hardware required to move traditional manufacturing to online processes for the purpose of reducing production costs and transforming products into services.	GE
Product platforms	These product platforms earn profits by transforming a traditional product into a service via other platforms and charging rental or subscription fees on them. Due to their lower entry barriers, platforms are considered the ideal structure for extracting data and utilising that data to gain a competitive edge in the market.	Spotify
Lean platforms	These lean platforms attempt to reduce their ownership of assets to a minimum and profit by lowering costs as much as they can. These outcomes are also achieved by their notorious outsourcing of workers. However, lean platforms are derived from tendencies and moments, including increasing unemployment, a surplus population, the digital age, etc.	Uber

Source: Adapted from Srnicek, 2017a, p.36

When classifying digital platforms, research tends to take an economic perspective. However, given that these changing ways of accumulating wealth can also cause social problems the section below examines the social, political and economic effects of social media platforms as a subset of digital platforms.

2.2.2 Social Media Platforms

Over the last decade, social media has come to permeate every aspect of daily life with significant sociological, political, and economic implications (McCay-Peet & Quan-Haase, 2016). Social media platforms, like Facebook, are the agoras of today: the most important places for the exchange of views (Riemer & Peter, 2021; Everett, 2018). However, social media platforms exhibit their own characteristics and they are different from digital platforms.

Definition of social media platforms and their differences from digital platforms

Gillespie (2018, p:18–19) defines social media platforms as 'online sites and services that:

a) host, organise and circulate the shared content or social interactions of users for them,

b) without having generated or commissioned (the bulk of) that content,

c) built on an infrastructure, beneath that information circulation, for processing data for advertising, profit and customer service.'

Social media platforms are at the centre of a dynamic environment with continuous social and economic changes. They create a connecting area for information and

communication (van Dijck, 2013). Users are ‘aggregated’ through uploading, promoting, and advertising user-generated content on various social media platforms. Through comments, social referrals, search rankings, and page visits, these user activities produce relevant data on the engagement and retention of users (Nieborg & Poell, 2018). In these respects, social media platforms as an infrastructure model offer a technological framework upon which others can construct, with a view to facilitating connections and growth on the sites, apps and data of others. Concurrently, the economic models of social media platforms are predicated on the preparation of external data for their databases (Helmond, 2015).

This section has provided a literature review of the definitions of platforms, digital platforms, and social media platforms. Facebook is the research focus of this thesis. The key innovations in this aspect revolve around platforms utilising data, algorithms, and machine learning, to better understand their users, and then on-selling this information to advertisers so that they can more precisely target their campaigns than would be possible without that information (Flew, 2021a). In this regard, section 2.2.3 reviews the literature on data, algorithms, and platformisation.

2.2.3 Data, Algorithms and Platformisation

Technology companies use data, algorithms and machine learning to cater to the requirements of their users, to push targeted advertising to earn money (Flew, 2021a), and to stimulate the activity and engagement of users to increase user stickiness. In this context, data and algorithms are the foundations of digital platforms.

Data and algorithms

The competitive advantage of the digital economy is grounded in data ownership (Srnicek, 2017b). Data is critical to multiple stakeholders and technology companies need to rapidly develop new methods for collecting and analysing it (Srnicek, 2017b). Furthermore, analysing and extracting user data is seen as an effective method by which social media platforms can further develop their current products and services (Atal, 2021). Data ownership can also be expanded through the acquisition of other companies in the era of digital platforms.

Another result of the broadening scope of data collection and analysis (across all social networks, and including those involved with news and politics) is the monitoring, capture, and exploitation of every communication between people and machines (Smyrnaios & Baisnée, 2023). The digitisation process has also triggered a societal transformation, especially in regard to the possession and use of global economic power (Ulbricht & Yeung, 2022).

Algorithms have become a crucial feature of people's participation in public life (Gillespie, 2014). They play an important role in selecting what information is considered most relevant to people. Gillespie held that algorithmic organisations have become common in a media environment dominated by platforms. Consequently, users are now reliant on algorithms in the same ways that they relied on experts in the past, despite their knowing little about the mechanisms underlying algorithmic decisions. During data collection, algorithms translate the tastes and preferences of users into relational databases that, in turn, inform user behaviour (van Dijck, 2013). However, these uses of various algorithms and interface features to curate content and

steer user activities remain opaque to the users (van Dijck et al., 2018).

Martens (2016) argues that we have witnessed 'the loss of individual autonomy, transparency about the workings of these algorithms and accountability of algorithm operators' due to the vast amounts of data collected, the algorithms used, and the influence of these algorithms on human behaviour and decision-making (p. 36). In particular, more extreme and outrageous content, like fake news, tends to receive more engagement (quoted by Riemer & Peter, 2021, p. 413). User-generated content on digital platforms is soaring. Instead of being connected by community ties, people are linked by social networks and recommendation algorithms. As pointed out by Gillespie (2020), the consequences of online harms increasingly exceed the platforms where they occur. The issues on digital platforms are complicated, involving not only technological issues but also social issues, even though machine learning offers a variety of mathematical approaches for solving a wide range of issues, such as moderating misinformation and hate speech (Taherdoost, 2023).

Platformisation

APIs and data flows between multiple platforms are regarded as relevant factors for the expansion of digital platforms, as open APIs and data play an important role in the development of digital platforms. 'Platformisation' is an emerging term used by researchers to refer to the consequence of the development of digital platforms across a variety of business and social realms. An API is a form of platform architecture that allows third parties to program a website using a software interface. The API is accessed "from outside the core system", which means that "the app code of the

developer lives outside the platform" (Helmond, 2015, p. 5). Moreover, data flows and pours not only build a bridge between digital platforms and third parties, they also work as data channels to make external web data platforms ready.

Scholars understand the term 'platformisation' via the infrastructure and political economy perspectives. From the infrastructure perspective, 'platformisation' refers the rise of platforms as the predominant infrastructural and economic model of the social web and the expansion of social media platforms into other spaces online (Helmond, 2015, p. 5). From the perspective of political economy, "platformisation encompasses the outwards extension of a platform into other websites, platforms and apps, and its inwards extension, with third-party integrations operating within the boundaries of the core platform" (Nieborg & Helmond, 2019, p. 202).

In regard to how the political economy of cultural industries changes via platformisation, "platformisation can mean that the economic, governmental and infrastructural extensions of digital platforms penetrate web and app ecosystems, which exerts a fundamental influence on the operations of cultural industries" (Nieborg & Poell, 2018, p. 4276). From this perspective, platformisation transforms simple single and two-sided markets into complicated multi-sided ones. The economic position of cultural producers is profoundly affected by this shift, as it brings novel economic mechanisms and managerial strategies.

The 'digital society' (van Dijck et al., 2018) and 'digital economy' (Srnicek, 2017a) have come to be dominated by data, algorithms, and infrastructures. Data assets are

integrated due to the emergence of platformisation (Andersson Schwarz, 2017). Importantly, the issue of media concentration becomes complex. Digital platforms have multiple roles to play owing to the formation of platformisation. Apart from hosting and curating media content, they also run their own business functions, including advertising networks, identity services, data intermediaries, social networking (van Dijck et al., 2019). In these respects, platformisation poses challenges for the regulation of multi-user digital platforms.

2.2.4 Conclusion

The issues associated with digital platforms have been enlarged and expanded by the development of digital platforms, including platformisation. Platformisation, which enables data to flow and be shared across different digital platforms, has exerted effects on the formation of large-scale, winner-take-all business models, posing challenges for the governance of social media platforms. This is because social media platforms have their own terms and conditions for content moderation, but self-regulation in content moderation has not been successful. Since social media platforms operate globally, national governments find it difficult to introduce laws to regulate social media platforms. Given the definition of digital platforms, the differences between digital platforms and social media platforms should be taken into account, as they have different profit models and operation methods. This study focuses on social media platforms. Social media platforms provide users with a way to post information on them and open up the path for users to share information in the

public space without geopolitical and temporal limitations. Nevertheless, although the benefits of social media platforms for allowing the public to share information online may be obvious and even seem utopian at times, the perils of social media platforms are also painfully apparent (Gillespie, 2019). The emergence of platformisation allows many different stakeholders to access and use social media. This state of affairs raises questions that are central to disputes concerning public value creation in the platform society: whose interests do the platform's activities serve; which values are at stake; and who benefits?

Social media platforms exhibit many characteristics, including connectivity, data, algorithms, and machine learning. Hence, multiple stakeholders also employ different methods to deal with controversial issues on social media platforms, including: introducing laws about the regulation of data in the European Union (EU); adopting a face-checking system in Japan; proposing data sovereignty; and establishing Meta's Oversight Board. If regulating social media platforms is necessary, it is crucial to learn about their characteristics, including connectivity, how information is distributed and spread on social media platforms, data, algorithms, machine learning systems, and AI-related problems. To understand and govern social media platforms, it is effective to propose targeted, implementable and rational solutions. In this regard, section 2.3 presents a review of the literature on platform regulation.

2.3 Platform Regulation by Multiple Stakeholders

Technology companies have established rules for how users interact with one another and designed-platforms tools, such as recommending algorithms that decide what

information is accessible to the general public. In other words, technology companies play a vital role in moderating content on social media platforms. The term ‘content moderation’ means systematically checking messages and making decisions on what speech to allow and what to block (Gillespie, 2018). Content moderation can be an opaque and secretive process (Gillespie, 2018; Suzor, 2019; Gorwa et al., 2020) that lacks public accountability and transparency. The moderation of user-generated content goes beyond content removal, to include removing content and suspending the accounts of users. Reducing the visibility of problematic content is also a common way of governing digital platforms (Gillespie, 2022). Nonetheless, moderating content on digital platforms is difficult. When the public puts pressure on technology companies to engage in moderating content, technology companies respond by recruiting human reviewers to moderate content (Gillespie, 2019).

In the digital age, setting rules for the regulation of digital platforms also means that whoever makes the rules can make them in their favour. Multiple stakeholders on digital platforms serve different interests by proposing and prioritising different methods of regulating digital platforms. Under the circumstances, the regulation of platforms goes beyond “who decides who decides” (Zuboff, 2019, p. 231) to “who sets rules and what rules will be” (Haggart, 2020, p. 322).

Currently, the social media platform regulation takes several forms, including: self-regulation; external regulation; co-regulation; and algorithmic regulation. This literature review explores the differences between governance and regulation, as different definitions embody different priorities. Different methods of moderating

content, what the stakeholders stand for, and whose interests they serve, are also reviewed.

2.3.1 From Governance to Regulation

Multiple disciplines have studied governance. This thesis focuses on media governance, particularly the governance of social media platforms. This section distinguishes between the terms governance and regulation. Since multiple stakeholders on social media platforms serve different interests, the formulation of regulatory rules for social media platforms has also changed and become more complicated.

Freedman (2008) thought that media governance means the full range of formal and informal, national and supranational regulations, and centralised and dispersed mechanisms aiming to organise media systems. As an analytical notion, media governance helps define, clarify and criticise media policy and governance (Puppis, 2010). The definition of media regulation is different from that of media governance. Media regulation refers to the responsibilities assumed by governments (Puppis, 2008). The digital society primarily depends on digital platforms for its economic, political, educational, community, health, and other activities. Therefore, the desire for reliable, secure, and effective communication platforms has significantly increased (Lunt, 2012). On account of this, the changing media industry over the past decade has also been illustrated by its conversion from media regulation to media governance (Puppis, 2008).

From a conceptual perspective, the emphasis on a process is a major distinction between regulation and governance (Hofmann et al., 2017). In a digital society, distinctions between rule-makers and rule-takers have been obscured by the transformation of regulation (Hofmann et al., 2017). This blurring of distinctions has captured the interest of scholars investigating the differences between regulation and governance. Some scholars argue that governance is best comprehended as regulation (Feick & Werle, 2010). However, other scholars indicate that governance is used more commonly than regulation (Flew, 2021a). The concept of Internet regulation initially emerged in the mid-1990s and, with the evolution of technology and society, the definition of Internet regulation has greatly expanded (Hofmann et al., 2017).

Julia Black asserted that regulation is a dynamic cooperative approach to problem-solving. It involves an effort to influence the conduct of others and extends beyond the government (Black, 2005). Based on this opinion, the implications of regulation exert a greater impact on individual behaviours. While governance is more widespread in the age of digital platforms, regulation suits the context of the Internet. Before these debates over the definition of governance and regulation, digital platforms, the main way to disseminate information in the digital era, received a great deal of attention in regards to regulation (Gillespie, 2018). Nowadays, technologies owned by technology companies are challenging the rules of data flow management by platform-designed tools (McAfee, 2009). In this context, the public is overwhelmingly convinced that regulation has the potential to work as a flexible approach to managing markets. However, individuals express less agreement with the

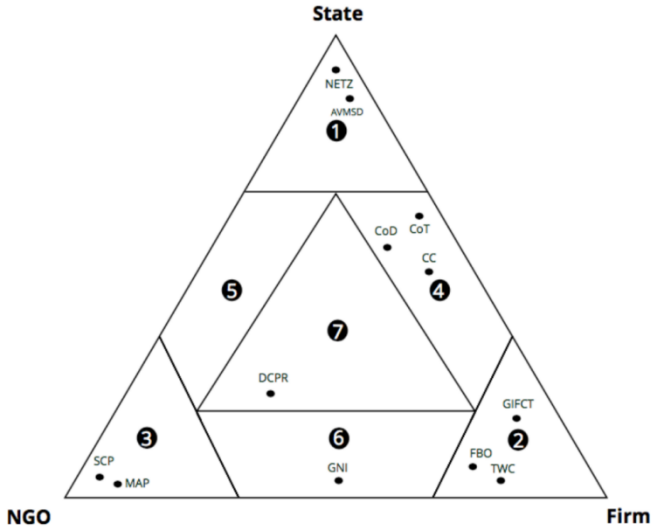
application of regulatory agencies to social and cultural issues (Lunt & Livingstone, 2012). In terms of governance, Mayntz (2004) argued that governance can be defined to regulate the entire system, including all community rules in society.

2.3.2 Platform Governance Triangle

Media regulation has captured the interest of communication scholars. Content moderation on social media networks has emerged as a disputed issue across the public and research agendas (Medzini, 2022; Puppis, 2010). Currently, in the platform ecosystem, 'regulating platforms', such as moderating misinformation, are viewed as a system of social, political and economic relationships used to regulate the interactions among key stakeholders, such as users, technology firms, and governments (Gorwa, 2019a). In light of this, different rules or types of regulation proposed by different stakeholders represent different interests of these groups.

Gorwa developed the concept of a platform governance triangle by expanding on the theory of the governance triangle (Abbott & Snidal, 2009). This concept contributes to our understanding of how governments, civil society, and digital platforms are interrelated. Governments, companies and non-governmental organisations, and civil society organisations are distributed across seven categories in the platform triangle. The actors responsible for regulating platforms will transfer from the platform itself to relevant stakeholders, including governments (introducing related policies) and civil society (including non-governmental organisations).

Figure 1: Platform Governance Triangle



Source: Gorwa, 2019b, p. 7

In the platform triangle, the research topic, Meta's Oversight Board, is located in the self-regulatory section. From section 6, it can be seen that the platform triangle tries to regulate digital platforms by seeking to incorporate civil society, despite its intention to prevent the state from being an actor in decision-making in sections 1 and 4 of the platform triangle.

In the original platform governance model proposed by Abbott and Snidal, the distribution of the three angles is more 'balanced'. However, when it comes to the platform governance triangle, platform participants are vastly more powerful than the state or civil society in most cases. The overall visual metaphor is still useful, but its implied rough equivalence of power simply does not exist in reality. Given this, at this stage, content moderation becomes a problem because social media platforms are willing to share their method of moderating content with other stakeholders, such as national governments or non-governmental organisations. Meta's Oversight Board

was established by Meta, which is the initial way for social media platforms to explore a new way to moderate content with a third party. It has broken the monopoly of a single platform, such as Meta, in the regulatory model of content moderation.

2.3.3 Current Types of Regulation

Different types of regulation are adapted in platform regulation. In the framework of the platform governance triangle, governance or regulation has three governance frameworks: self-regulation; external regulation; and co-regulation (Gorwa, 2019b). In this triangle, the self-regulation implemented by technology companies stands for the interests of technology companies. External regulation represents the interests of the government, which sets rules for public interest and value. Co-regulation opens the way for negotiating space among multiple stakeholders, including the government, NGOs, and the technology companies. Furthermore, platformisation and the scale of user-generated content on social media platforms have brought new challenges to platform regulation, since moderating content at a large scale requires human moderators. Consequently, platform regulation also requires technological implementation of large-scale moderation of content. This type of regulation is known as ‘algorithmic regulation’. Next, this literature review explores each types of regulation rather than arguing which ones are good or better.

Self-regulation

Self-regulation can be defined from private and industrial perspectives. For example, Medzini (2022) suggested a new concept of enhanced self-regulation based on his

discussion of self-regulation from private and industrial perspectives. However, Coglianese and Mendelson's (2010) interpretation of self-regulation is used for reference: any regulatory system with a regulatory target, at the individual or corporate level, or occasionally by an industry group identifying targets, implies orders and penalties for themselves. This definition can also apply to the regulation of social media: part of the question is definitely how platforms are governed, and a question of equal importance is how platforms govern (van Dijck, 2013).

In 2010, public shocks led to criticisms of social media platforms, including the flow of misinformation and toxic content. In response to public shocks, the public interest rationale for self-regulation to meet the requirement of self-regulation would be a more affordable way of rising to challenges than traditional approaches (Ogus, 1998). Digital platforms took action and called for regulation to provide users with a better online environment. In 2018, Mark Zuckerberg acknowledged these calls for action, noting that these platforms have a bearing on every aspect of the public's lives, and that "the question was not whether regulation should be adopted, but what the right form of regulation should be" (Zuckerberg and the Senate Commerce, Science, and Transportation Committee, 2018).

Regarding the self-regulation of technology companies, Facebook's content moderation policies have been shaped by two opposing forces: a centralising force from dominant platform policies like the community standards of Facebook, and a splintering force from political circumstances and national speech/election laws. These forces have led to a two-tiered approach to moderating content across the

examined jurisdictions (Ahn et al., 2022). A consensus was reached that the government regulation of Internet platforms was threatening (Bossio et al., 2022) because governments prioritised the needs of the public above the technology companies' financial interests. Beyond that, the expansion of digital companies worldwide offered an opportunity for self-regulation. The dilemma of regulating digital platforms at the national level contributed to self-regulation by the digital media sector (Cherubini & Nielsen, 2016). Self-regulation would be the method of adjusting policies across different cultures due to the global nature of social media platforms and the state-level implementation of legislation for digital platforms.

External regulation

External regulation is a method by which the government introduces laws or rules to regulate misinformation. This action has a huge impact on firms' and industries' operational strategies and profitability (Chin et al., 2022). While self-regulation may be a promising approach for regulating digital platforms, self-regulation has been criticised from a public interest perspective (Flew & Gillett, 2021). Proponents of external regulation felt that government engagement would increase the transparency of information and rebuild the trust of the public in social media platforms.

A number of countries, including China, have implemented an authoritarian internet, whereby monitoring and identification technology are claimed to benefit social cohesiveness and security through their tracking of crimes, deviance, and other issues (O'Hara, 2018). For example, China constructed an electronic 'Great Firewall' to

protect its ideology (Watts, 2006). Additionally, the designation “techno-nationalist model” has been used for maintaining the economic independence of China (Wang & Gray, 2022, p. 82). Foreign companies need to comply with licensing requirements to operate their own businesses in China, as well as with limitations on foreign capital and ownership and technological Internet censorship (Mueller & Farhat, 2022). These Chinese government policies were implemented to protect the ideology and economy of China.

The EU has imposed similar regulations. The EU Cyber Agenda established digital sovereignty as an essential goal with the objective of constructing a free, secure, and resilient cyberspace (Calderaro & Blumfelde, 2022). The concept of ‘digital sovereignty’ refers to the EU attempts to cover for the deficiency that has arisen over the past decade as a result of inadequate innovation in the operation of software and hardware progress (Kaloudis, 2022, p. 275). To resolve these problems, digital sovereignty was proposed, which required member states to agree to comply with EU legislation. However, no consensus has been reached among these member states, which causes problems with non-compliance (Bellanova et al., 2022).

The reason for the differences in regulatory systems between China and the EU lies in their distinct political ideologies. Regarding China, the implementation of an authoritarian internet means that content moderation is a top-down model. Nevertheless, this model is difficult to apply in other countries, including the EU, as the EU places a stronger focus upon democratic rights of citizens. Due to the strong influence of social media platforms and the difficulty of reaching consensus at the

national level, non-compliance is not useful when facing content moderation on social media platforms. Social media platforms operate on a global scale, but the inconsistencies in regulatory policies across different countries have made the standardisation of content moderation extremely challenging. This is the core issue addressed in this thesis.

Co-regulation

Gorwa has argued that "Steps towards 'co-governance' look for a third way between the two former approaches" (2019a, p. 864). This way generally involves non-governmental organisations and civil society organisations in some form of multi-stakeholder governance arrangement (Flew, 2021a). Unlike top-down command and rule-making or industrial self-regulation, co-regulation develops a way for social media networks to moderate content and related issues. It is a way whose boundaries call for more clarity (Napoli, 2015). To take one example, the General Data Protection Regulation regulates the data practices of platforms in the EU (Popiel & Sang, 2021).

Due to unavoidable government involvement in governing social media, a new system of institutions that allows others to participate in forms of regulation may be critical to its ongoing operation and prosperity (Cusumano et al., 2021). This collaborative approach is seen as a bridge that brings together multiple stakeholders and balances their divergent viewpoints. For instance, Meta's Oversight Board privately attempted to establish and circumscribe outside oversight on particular issues regarding content moderation by appointing members from the civil society and academic communities

(Arun, 2020). Popiel and Sang stated that “the initiative makes the line between co-governance and self-regulation blurry and raises key questions about how these classes of arrangements will work in practice” (2021, p. 3). It is industrial self-regulation. As mentioned previously, self-regulation is a better (and cheaper) approach to addressing problems than conventional regulation from the government. Nevertheless, it also calls for suggestions from the Oversight Board formed by members from various organisations.

Scholars define self-regulation and co-regulation separately. However, Meta’s Oversight Board blurred the boundaries between traditional self-regulation and co-regulation. This practice of moderating content on Meta fills the research gap in studying the regulation of social media platforms and their implementation on other social media platforms.

Algorithmic regulation: Turning technology for moderation at scale

With the continuously expanding scale of user-generated content on social media platforms, companies and legislators intend that technical tools can deal with issues of content moderation (Gorwa et al., 2020). In 2017, the term ‘algorithm regulation’ proposed by Yeung (2018), refers to decision-making systems that regulate an activity field and are leveraged to manage risks or evolution directions through the preservation of computational knowledge. These systems achieve these objectives by logically gathering data directly emitted from various dynamic components associated with the governed context to recognise and, if necessary, self-refine the operation of

the system to meet pre-designed goals.

Gritsenko and Wood (2022) studied the regulation of three policy challenges (speeding, misinformation and social sharing) to demonstrate what happens when algorithms enable collaboration in forms of external regulation, self-regulation, and co-regulation. The authors found that the application of algorithms meant that the space for controlling the directions of actors decreased, but efficiency of moderating content was improved. These groups and politicians have both been digging for technical ways to cope with issues on digital platforms, like misinformation and harmful content, as a consequence of the increased pressure from the government on technology companies. Social media networks, such as Facebook, YouTube and Twitter, are progressively increasing their use of algorithmic content moderation systems, which are automated hash-matching and machine learning methods (Gorwa et al., 2020).

For the moment, algorithms are a part of public life, but the debate about their effects remains in its early stages (Gritsenko & Wood, 2022). Algorithmic governance was categorised by Just and Latzer as software as institutional governance, at least conceptually enables the notion that algorithms could be integrated into market, hierarchical and platform modes of regulation (2017). In the view of users, algorithmic audiencing, which refers to users' speech chosen by algorithms, is interfering with the right to free speech in a way that was not previously possible. Algorithmic audiencing will enhance or manage speech on social media for the purpose of making profits, which in turn negatively affects fair and open

communication in public discourse. The fundamental issue is that black-box algorithms select audiences who are recipients of what is shared. As a consequence, the free speech issue shifts from what can be said to what will be heard. To moderate content on digital platforms, the issues must be addressed from the perspective of audiences (Riemer & Peter, 2021).

The role of algorithms in moderating content on social media is of increasing importance. In the online world, all social communication is fully collected as data, which has since been rendered economically exploitable and predictable by algorithms. In this context, algorithmic governance is paramount to the operation of public and private institutions (Schuilenburg & Peeters, 2021): "algorithms are a highly economic governance tool because of reducing dependence on human deliberation and hence governance costs" (Schwarz, 2019, p. 127). However, in the social context of algorithmic governance, Issar and Aneesh highlight three general areas in which the social negotiability of processes is menaced, namely the problems of power (surveillance), discrimination (social bias), and identification (system identity) (2022, p. 1). From this perspective, algorithmic regulation is not enough because some complex events still require human intervention and decision-making.

Trust

Trust is a key factor in human interaction. Trust has been studied in sociology, psychology, management, marketing, and other disciplines (Das & Teng, 2004; McKnight & Chervany, 2002; Kim & Lee, 2020; Flew & Jiang, 2021) despite the

difficulty in defining it conceptually. Due to declining trust in social media platforms, especially in relation to the spreading of toxic content, like misinformation and fake news, trust and technology have become major topics of discussion in the context of technology and media policy. In this regard, Sztompka argued that "without trust, people would be paralysed and unable to act" (1999, p. 8) and Luhmann noted that "one would even be prevented from getting up in the morning if trust is completely absent" (1979, p. 4). Furthermore, the spotlight placed on trust indicates that sociology has not become insensitive to huge social challenges or given up discovering truths also relevant to society (Sztompka, 1999).

In the digital age, trust as a glue holding society together only shifts instead of disappearing (Botsman, 2017). Botsman (2017) classified trust into three components: local, institutional, and distributed trust. Local and institutional trust function well in the older social context, while distributed trust is developed in the context of rising technology. The undergoing social context means rewriting the rules of human relationships. Under the circumstances, the problems of accountability are becoming complex. In addition, it is difficult to find responsible stakeholders as they offer services without possessing any assets or hiring any providers. Traditional regulation, like regulating assets or putting pressure on responsible parties, loses the possibility of implementation. In light of this, the key point would lie in proposing other rules to rebuild trust. The main similarity among multiple stakeholders on digital platforms lies in that the public expects platforms to minimise the occurrence of negative events, lessen uncertainty, and provide them with assistance in the event that negative events

occur (Botsman, 2017). This is only the key point for the expectations of the public. The other thing would be how to meet the requirement and what factors play a role in rebuilding public trust. It is necessary to realise and fulfil the trust hierarchy of needs: identity, security, safety, compatibility, and belonging (Botsman, 2017).

Studies of trust in the digital age that have focused on the discipline of communication, and communication have covered a broad range of areas, including trust in journalism, institutions, advertising, and traditional media (Flew & Jiang, 2021; Kim & Lee, 2020; Botsman, 2017). Studying trust in communication would involve a combination of political, economic, technological, and social effects, as well as issues arising from these effects. This view was also proposed by Muir (1994) and developed by Söllner. Sollner proposed the idea of a trust network, particularly for recognising and evaluating the various prevalent relationships of trust while studying complex technological systems (Söllner et al., 2016). Meanwhile, academics, policymakers and the general public are becoming increasingly concerned about whether the arsenal of national and supranational regulatory tools, such as antitrust, competition and privacy laws, is flexible enough to address the digital dominance of technology companies (Moore & Tambini, 2018). Given this, this project would be narrowed down and further developed into two parts: first, the political aspect; and second, the factors of rebuilding public trust in the age of digital platforms. From this perspective, this research intends to explore ‘who governs’: specifically, in relation to moderating content on social media platforms, who sets the rules and what interests they serve? As van Dijck et al noted, “besides the ‘how’ of regulation, ‘who’ is

responsible" should also be addressed (2019, p. 10).

2.3.4 Conclusion

To conclude, different stakeholders exert different power and pursue different interests on digital platforms. The power relationships between, and the conflicts among, stakeholders result in different types of regulation. This section has expounded the differences between regulation and governance, followed by four kinds of regulation. Despite setting rules for different interests and serving different groups, different stakeholders have the same purpose of moderating content in this technological environment, including moderating content at scale and earning the trust of the public by implementing different methods. This thesis discusses content moderation that different stakeholders pursue to align with legitimacy when making decisions about conspiracy issues, which is what Meta's Oversight Board is working for.

2.4 Methodology

This section provides a summary on research methodology of this study. In view of the research questions summarised in Chapter 1, the research methodology quoted theoretical frameworks of digital platforms as well as regulations set for these platforms. Specifically, this study utilised a qualitative research approach, mainly based on semi-structured interviews and document analysis, which can gather data effectively. Both methods provided empirical data for this study. The semi-structured interviews were conducted with academics, industry experts, and members of non-

governmental organisations. Different perspectives were delivered to make up for the contextual information that the case texts lack.

2.4.1 Document analysis

In general, document analysis is applied in conjunction with other approaches of qualitative research (Bowen, 2009). Because of the power relations between multiple stakeholders and the presence of various stakeholders in a single platform, this method prevails. Document analysis of news articles was conducted on Meta's official website for a deep understanding of the regulatory history of Facebook from 2010 to 2020, which delivers the underlying governance logic and power dynamics. Document analysis aims at investigating the regulatory history of the platform in addition to the unequal power dynamics between the platform Facebook and its users between 2010 and 2020. Following research on its regulatory practices, this study focuses on concerns brought by contemporary social media platforms, including misinformation.

Power dynamics among stakeholders can be studied through document analysis, which helps to deal with the research concerns about how Facebook users' trust declined from 2010 to 2020. In total, 820 news articles concerning Meta were collected from the official website into an Excel spreadsheet. Through data connection during the study period, 40 events were selected to examine the regulatory history of Facebook from 2010 to 2020 based on 232 articles selected from the 820.

The selection criteria for the 820 news articles on Meta's official website are news

articles focusing on Facebook. However, after collecting these 820 news articles, this was a large amount of information for document analysis. In order to narrow down the news articles, news articles about the development of Facebook, especially those related to content moderation, were chosen. After screening, the number of news articles was reduced to 232. The 40 events were selected based on important developments of Facebook and significant events faced by Facebook from 2010 to 2020.

Correspondingly, 40 events demonstrated how Facebook evolved and the causes of public trust decline. The regulatory history of Facebook was reviewed with the ACCC report (2019) as a case study on government regulation from 2010 to 2020. In addition, Mark Zuckerberg's public speeches were also studied to provide further insights into the regulation methods taken by Facebook between 2010 and 2020.

2.4.2 Semi-structured Interviews

A semi-structured interview is defined to obtain descriptions of the life world of a participant with respect to interpretation about the meaning of the phenomena involved (Kvale, 2008). Qualitative research interviews mainly function to reveal the lived reality of participants, understand the world from their perspectives, and highlight the importance of their experiences (Kvale & Brinkmann, 2015). While this method is not applied in a generalised manner, it still shows powerful potential. Semi-structured interviews are also structured enough to discuss specific dimensions of research questions, with space left for study participants to express new meanings centred on the topic of this study (Galletta, 2013). Participants are encouraged to

propose comments about questions during semi-structured interviews in addition to open-ended discussions. Participants' responses are kept confidential and used as first-hand information.

The semi-structured interviews adopted in this study managed to collect information from staff in non-governmental organisations and other stakeholders, with participants from industry, academia and non-governmental organisations who were invited to provide a diverse range of data targeted at the research questions. In view of the huge number of stakeholders for social media platforms, the standards to select participants were based on the concept of a platform governance triangle. See Appendix I for details of the interviewees.

After the identification of potential participants, potential interview questions were designed according to theoretical frameworks, digital platforms, and platform governance, with 12 interview questions set, including two for Meta Oversight Board participants and one for civil society participants. For the purpose of getting more information from the participants, interview questions were structured in a way that developed from open-ended to closed questions.

2.4.3 Ethics Approval

The interviews were conducted under the guidelines of the Human Ethics Committee of the University of Sydney (Project No. 2023/660) with approval from the committee in early 2024. Participants provided written permission (a consent form), with which the researcher was authorised to use their data in the study. These concentrated first-

hand data are set under the premise that respondents' responses are grounded in the given questions, all of which concern the subjects of this research and theoretical frameworks.

2.4.4 Semi-structured interviews and document analysis

Online and offline methods were combined to carry out the fieldwork. Online interviews were conducted via the Zoom platform, which managed to facilitate the communication, thereby ensuring that audio, video and text records were integrated to fully retain detailed information during subsequent processing. Face-to-face interviews were recorded by Otter.ai software, which was transcribed into text quickly, enhancing the accuracy and traceability of the data.

With regard to the topic of study, seven participants were interviewed from August to December 2024. Each interview took 45 to 90 minutes. This interview data supplemented the case studies and literature reviews, thereby enriching the researcher's understanding of, and reflections on, the research questions involved in this study.

To recruit the interviewees, the email was sent in mid-August 2024. Initially, two interviewees showed interest in the project and replied promptly to the email. However, there was no reply after sending three emails to three researchers. Therefore, under supervision, I revised the email content and sent it to other researchers. Finally, the rest of the researchers replied to the emails. It is easy to find the email addresses of the researchers; however, when recruiting the non-

governmental organisation members, I found it difficult to locate their email addresses. So I left the message on the website, and a member of a non-governmental organisation replied to my message via email. In addition, I had sought to interview industry participants, including Meta representatives, but did not get responses to my requests in time.

After completion of the fieldwork, the researcher organised and analysed the data. All interview recordings were transcribed into written form. After the transcription work, open and thematic coding were conducted on the data collected between August and September 2025. Regarding the theoretical framework, the thematic analysis took into account two dimensions: digital platforms and media regulation, whereby delivering thematic coding directional guidance, which can deliver insight from the analysis of responses from different participants.

As for the theoretical framework, the thematic analysis of the present study strictly abides by two major dimensions: media governance and digital platforms. This framework offers thematic coding directional guidance and serves as a reference point for the analysis of responses from different interviewees. Under the dimension of ‘digital platforms’, for instance, the focus is especially on how interviewees depicted the role that the platforms played in algorithmic recommendation, information distribution, as well as the construction of discourse flows. The analysis is focused on the inherent features of platforms, the non-transparency of algorithms and the angles of artificial intelligence. ‘Media governance’ is further divided into external regulation, self-regulation and co-regulation, so as to examine the advantages,

disadvantages and limitations of these models according to the interviewees' responses.

The thematic analysis in this study mainly follows the theoretical framework, but a few derivative themes extending beyond the framework appeared during the particular coding process. These themes usually stemmed from the interviewees' references to current affairs, particular figures or specific cases. For instance, multiple interviewees frequently mentioned the connection between Trump and the US presidential election. This reveals the complicated interplay between digital platforms and political figures, disclosing the intricate entanglement of political propaganda, misinformation and platform governance in reality. These findings prompted the incentivised researchers to maintain the precision of theoretical frameworks while embracing openness and flexibility, which accommodate new issues emerging from data.

Furthermore, during the thematic analysis process, priority was given to a strategy of cross-case comparison. By comparing the responses of different groups horizontally, consensus themes can be identified, and areas of conflict can be highlighted. For instance, academics are likely to emphasise the opacity of algorithms and their influence on contemporary users, whereas civil society representatives pay closer attention to the necessity of legal safeguards and external supervision. Academic representatives may be prone to highlighting the responsibility of platform algorithms in the dissemination of information, while civil society representatives emphasise the significance of algorithmic transparency and accountability mechanisms. Regarding these differences, they are considerably valuable at the analytical level. This is

because they showcase the different positions among multiple stakeholders and empirically verify subsequent discussions on platform governance models.

In addition to semi-structured interviews, document analysis was utilised to further the identification of tensions between the official and public discourse relating to platforms. The data collected through document analysis is mainly based on the analysis in Chapter 4, which focuses on the rhetorical strategies of Meta within policy documents, press releases, and official statements. The rhetorical strategy adopted by Meta in dealing with public events was that Mark Zuckerberg and the official website published official statements to respond to "public outcry". In the article of "Suspending Cambridge Analytica and SCL Group from Facebook", Paul Grewal (VP & Deputy General Counsel) responded that "Protecting people's information is at the heart of everything we do, and we require the same from people who operate apps on Facebook" (Grewal, 2018). However, in fact, there was no significant improvement in content moderation.

In addition, Chapter 4 draws on the ACCC report to illustrate the imbalance of power between social media platforms, such as Facebook and Google, and their users. The concentration of media power in platforms like Facebook has created significant regulatory concerns, especially regarding user privacy, platform transparency, and the responsibility of social media companies to moderate content. After two decades of largely ineffective self-regulation, there is a strong case for greater government involvement in regulating social media platforms, including through legislative measures such as Australia's News Media Bargaining Code. In light of this, the

ACCC report could show why Meta's Oversight Board was established in the lacuna of communication policy. This chapter illustrates Meta's construction mechanism of governance legitimacy via discourse, and reaction means towards external policy pressures and social opinions.

However, the acquisition of data is limited in certain aspects. Meta did not provide any internal documents concerning this study. For this reason, the research process relied on publicly available sources, such as media coverage, policy statements, annual reports, and public rulings from supervisory committees. While this sets limits to insight on what the company has been thinking internally about the Oversight Board, external statements provide sufficient empirical evidence to draw valid conclusions.

This section has outlined the research methodology and research design for this study, with a focus on how this approach maintained scientific rigour and objectivity under complicated research subjects and limited data availability. Through a combination of document analysis and semi-structured interviews, this study established a logical chain of cross-validation between different analytical pathways and data sources.

2.5 Conclusion

This study is concerned with governance questions arising in relation to digital content hosted on proprietary platforms (i.e. platforms possessed and managed by companies). These questions loop back to societal issues like declining trust in news and information, which in turn provide a reference point as to whether different

governance strategies work. Connected to this is the issue of who undertakes regulation: governments, companies themselves or 'third parties' like Meta's Oversight Board? Based on previous studies, this study focuses on 'who governs'. Thus, the two parties would be the governments and other parties. In light of this, studying Meta's Oversight Board can help explore how to handle issues on social media platforms within a legal framework on a large scale and who will set rules in this digital age.

Chapter 3: Platform Power, Platform Policy, and Declining Trust: The Development of Facebook's Regulation

3.1 Introduction

This chapter explains the rationale behind the establishment of Meta's Oversight Board by analysing the development of Facebook's regulation, including the platform's rising power, declining trust in the platform, and the platform policy in Australia. First, the chapter presents an introduction. Second, the chapter outlines the history of Facebook's regulation between 2010 and 2020 by classifying and discussing the key events during the period, drawing on the framework created by Flew (2021a), including the economic framework, the policy framework, the cultural framework, and the digital framework. Last, the chapter explores the reasons behind a breach of trust in Facebook with its evolution and content moderation, and analyses Meta's Oversight Board as a method for moderating content on Facebook.

With the number of Facebook users surging from 400 million worldwide in 2010 to 2.6 billion in 2020, Facebook has considerable influence over public discourse and in the public sphere. The rising number of users with the evolution of technology for collecting users' data by human-designed infrastructure, and user-generated content on Facebook, raised questions regarding the regulation of toxic content posted by users (public discourse), and regarding how for-profit data-driven companies regulate human conversation (public sphere). Imbalanced power relationships, shaped by human-

designed algorithms and terms of services on Facebook with Mark Zuckerberg's public statements, between users and Facebook resulted in ineffective self-regulation on Facebook during the period.

For Facebook, trust is crucial to the preparedness of users to engage with the social media platform. However, the platform has been consistently criticised for disregarding user privacy and other breaches of trust (Flew, 2022). Policy-makers and multiple stakeholders have also paid closer attention to the trust issues on Facebook due to its ineffective self-regulation. During the period when criticism of Facebook's regulation increased, governments introduced external regulation through policies requiring the moderation of content on Facebook. Or the government collaborating with technology companies through the method of co-regulation.

From rising platform power and self-regulation implemented by Facebook to declining trust or breaches of trust, and external regulation by governments, different stakeholders play their roles in shaping debates about how to better moderate content on Facebook, with introducing the emergence of the Meta Oversight Board as a form of quasi-self-regulation by a semi-independent group of experts. This chapter addresses three critical questions regarding the regulatory history of Facebook:

1) How did events involving Facebook and Mark Zuckerberg's public statements influence the self-regulation of Facebook between 2010 and 2020?

2) How was the government involved in regulating content on Facebook between 2010 and 2020?

3) How did the regulation evolve between 2010 and 2020 behind the establishment of Meta's Oversight Board, and its role in content moderation?

These three questions aim to address the background to the establishment of Meta's Oversight Board, the traditional self-regulation has been ineffective, and the national states have not yet found the right way to regulate social media platforms. Therefore, Meta's Oversight Board was established under this context. After that, this chapter will elaborate on the actual operations of Meta's Oversight Board. To address the rationale behind the establishment of Meta's Oversight Board, this chapter draws on the framework created by Flew (2021a), including the economic framework, the policy framework, the cultural framework, and the digital framework. This chapter focuses on the economic and policy frameworks to analyse the context for establishing Meta's Oversight Board. By exploring power relationships between Facebook and users as well as co-regulation between the government and Facebook, such as the Australia Competition and Consumer Commission report (2019) in Australia, this chapter fills the gap in communication policy for regulation because Meta's Oversight Board emerged in a lacuna for policy. To be more specific, from the self-regulation by Facebook to co-regulation between Facebook and the national state, the current digital society or digital economy, including the platform ecology or related regulation, is the result of the interaction of the four factors (economic, policy, cultural and digital frameworks) under the framework.

3.2 A Brief History of Facebook's Regulation between 2010 and 2020

Between 2010 and 2020, the dominant "hands off" agenda of global politics about platforms was challenged from a variety of perspectives, with massive growth in digital platform companies (Flew, 2022, p. 308). To understand where Facebook is today, politically and socially, it is necessary to understand how the decisions made in the past still linger in the system, which is also referred to as the path dependency of institutions (Bucher, 2021). This section introduces Facebook's regulation between 2010 and 2020. Created in 2004, Facebook briefly stated its mission to be empowering people to share information (Napoli, 2015). To protect free speech, social media platforms like Facebook have notoriously been granted a "safe harbor" approach to regulation, which exempts them from liability for the content they host because they are not regarded as content creators or editors (Gillespie, 2018). This approach led to very limited content moderation on Facebook as it disrupted the balance between expressing ideas and human rights on the social media platforms. Although these platforms have gradually implemented their respective self-regulatory mechanisms to address issues such as misinformation, they are not regulated like other media, including print, broadcast, or digital news platforms, for which there are mature and effective regulation models (Siapera, 2022).

Based on the work of Helmond, Nieborg, and Van Der Vlist (2019), this chapter divides self-regulation on Facebook into four stages: the initial regime (2004–2009); the thin self-regulation regime (2009–2012); the strengthened self-regulation regime (2012–

2018) (Helmond et al., 2019); and the proposal for the establishment of Meta's Oversight Board (2018-2020). This section focuses on the last three stages.

Table 4: Classifying Facebook Events Between 2010 and 2020

Timeline		Event
1	April 9, 2012	Facebook to Acquire Instagram (Meta, 2012a)
2	May 17, 2012	Facebook Announces Pricing of Initial Public Offering (Calif, 2012)
3	August 2, 2012	Introducing Facebook Stories (Meta, 2012b)
4	October 4, 2012	One Billion People on Facebook (Zuckerberg, 2012)
5	July 22, 2013	Feature Phone Milestone: Facebook for Every Phone Reaches 100 Million (Makavy, 2013)
6	August 21, 2013	Technology Leaders Launch Partnership to Make Internet Access Available to All (Meta, 2013)
7	August 27, 2013	Global Government Requests Report (Stretch, 2013)
8	February 19, 2014	Facebook to Acquire WhatsApp (Meta, 2014a)
9	March 7, 2014	Announcing the Public Content Solutions Program (Morgan, 2014)
10	March 25, 2014	Facebook to Acquire Oculus (Meta, 2014b)
11	April 24, 2014	Announcing FB Newswire, Powered by Storyful (Mitchell, 2014)
12	November 3, 2015	New Milestones in Artificial Intelligence Research (Schroepfer, 2015)
13	October 21, 2016 Napalm girl	Input From Community and Partners On Our Community Standards (Kaplan & Osofsky, 2016)
14	April 6, 2017	Working to Stop Misinformation and False News (Mosseri，2017)
15	June 26, 2017	Facebook, Microsoft, Twitter and YouTube Announce Formation of the Global Internet Forum to Counter Terrorism (Meta, 2017a)
16	September 6, 2017	An Update On Information Operations On Facebook (Stamos, 2017a)
17	September 21, 2017	Facebook to Provide Congress With Ads Linked to Internet Research Agency (Stretch, 2017)
18	October 2, 2017	Improving Enforcement and Transparency of Ads on Facebook (Kaplan, 2017)
19	October 3, 2017	Promoting October Cyber Security Awareness Month

		(Stamos, 2017b)
20	November 22, 2017	Continuing Transparency on Russian Activity (Meta, 2017b)
21	December 4, 2017	Update on the Global Internet Forum to Counter Terrorism (Meta, 2017c)
22	December 8, 2017	Sharing Facebook's Policy on Sexual Harassment (Sandberg & Goler, 2017)
23	December 18, 2017	Reinforcing Our Commitment to Transparency (Sonderby, 2017)
24	March 21, 2018	Cracking Down on Platform Abuse (Meta, 2018a)
25	April 4, 2018	An Update on Our Plans to Restrict Data Access on Facebook (Schroepfer, 2018)
26	April 9, 2018	Facebook Launches New Initiative to Help Scholars Assess Social Media's Impact on Elections (Schrage & Ginsberg, 2018)
27	May 15, 2018	Reinforcing Our Commitment to Transparency (Sonderby, 2018)
28	May 23, 2018	Facing Facts: Facebook's Fight Against Misinformation (Hegeman, 2018)
29	November 15, 2018	A Blueprint for Content Governance and Enforcement (Zuckerberg, 2018)
30	June 30, 2019	A Second Update on Our Civil Rights Audit (Sandberg, 2019)
31	October 14, 2019	European Court Ruling Raises Questions About Policing Speech (Bickert, 2019)
32	December 5, 2019	Taking Action Against Ad Fraud (Romero, 2019)
33	December 12, 2019	Ready for California's New Privacy Law (Meta, 2019b)
34	December 12, 2019	An Update on Building a Global Oversight Board (Harris, 2019a)
35	December 17, 2019	Helping Fact-Checkers Identify False Claims Faster (Silverman, 2019b)
36	December 20, 2019	Removing Coordinated Inauthentic Behavior From Georgia, Vietnam and the US (Gleicher, 2019)
37	January 6, 2020	Enforcing Against Manipulated Media (Bickert, 2020)
38	January 15, 2020	Facebook Disaster Maps Help Those Affected by Australia's Bushfires (Meta, 2020a)
39	August 20, 2020	Facebook Files Official Comments on Data Portability with Federal Trade Commission (Meta, 2020b)
40	August 31, 2020	An Update About Changes to Facebook's Services in

		Australia (Easton, 2020)

3.2.1 Thin Self-regulation Regime (2010–2012)

In his 2012 annual report, Mark Zuckerberg concluded that "we connected one billion people, we transformed our products and business to be primarily focused on mobile, and we transitioned to being a public company" (Facebook, 2012, p.1). When identifying the profits of acquisitions, including the acquisition of Instagram in 2012, Facebook put its comments in the "risks factors" section, which elaborated on the plan to make more acquisitions despite potential challenges, such as the demand for more significant management efforts, the disruption of business, the dilution of stockholder value, and the adverse impact of acquisitions on financial results. At this stage, external regulation by the government is regarded as an intervention in Facebook's innovation and development.

However, modern mega-corporations have disproportionately concentrated economic and political power in the hands of people who do not stand accountable for anyone's loss (Zingales, 2017). In particular, platforms operate in winner-take-all markets due to characteristics like economies of scale, network effects, the accumulation of big data, switching costs, and lock-in properties (Popiel, 2020). In this context, entrepreneurs might engage in lobbying and corruption, so as to gain a key first-mover advantage and retain their power (Zingales, 2017).

Against the backdrop of laws of "safe harbor" and a loose regulatory environment, the

self-regulation of Facebook was limited due to the lack of policies for content moderation. Furthermore, the strategy of Facebook for connecting people aimed to cater to users' preferences to encourage their engagement, and the technology was not advanced enough to design and guide users' behaviours.

At this stage, due to the lack of regulatory awareness and weak regulatory measures, social media companies were willing to attract more users to engage in social media platforms. As Participant 5 noted:

> I think digital platforms were a solution to the to a problem of trust. The people felt when interacting online or over the Internet, or even they're interacting to say, book a place to stay right? That there was a trust issue and a social deficit ……And I think digital platforms in a way emerged to address that. But I think also create new kind of social issues in trust, too.

Furthermore, regarding the operation of content moderation at scale, participants also pointed out:

> And I think one of the things that emerged in moderation and in the Internet, in the 1990s and 2000s was the people who ran the email lists, or the people who were called the Sysops. The system operators. Left out of the. policy. But actually they were doing a lot of the work right. So I think you

need those people. And then you need community organizations or other groups as well……. Scale. I mean, that's the problem.

The content on social media platforms is massive, making it difficult for tech giants to remove harmful content. Due to the huge amount of information on the internet, especially on social media platforms, content moderators do not have the energy to moderate all this information. Therefore, moderating content moves to the next stage, from 2012 to 2018.

3.2.2 Strengthened Self-regulation Regime (2012–2018)

Between 2012 and 2018, Facebook strengthened its self-regulation due to the changing communication environment, including technological advancements, widespread 'fake news', and the public's criticism of data misuse, including the function of "trending topics". Social media platforms, such as Facebook, Twitter, and YouTube, were increasingly attracting controversy, especially after the 2016 election, which naturally motivated them to pay more attention to the public interest. In other words, they began to see themselves as emerging media platforms, who would be held responsible for the global public good (Napoli, 2019). After 2016, these platforms undertook multiple initiatives, ranging from the adoption of new advertising tools to the adjustment of their interactions with political campaigns, which appeared to hinder regulatory actions while effectively preserving their highly profitable business models (Gorwa, 2019a). On the other hand, the emergence of the platform society raised challenges, mainly due to the detachment of social processes from conventional (often nationally constrained)

regulatory frameworks (Nash et al., 2017). It was only after reports of foreign interference emerged during the 2016 US presidential election, and the subsequent scandal involving the political consulting firm Cambridge Analytica's harvesting of 87 million Facebook users' data, that platform companies began to undergo concentrated political scrutiny. This situation has sparked an increasingly strong public opposition, and critics have raised key concerns about the business model of digital platforms: the large-scale spread of disinformation; immense unregulated data collection subject to breaches and non-consensual sharing with third parties; siphoning profits from the already struggling news sector transitioning to digital advertising; the sociopolitical consequences of inherently flawed content moderation at scale; and growing concerns about market power (Popiel, 2020). In light of this, the failure of content moderation requires new regulations to address the power issues of the platforms.

At this stage, Facebook continued to expand its business, including acquiring WhatsApp and Oculus. As a result, social issues became more complex on Facebook, and the platform was criticised for the selection of its "trending topics" by humans who had potential political orientations rather than by an algorithm (Gillespie, 2016). In response, Facebook announced improvements to its trending functions (Cathcart, 2017). In addition, the transparency of algorithmic operations resulting from human-designed technology faced public criticism (Meta, 2017b; Kaplan, 2017; Sonderby, 2017), and the debate about content moderation has become increasingly intense (Sandberg & Goler, 2017).

Some the criticism of Facebook's content regulation arose from public shocks, where an incident generated enough criticism of a platform to gain significant visibility, led to controversies about the basic operations or effects of a platform, and prompted platform owners to act differently (Ananny & Gillespie, 2016). In other words, the platforms have faced strong criticisms in terms of their impact on the public, from individual users, the media, and the broader public. Such criticism ranges in scope, from strongly worded user complaints all the way to public outcries that dominate news cycles. Public outcries followed public shocks. Public outcries could be tackled effectively through crisis communication (Ananny & Gillespie, 2016). To deal with the public backlash and meet Facebook's expectations, Mark Zuckerberg, leveraging his position as the founder of Facebook, attempted to wrest back control of public discourse about events on Facebook. Rhetoric here is Mark Zuckerberg's call for external regulation for Facebook. In response to the US Congress and Hearing issue of 2018, Mark Zuckerberg said, "the real question, as the Internet becomes more important in people's lives, is what is the right regulation, not whether there should be or not" (quoted in Flew, 2021a, p.150). It was not until 2018 when the Cambridge data breach events raised concerns about content moderation on Facebook that Mark Zuckerberg admitted Facebook's call for the right regulation. In other words, Facebook's self-regulation cannot meet the demand of moderating content on its platform at this stage. Nevertheless, as pointed out by Kate Klonick (2018), Facebook has already become the de facto governor of online speech, which requires additional and far more powerful institutional accountability mechanisms.

In light of this, Facebook's regulation lagged behind the public's requirements. Although these platforms benefit from extensive legal exemptions by protecting the policy of safe harbour, scholars and politicians have expressed concerns about the absence of democratic accountability and the risks of private censorship under their self-regulation (Helmond et al., 2019). Moreover, articulating the algorithm as a distinctly technical intervention was beneficial for information providers' effective response to bias, mistakes, and manipulation. The political landscape was being reshaped by a new category of information power, which was concentrated in these huge databases of user activity and preference (Gillespie, 2014). For this reason, self-regulation opened its new chapter in content moderation within a legal framework and called for government involvement from 2018 to 2020.

3.2.3 Proposal for the Establishment of Meta's Oversight Board (2018-2020)

The document analysis highlighted platforms' efforts to maintain a nuanced balance between the need for information shareability and the imperative to control some types of information by conducting content dissemination controls under the guise of security and developing a sophisticated mechanism of control on this basis (Siapera, 2022). Although the outcomes of these efforts are still unknown, they suggest Zuckerberg's belief that many stakeholders in self-governance no longer see self-regulation as a feasible long-term solution. Many called for an increasing government role in the wake of numerous public relations scandals, widespread privacy breaches, and mounting

concerns about polarisation and false information (Gorwa, 2019b). After the Cambridge Analytica Scandal in 2018, calls for the government's involvement in content moderation on Facebook were a belated recognition of the necessity of regulation of some sort by Zuckerberg, and the proposal that the Oversight Board be set up as an alternative to government regulation.

During the COVID-19 pandemic, 46% of consumers trusted premium news sites as a preferred source of information over social media platforms (PR Newswire, 2020). And 47% of consumers said that they would make a future purchase from an ad found on a premium news site, compared to 38% on social media (PR Newswire, 2020). In order to regain public trust and address data privacy issues on Facebook, Facebook took systemic measures, including responding to platform abuse (Meta, 2018a), implementing restricted data access on the platform (Schroepfer, 2018), improving transparency (Sonderby, 2018), and fighting misinformation (Hegeman, 2018). The most important and meaningful proposition was to establish an independent Oversight Board for Meta (Zuckerberg, 2018).

Platforms attempt to incorporate external community support into their governance procedures to establish legitimacy and trust. In response, Facebook founded what Mark Zuckerberg termed a "supreme court", allowing external appeals of content policy rulings (Zuckerberg, 2018; Gorwa, 2019b). It appeared that there was an inherent tension between government regulation of public communication and implied rights to freedom of speech. At the same time, there were growing concerns about social media

platforms' capacity to behave in the public interest given their profit-driven business models (Helberger et al., 2018). Moreover, after twenty years of limited regulatory oversight, the digital platform market was now under increasing scrutiny worldwide (Popiel, 2022), and there were growing demands for policy intervention. In this context, self-regulation within a quasi-legal framework, such as Meta's Oversight Board, appeared to represent a new way to adhere a balance between external regulation and self-regulation.

By analysing the events and Mark Zuckerberg's public statements on Facebook, this section shows the power imbalance between Facebook and users due to Facebook's dominance in the design and control of the public discourse. The prominent transformation during the period was from "we are not a media company, we are a tech company" (Segreti, 2016) to "what is the right regulation, not whether there should be or not" (Zuckerberg, 2018). However, power relationships among the multiple stakeholders were complex and they were influenced by the economic, political, cultural, and technological context. Using Flew's framework, the next section discusses how economic and political factors changed the social context and explores the underlying reasons for the evolution of Facebook's self-regulation.

3.3 Discourse, Institutions, and Power Relations

The regulatory issues relating to Facebook and the related policies implemented by Facebook did not occur in a vacuum; they arose from a complex and rapidly evolving economic, political, social, and technological landscape.

This section uses Flew's framework (2021a) to discuss the type of market power transformed from media business to platforms, especially Facebook, and the transforming market platform power and how it can be shaped and reshaped by media policy. By examining discourse use within institutional contexts, this study revealed the intricate connections between economic power and social structures.

In this section, the relationship between discourse and institutions can be interpreted as follows:

1) Discursive practices: Institutions are manifested through their discursive practices, including speech.

2) Institutionalisation of discourse: Institutions codify and institutionalise certain discourses such as legal texts. In turn, these discourses legitimise and reinforce institutional norms.

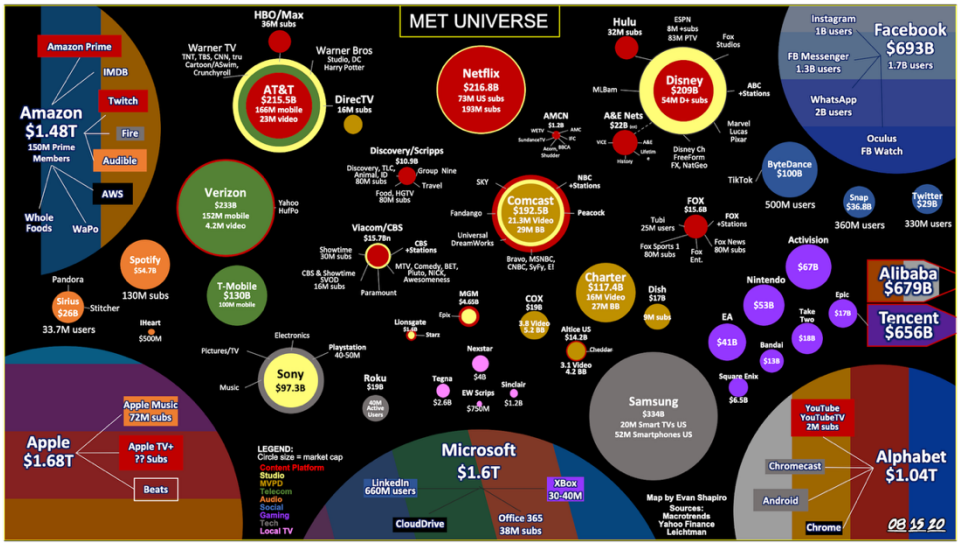
3) Discourse as a tool of power: Institutions use discourse strategically to maintain authority, control narratives, and influence public opinion. However, critical analysis reveals underlying power dynamics and the role of discourse in sustaining or challenging institutions.

3.3.1 The Economic Framework

Within the economic framework, institutions with economic power can shape discourse to align with their interests, including the policies of social media companies. In this section, economic power prevails over user power. In other words, media users exercise

and operate power on social media platforms, such as Facebook, and the evolving media landscape reflects this power dynamic from 2010 to 2020. In 2010, traditional media businesses dominated the market and had the largest user base. By 2020, the world's most powerful media companies had transformed from traditional media businesses to digital platforms. In 2020, the world's largest media businesses included traditional organisations such as Disney ($209 billion), Samsung ($334 billion), and Sony ($97.3 billion) (Shapiro, 2023).

Figure 2: Media Universe Map 2020



Source: https://eshap.substack.com/p/media-universe-maps-2020-2023

However, the evolution of technology and the changing digital environment resulted in a different brand ranking by 2024: big technology companies came to dominate (Shapiro, 2023). By 2024, the market value of Disney had declined to $167 billion, while the market value of Sony and Samsung remained unchanged. In contrast, technology companies' market values had rapidly increased. Apple's market value had

increased from $1.68 trillion to $2.85 trillion, and Microsoft's had increased from $1.6 trillion to $2.8 trillion.

This section addresses the question of how changing media shaped the discourse given their interests in earning profits, including from media power to platform power, data-driven business, and platformisation. Technology companies implement corporate policies when developing technologies, including data collection and an open API, so as to form the platform, thereby reshaping the economic environment and media environment, so as to align with their interests. To understand the operation of media businesses in the age of digital platforms, this section analyses the transformation of media power into platform power and the economic environment in which the businesses operated, including data-driven business and platformisation.

Media power and economic power are interlinked and mutually reinforcing. Consequently, understanding the economic environment is essential to analysing the operation of media businesses (Flew, 2018a). In the age of digital platforms, news organisations around the world relied on digital platforms aligned with their business models, distribution infrastructures, and production practices, with the increasing number of users on digital platforms (Poell et al., 2023). In this context, platform power has fundamentally reshaped the media environment and, by extension, influenced how the public gets news, how news is produced, how politics works, and how the public connects with one another (Nielsen & Ganter, 2022). In the 21st century media environment, from traditional and internet environments to digital platforms, especially

Facebook, surprisingly, their competitors are not news organisations but other platforms. In light of this, previous media power has transformed into platform power, changing the public's information and news-related habits and behaviours. In other words, the ability to exercise citizenship, which includes socio-technical, political, and economic mechanisms for the distribution of information and communication resources (such as the capacity to frame current events and to make one's ideas visible to others), relies on the political economy of the media (Smyrnaios & Baisnée, 2023).

In this vein, data-driven companies and digital platforms have established themselves as new intermediaries in information-goods markets that were already oligopolistic. The digital economy has experienced a progressive shift in market power along the value chain, from product and content producers to service providers and distributors (Nuccio & Guerzoni, 2019). In Australia, the ACCC concluded that a data-driven business on digital platforms – particularly the production and distribution of news on digital platforms outside the legal framework of regulation – was at odds with public-interest journalism (ACCC, 2019). Moreover, advertising on digital platforms has evolved into a highly complex and interconnected global ecosystem with technological development and practices driven by automated systems and applications of data and analytics (Van Der Vlist & Helmond, 2021). Platforms operate in a global winner-take-all market, in which there is a massive concentration of platform power, which can be abused. Consequently, there is a requirement for the regulation of platform power (Evens & Donders, 2020). A winner-take-all market could be aligned with information complementarities. The value of the data derived

from Facebook and Instagram together (i.e., combined and compared) is likely to be higher than the sum of the values of the data derived from Facebook and Instagram separately. If you add market power effects, the momentum towards concentration might be irresistible (Zingales, 2017). By prioritising profits over the public interest, Facebook has experienced declining user engagement and increased political and regulatory scrutiny of its business practices in some markets (Nielsen & Ganter, 2022). In light of this, with organisations potentially adopting different data-driven campaigning practices, it is important to ask which forms of data use are seen to be democratically acceptable or problematic.

Critical political economists have observed that the power dynamics among complementors, end users, and platform operators are extremely volatile and intrinsically unbalanced, because operators bear sole responsibility for the techno-economic evolution of a platform (Hurni et al., 2022). This situation arises because globally running platform companies began to increase their role in both public and private life, transforming the economic environment and domains in society, including the news. This transformation can be understood as a process of 'platformisation' (Poell et al., 2019). These businesses benefit from substantial economies of scale and, especially, scope of operations as a result of platformisation through the use of open APIs among platforms, which allows them to take advantage of massive information assets (Mansell, 2015). Furthermore, with the rise of a platform-and-data-driven ecosystem (van Dijck et al., 2019), a mix of traditional currencies (attention and capital) and contemporary ones (data and users) became essential to creating a

common set of platform rules to govern digital interactions and transactions.

Digital platforms' services pose new policy challenges because their scale and ubiquity mean that they are increasingly integrated into the political, social, and economic areas of public life (Popiel, 2020). In other words, Facebook's significant economic power means that it can shape discourse to align with its interests. Discourse, in turn, influences institutions, such as Facebook, by calling for the introduction of policies that align with their interests. The next section discusses the increasing role of government in platform regulation, using the ACCC report (2019) as a case study to understand the relationship between institutions and discourse.

3.3.2 The Policy Framework

The power of digital platforms with their problem-solving abilities enables the potential for co-regulation between national governments and global technology companies in the public sphere (Flew, 2022). As proposed by Meta's CEO Mark Zuckerberg (2019), there should be 'a more active role for governments and regulators'. In this context, by conducting a case study using the ACCC report and analysing the deal between Facebook and media businesses in Australia, I argue that the growing role of governments in regulating platforms can be regarded as a shift towards co-regulation, which also arise from the process of policy implementation and policy silence.

The regulation of digital platforms has returned to governments after two decades of inadequate self-regulation, which failed to address public concerns (Flew & Wilding, 2021). From this perspective, communication policy is intended to address the exercise

of power on social media platforms as a constant issue, including both market and political power. Regulatory frameworks are designed to guarantee that, regardless of the power that digital platforms may possess, their use is consistent with the interests of citizens and consumers (Mansell, 2015). To achieve these outcomes, regulatory action has represented a response to the scale and dominance of Facebook (Andrews, 2019). To address power concentration and imbalance, expanding the scope of legal frameworks to encompass the sociotechnical and political-economic connections in which these frameworks are rooted is necessary (van Dijck et al., 2019). Furthermore, the increasing public enquires provided evidence for a transition from corporate to state governance of platforms (Flew & Gillett, 2021). In this regard, the EU has introduced regulations such as the General Data Protection Regulation and the Digital Service Act. Similarly, the Australian government engaged in an enquiry, via the ACCC, with the ACCC's final report released in 2019. Due to the failure of self-regulation, there appeared to be an emerging consensus about the shift from self-regulation to co-regulation among businesses, policymakers, and civil society players (Stockmann, 2023).

The concept of co-regulation has long been present in regulatory theory. Co-regulation is based on the idea that, while regulators are responsible for the formulation of the general rules and laws, the industry manages the application of these regulations under governmental and parliamentary supervision (Flew, 2018b). Co-regulatory agreements between states and platforms provide a solution for regulating platforms that overcomes the challenges posed by the jurisdictional issues arising from platforms, as well as the

amount and size of information disseminated across different platforms (Flew, 2018b; Popiel & Sang, 2021). First, co-regulatory scope and scale are beneficial to nation-states because they allow them to deal with a small number of large corporations rather than external regulation when addressing issues such as cybersecurity. Second, structural segregation and strong economic rules might be more desirable than, or precede any, co-regulatory arrangements, thereby making co-regulatory parties, that is, policymakers and platforms, more dependent on one another (Popiel & Sang, 2021).

Supervised by the Australian Communication and Media Authority (ACMA) under a co-regulatory framework (ACCC, 2019), the Australian Mandatory News Media Bargaining Code, which is designed to deal with imbalances in commercial relationships between digital platforms (Facebook and Google) and news companies, positions the government as a mediator, to facilitate negotiations between digital platforms and media businesses (Flew et al., 2021). Following the failure of initial efforts to encourage the relevant industries to create a voluntary code, in 2020 the Australian Government urged the ACCC to formulate an obligatory code of conduct, which would address evident power imbalances by requiring platforms, including Google and Facebook, to pay for content. The code permits an independent arbitration panel to decide payments in the event that a digital platform fails to reach an agreement with a news organisation (Bossio et al., 2022). Since then, the introduced Australian Mandatory News Media Bargaining Code by the Australian Federal government have begun to contribute to agreements between Facebook and media businesses.

Co-regulation of Facebook also resulted from policy implementation for technological reasons, including regulation scale adjustments for content moderation. Focusing on the three connected forces – technological, economic and power relations – that propel the implementation process, a political economy model of implementation was created, with key components, such as policy-making, the selection of policy instruments, the identification of critical actors, driving forces, the service delivery system and the policy output (Hasenfeld & Brock, 1991). The phenomenon of platformisation enables more stakeholders to exercise their power on digital platforms. As a result, the proliferation of 'stakeholders' is perhaps more closely linked to the growth of the media industry than it is to their capacity to alter the power balance in a decision-making situation (Freedman, 2013). In light of this, any risk of excessive influence from private interest groups would be mitigated by the transparency and accessibility of the policymaking process, which would enhance the stability of the system (Freedman, 2010). Voices and perspectives engage in a transparent, vigorous, and non-discriminatory dialogue to come to a consensus about policies that prioritise the interests of the majority over those of the few (Freedman, 2013). The media's increasingly active role as highly powerful policy participants further challenges the idea that contemporary media policymakers embody the pluralist concept of a competitive and dynamic bargaining arena (Freedman, 2013). Digital platforms such as Facebook and Google influence the policy-making process because these technology companies take a central role like the conduit, the pipeline and the platform for communicating and exchanging information, as well as actors in moderating content,

particularly in the digital era as I mentioned earlier. Platforms, therefore, represent a fundamentally different information configuration from a material, institutional, financial, and social standpoint (Gillespie, 2019).

Policy silence does not equate to 'doing nothing' and 'negative policy', and it does not imply that policymakers are inert or unwilling to intervene. Rather, policy silence alludes to a strategic choice that the state's specific role as a policy maker is the most effective means of promoting hegemonic interests and legitimising fundamental beliefs (Freedman, 2010). In this sense, regulation should be viewed as a process rather than an event. The cycle begins with the identification of public issues, which may prompt corrective action, and then moves through various regulatory and governance proposals, as well as their approval and implementation by the courts or regulators (Andrews, 2019). After publishing the final ACCC report, Nick Clegg, the then Vice President of Global Affairs and Communications for Facebook, stated that, 'while we have concerns with some of the final recommendations, we welcome the ACCC's report and are fully committed to engaging in the consultation process while ensuring that we can continue to deliver the benefits of technology to the millions of Australians who use Facebook's services' (2019). However, a Meta spokesperson declined to comment on the latest news about the bargaining code on Meta. This month the company defended its decision, stating that people were not using Facebook for news content and that it should not be the responsibility of global technology companies to solve the issues plaguing news media (Taylor, 2024). Meta did this in Canada in July 2023, in response to a similar government regulation whereby Canada's news ban is still going (Purtill, 2024).

In conclusion, this analysis of co-regulation of Facebook in Australia, using the ACCC report as a case study, has clarified the rationale for the regulatory measures, the process of policy implementation, especially the technological factors when implementing the policy, and the concept of policy silence. Regarding the regulation of Facebook, it should be negotiated with other stakeholders on social media platforms and be applicable when implementing the regulation. The recommendations of the Australian Competition and Consumer Commission (ACCC) for regulating market power and Facebook's withdrawal of its news feeds in Australia provide interesting research cases for studying the regulation of social media platforms. The case indicates that tech companies have the power to resist the platform policy introduced by the national states, as they can choose to withdraw part of their business in one country. Given this, it raises the question of the difficulty in moderating content on social media platforms. Although the trust in social media platforms has obviously declined, the ineffective platform policy and the lack of cooperation from social media platform companies have led to worse results in regaining public trust in the current era. The analysis of the ACCC report reveals whether Facebook's self-regulation within a legal framework, such as Meta's Oversight Board, provides obvious and effective case studies when external stakeholders' collaboration with Facebook ends with Facebook's withdrawal from the local business market.

3.4 Meta's Oversight Board

This section follows a three-part structure, commencing with the issue of declining trust in Facebook, then exploring the reasons for this decline (including the policy's

silence and the lawlessness prevalent in social media, including Facebook), then concluding with the introduction of Meta's Oversight Board, the regulation of Facebook within a legal framework, and the challenges encountered in operating Meta's Oversight Board.

3.4.1 Declining Trust in Facebook

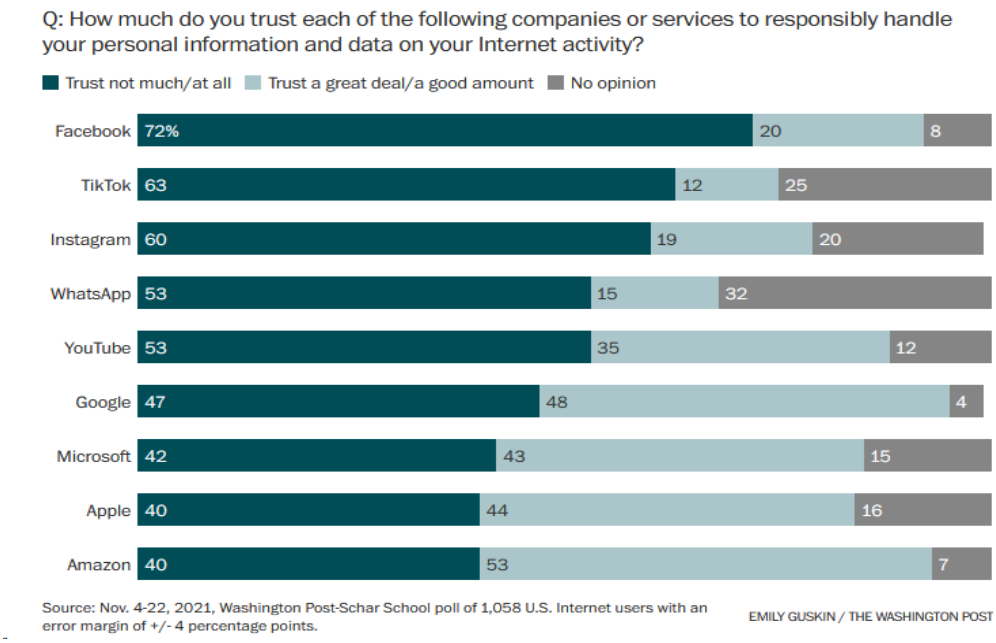
Due to the declining trust in Facebook, particularly after the Cambridge data breaches in 2018, Meta's Oversight Board was born within a lacuna for platform policy in 2020. Declining trust in Facebook manifested in two ways: first, in the public's distrust of the technology adopted by Facebook; and second, in the public's distrust of Facebook itself. This section discusses these two characteristics of declining trust. From a technological perspective, several factors contributed to public apprehension including: the open API for platformisation; the algorithms for ranking news feeds; and the tools for collecting data (datafication). These technological functions had disrupted the original public order and led to a breakdown in the traditional public sphere. These public concerns also related to other issues on Facebook such as the proliferation of 'fake news'. The spreading of misinformation on social media platforms positioned Facebook as a primary conduit for the dissemination of information, thereby shaping the public discourse about content moderation at the platform level. To be more specific, there was an urgent need to build a new order and rules that could be voluntarily followed by multiple stakeholders. By doing so, it will become possible to rebuild public trust in social media platforms in the future.

Technology

The issue of ‘trust and technology’ is poised to be a focus of discussions about technology policy in the near future (Bodó, 2021). In particular, new technologies are increasingly organised in ways that obscure the identities of creators, promoters, and middlemen, making it even more difficult to regulate their actions and decline or withhold trust with discrimination (O’Neill, 2020). In addition, recipients of digital content often struggle to detect the procedures of specific claims, their objectives, funding sources, or underlying objectives and agendas. For example, the phenomenon of ‘platformisation’ in the online world enables the proliferation of intermediaries at low cost but it does not ensure that they are held accountable for the outcomes arising from their collective actions.

Research (see Figure 3) suggested that the services or technologies offered by Facebook had failed to earn user trust in the digital age. According to a Washington Post survey (Kelly & Guskin, 2021), 72% of platform users expressed little or no trust in Facebook to responsibly manage their personal information and the data related to their activities on the Internet. In addition, 8% of platform users offered no opinion on whether they trusted Facebook in this regard.

Figure 3: Trust in Social Media Platforms



Source: https://www.washingtonpost.com/technology/2021/12/22/tech-trust-survey/

Meta published several sets of public materials relating to the relationship between algorithms and news feeds between 2010 and 2020, ranging from the operation and the designation of "Trending topics" (Osofsky, 2016) to three updates to "Trending topics" (Cathcart, 2017). However, these publications did not effectively enhance Meta's transparency or accountability. The Cambridge Analytica scandal in 2018 generated a significant expression of intense mistrust towards technology companies, especially Facebook, which some called the 'techlash' (Zuckerman, 2021). The decline in trust in technology and technology companies, especially in the case of Meta, is caused by the misuse of technological tools for data collection and the consequential misuse of data. In response, Facebook took steps to improve technological tools to prevent misinformation, including the publication of the 'Facing Facts' short film, the launch of a news literacy campaign, and the introduction of measurement of misinformation via an academic commission (Hegeman, 2018). Facebook also introduced an

explanation feature for the algorithms behind the News Feed with "Why Am I Seeing This?" (Sethuraman, 2019). What happened with Cambridge Analytica was a breach of Facebook's trust. More importantly, it was a breach of the trust people placed in Facebook to protect their data when they shared it (Meta, 2018a). In an attempt to prevent a similar issue in 2018, Facebook announced that 'we wanted to bring you all together, now that the UK General Election is underway, to set out the range of actions we are taking to help ensure this election is transparent and secure – to answer your questions and to point you to the various resources we have available' (Meta, 2019c).

Despite these general statements, Zuckerberg's reaction to public events has typically followed a predictable cycle: denial, acceptance, a commitment to change and an apology, but no subsequent enactment of change (Bucher, 2021). This pattern has also applied to the responses to Facebook's handling of public issues. Despite the reforms driven by Facebook's dominance, and regardless of the Facebook leadership team's commitments, the public is becoming increasingly distrustful of Facebook's sincerity in dealing with these issues. Facebook now garners little trust, partly because it has previously failed to meet expectations and partly because its content moderation decisions have been perceived as inconsistent, contradictory, arbitrary, and profit-driven (Andrews, 2019).

Facebook itself

With public discourse concerning increasing platform power (dominant market power and the control of cultural production) moving to social media platforms, including

cultural production and social interactions, it has become imperative to evaluate the decisions that moderators make (Gillespie, 2019). Entering the digital age of public discourse via the Internet also means that the decisions made by leaders of social media platforms have exerted broad effects on the public, since their voices can reach millions or billions of users. However, the lack of gatekeepers or filters controlling the flow of information on social media platforms – aside from human-designed algorithms that direct flows based on user activities and the maximisation of views and shares – means that social media platforms risk becoming breeding grounds for fake news (Flew, 2021b). Since social media platforms serve as intermediaries for information dissemination, the spread of misinformation, disinformation, and 'fake news' is often related to the trust crisis on these platforms (Flew, 2022). This is why, in April 2018, after the Cambridge Analytica scandal captured global public attention, Facebook embarked on its biggest-ever advertising campaign to try to rebuild trust (Andrews, 2019). The origin of this 'crisis of trust' was the inadequate remedies offered for the problems arising in a changing environment, including the difficulties professions and professionals faced in maintaining and discharging their obligations in changing institutional landscapes (O'Neill, 2014). According to Schlesinger (2020), the lack of effective regulation is at the heart of public concerns about the spread of harmful content for individuals as well as institutions, which is also the current regulatory concern.

A Gallup report (2021) showed that, as a generation that grew up with the internet, young people relied on social platforms but did not trust social media. According to

the report, young people between 15 and 24 years old constituted the majority of social media users (45%), with people aged 40 and older constituting only 17% of users. In order to keep up with the latest on current events, young people relied on internet and social media resources. Of all the organisations they were asked about, young people were least inclined to trust social media platforms to give them accurate information. A median of 17% of young people indicated that they trusted social media platforms. Young people were twice as likely to place their trust in the accuracy of national (37%) and international media (36%) than social media platforms.

Contemporary 15- to 24-year-olds are growing up in a digital society. They have always had access to the internet, and young people have also always had social media. Furthermore, it is becoming more and more difficult for individuals of all ages to distinguish fact from fiction in what they see, read, and hear, given the current climate of misinformation and disinformation.

Given that young people’s decisions about their lives are shaped by information, which can be true or untrue, it is important to know what sources they turn to for news and which institutions they most trust to keep them informed. According to O’Neill (2014), the action can be interpreted as placing and refusing trust in others’ promises or truth claims. This includes evaluating the available information, including the institutions’ track record, its public communications, and the likelihood of the institution keeping its word. Facebook’s historical handling of disinformation has served as a model for young people's decision-making.

In conclusion, the issue of declining trust in Facebook was due to the technology and the platform itself. Although Facebook had been working hard to improve its technology and public discourse, what is needed to rebuild trust is not only a change in technology or an improvement in the channels of information dissemination but also the establishment of appropriate rules and regulations, a common standard that applied globally, and a common understanding of moderating content within a legal framework.

3.4.2 Lack of Legitimacy in Regulating Facebook

In the digital age, technology and individuals can shape and be reshaped by each other, which can be observed in the declining trust in Facebook. In other words, the new social context has rewritten the imbalanced relationship between institutions (Facebook) that possess the technological tools and discourse (related policies). Given this, the relationship between "code" and "law" needs to be defined again in the digital era. As suggested by Lawrence Lessig (2006), code is law, which implies that the design choices made by engineers while developing the internet infrastructure work to limit what is likely to be done online, and those with power over the software and protocols the general public use daily are capable of shaping how the general public behaves. Nevertheless, a broad interpretation of the relationship between "code" and "law" is proposed. As governments place more emphasis on technology and the relationship between "code" and "law" has a much larger set of rules and laws that relate to digital technology, scholars observe that "law is code,"; that is, the rules humans want to have

as a society start to condition technology architectures (Filippi & Hassan, 2016). In this way, “code is law” has transformed into “law is code”. This transformation manifests not only in the evolution of technology but also in the varying roles in the regulatory context and in the interests of stakeholders. Thus, it can be inferred that regulating Facebook in this age requires new rules, including placing the moderation of content within a legal framework.

Indeed, Facebook turned to democratic discourse for legitimation to address declining trust in the platform (Schwarz, 2019). Furthermore, the unequal power dynamic between users and platforms is evident in the opacity of platform operations. The topic of where a platform’s obligations for content moderation end and where the user’s obligations begin is notoriously complicated (Helberger et al., 2018). In this regard, regulating Facebook within a legal framework has the potential to break the platform’s power and establish a new order for moderating content. As Democratic Sen. Dianne Feinstein said, “what we are talking about is a cataclysmic change…. You created these platforms and now they are being misused. You have to be the ones to do something about it or we will” (Timberg et al., 2017). The regulation of Facebook has progressed from policy silence to policymakers’ participation. Those with the most policy-making power have transformed policy silence – which involves the options not taken into account, the questions excluded from the policy agenda, the players not invited to the policy table, and the values considered unrealistic or undesirable – into positive policy responses.

Despite these changes, nation-state power to regulate social media platforms, especially Facebook, remains limited. Globally, governments have learned how to regulate the internet by targeting technology providers and users within their respective jurisdictions. However, these states' powers remain limited when it comes to regulating such a huge social media platform effectively. Indeed, a large number of internet service providers have the freedom to decide where to establish their corporations and which laws they need to adhere to (Suzor, 2019). Instead of obeying the law, these intermediaries, including Facebook, can also choose to withdraw their business from a territory.

To illustrate, while technology discourse is replete with visions of empowering the dissemination of information to everyone in all corners of the globe, this expansion of scope has not been accompanied by an expansion of trust. Also, more often than not, it is the expansion of scale that has given rise to the difficulty of building trust and moderating content on Facebook, including the human-designed tools, data collection, and user-generated content. This state of affairs raises the question of why national policies or co-regulation between business and government sometimes fail to reach a consensus. The media concentration in global technology companies like Facebook can be viewed as a strategy to unite users on a single global platform, Facebook. While independent versions of co-regulation in different nations decentralise the power of the platform, it also implies that the original consensus on Facebook is beginning to fragment into multiple consensuses among different frameworks of regulation on Facebook and among different nations in the post-globalisation context. Like Facebook, a US-based technology company, the strategy and the operating goal differ from other

countries' policies. Given this, the declining trust in Facebook has become a truth, and the organisations operating globally, such as Meta's Oversight Board, fill the gap in eliminating trust gaps and reaching platform consensus globally. This solution still requires a thorough examination of the current governance framework. In this regard, the next section analyses Meta's Oversight Board from the dimensions of its operation and drawbacks.

3.4.3 Regulating Facebook within a Legal Framework

The concentration of platform power on Facebook and its operation without effective regulation have resulted in declining public trust. Prior regulations, especially self-regulation, lacked sufficient safeguard measures to guarantee that the power of technology enterprises could be exercised in a more legitimate, humane, and equal manner (Suzor, 2019). In this context, the question arose: who could set the rules to regulate social media platforms in this age – the government or social media platforms themselves? Given the drawbacks of self-regulation and external regulation already discussed, it was hoped that Meta's Oversight Board could serve as a third way to regulate Facebook within a legal framework. This section begins with an introduction to Meta's Oversight Board, the issues related to it, and its operational difficulties.

Meta's Oversight Board

Before his appearance in the US Congress in 2018, Mark Zuckerberg acknowledged the need for regulation of the platform. According to Zuckerberg, with the internet becoming increasingly critical in human lives, the key question was what was the right

regulation, rather than whether or not there should be regulation (Zuckerberg and the Senate Commerce, Science, and Transportation Committee, 2018). In the same year, Zuckerberg pointed out in the post *A Blueprint for Content Governance and Enforcement* that, during the following year, a new way would be found for people to appeal content decisions to an independent body that would make transparent and binding decisions. This Oversight Board's objective would be to follow the principle of allowing people to make a voice while realising the reality of maintaining their safety (Zuckerberg, 2018). Apart from the proposal for establishing an independent oversight board to make decisions on the rectification of errors, Zuckerberg's posts also mentioned two other areas of focus. One area would be to release transparency and enforcement reports quarterly, and the other area would be to allow for more academic studies (Zuckerberg, 2018).

The first draft charter, published on January 28, 2019, indicated that the board would be a body of independent experts who would review Facebook's content decisions, with a focus on key and controversial cases. The body would transparently disclose its decisions along with the reasons for those decisions (Meta, 2019a). Moreover, the board also mentioned that it aimed to give oversight of how responsibility would be exercised and make Facebook more accountable (Meta, 2019a). From this perspective, the establishment of Meta's Oversight Board had the potential to enhance Facebook's accountability and transparency. Following the charter, Harris (2019b), Director of Governance and Global Affairs at Facebook, published an article titled *Establishing Structure and Governance for an Independent Oversight Board*, which explained issues

relevant to Meta's Oversight Board. Finally, Meta's Oversight Board was established on May 6, 2020 (Clegg, 2020). As a quasi-judicial body, overseeing Facebook's content moderation judgements, Meta's Oversight Board consisted of experts from all around the world (Flew & Lin, 2022). The Board was established with a governance structure and charter that guaranteed its independence from Facebook, and its operations were funded through an operating trust structurally separate from Facebook and administered by independent trustees after an initial gift of $US130 million from Meta to establish the entity (Clegg, 2020). Members of Meta's Oversight Board are listed in Appendix III.

The Meta Oversight Board began hearing cases in October 2020 (Harris, 2020a) and chose the first case to review in December 2020 (Harris, 2020b) as an extended case study for assessing quasi-judicial processes operating within a giant technology company. The Board's establishment, as a private-public oversight experiment aiming to resolve disputes in content moderation, gave Facebook an advantage over policymakers in regard to the definition of co-governance frameworks (Popiel, 2022). In this way, the Meta Oversight Board constitutes a very timely case study in whether greater engagement of non-governmental organisations and academics, within a framework of industry self-regulation, can achieve superior content moderation outcomes that serve the public interest better than external regulation by nation-state governments.

However, the establishment of Meta's Oversight Board had its critics (Flew, 2022).

After the announcement of its establishment, an ad hoc group of academics, activists, journalists and others announced the establishment of a Real Facebook Oversight Board. Headed by Guardian investigative journalist Carole Cadwalladr, this group intended to utilise their diverse platforms to reveal the multiple ways in which Facebook algorithms fostered content that was divisive, inflammatory, and extreme, amplified false and incorrect information, and facilitated deceitful political advertising (Halpern, 2020).

The establishment of Meta's Oversight Board was intended to rebuild public trust in Facebook. In 2019 and 2020, there was a huge amount of public distrust expressed towards social media platforms. While it remained unclear whether this distrust could be addressed via the establishment of this Board, the change allowed Facebook to have real discussions about human rights in a way that would actually have a positive influence on Facebook's policies and its broader approaches (Suzor, 2024). In this regard, over the following five years, Meta's Oversight Board released related case studies to consolidate the accountability and transparency of Facebook. The remainder of this section discusses the operations of Meta's Oversight Board and the difficulties it faced.

Choosing case studies

The Board had two major functions. First, the Board made binding decisions on individual pieces of content. Second, the Board made non-binding recommendations on company policy, across all of Meta's platforms (Oversight Board, 2023). For

submitting the case studies, Meta's Oversight Board opened the way to receive case studies from Facebook, Instagram and Threads users. If users did not agree with Meta's decisions to remove their content on these social media platforms, they could submit their case studies to Meta's Oversight Board to review their cases. Also, Meta could directly submit a case study to Meta's Oversight Board, including controversial cases. In regard to non-binding recommendations on company policy, the Board's policy advisory opinions constituted reviews of a selection of Meta's policies and enforcement mechanisms, such as those addressing health misinformation or privacy, and an exploration of how those policies could be improved. The disparity between binding and non-binding decisions was that Meta was not obligated to implement recommendations to binding decisions. Despite this, Meta was required to respond publicly to Board recommendations within 60 days, thereby establishing a degree of transparency unique to the Board's work.

In terms of authority for reviewing case studies, the board would review and decide on content following Meta's content rules and values (Oversight Board, 2023). The board had a full-time staff responsible for supporting its administration and operations. The staff's duties included reviewing submissions of case studies and coordinating outside research and statements for selected cases.

Difficulties running Meta's Oversight Board

Unlike auditors, who work with preexisting standards, Meta's Oversight Board did not only make recommendations to Meta, including policy and case studies, it also

evaluated Meta's adherence to those recommendations. In regard to these responsibilities, the Board has faced certain difficulties, including tracing misinformation and its independence from Meta.

To date, Meta's Oversight Board has released 150 case studies and four policy advisory opinions. Among them, seven case studies and one policy advisory opinion related to misinformation. Meta's Oversight Board co-chair, McConnell recognised these cases as the most difficult area for content moderation (quoted, Shapiro et al., 2023, p. 4). Facebook has long been using constantly changing technological tools, such as changing the algorithms and the digital context, to combat misinformation. Nonetheless, these issues always concern the public. Even though machine-learning algorithms were used, Facebook applied the tools with the rules of building credibility and trust with the audience (Kacholia, 2013). The distribution of misinformation on Facebook with the increased number of users raised concerns for the quality of public discourse. In 2017, Facebook introduced new tools against misinformation, including tips on the way to detect false news, such as checking the site URL and the source and searching for other reports on the same topic (Mosseri, 2017). The measures to address misinformation were improved in 2018, with their expansion to include penalising clickbait, links shared at a high frequency by spammers, and links to websites with low quality, also called "ad farms". Other actions were taken, including: cutting the News Feed distribution of entire pages and websites that shared false information repeatedly; the introduction of third-party fact-checker tools (Lyons, 2018); the combination of human reviewers and a machine learning classifier to compile misinformation signals (Meta,

2018b); and other solutions to fight false news at scale (Silverman, 2019a).

Meta's history in preventing the distribution of misinformation has shown the evolution of measures concerning human reviews and platform-designed tools. Nonetheless, the difficulties associated with tracing the sources of misinformation and with moderating content on Facebook at a large scale cannot be solved by a third party, such as Meta's Oversight Board. Facebook's fight against misinformation has become more complex since Meta released the new version of Meta AI: the assistant that users can ask any question when using Meta's apps and glasses (Zuckerberg, 2024). In this vein, Meta is relevant in Generative AI (Spencer, 2024), which can pose challenges in moderating content on Facebook, Instagram and Threads. In addition, the technological industry has changed with the evolution of technology, such as Generative AI, which may fundamentally change what it is possible to do automatically (Suzor, 2024).

In addition, the difficulty in running Meta's Oversight Board has also been due to its independence. The key problem is information asymmetries between Meta and Meta's Oversight Board, which affect Meta's Oversight Board's implementation in reality and communication between the two. Under these conditions, the corporation's leadership could do what it thought was best, regardless of the Board's recommendations. This aligned with common corporate practice in many domains of social responsibility (Shapiro et al., 2023). Specifically, Meta's Oversight Board faced two difficulties when moderating content. First, it faces difficulties when making decisions about what should be allowed on the social media platforms. Furthermore, the decisions were also

complicated by the difficulty the Board faced when trying to obtain information from the companies about how they moderated content (Suzor, 2024). When the Board was initially conceived, the focus was mainly on who the Board members would be and the legal framework they would use to make decisions. During the first year of the Board's operation, tracking Board recommendations consisted of noting whether Meta said they would agree to implement the recommendation or not. There was no mechanism for independent evaluations of Meta's implementation, whether Meta was misinterpreting the recommendation, or of what constituted sufficient evidence of implementation. The Board came to understand that it was struggling to adequately answer these questions. In response, in July 2021, the Board dramatically changed its approach to recommendations by creating an Implementation Committee and hiring a team to build an analytic and data-driven infrastructure to support that committee. The knowledge gained over the four-plus years after this work began is explained in detail in Chapter 4 of this thesis.

3.5 Conclusion

This chapter has analysed Facebook's self-regulation between 2010 and 2020. Increasing criticism over a decade of self-regulation resulted in co-regulation between the government and social media platforms, including Facebook, with the government taking on a larger role in regulating those platforms. Discussing the limitation of co-regulation, this chapter has outlined and evaluated Meta's Oversight Board, specifically the reasons for its establishment, its operations, and its challenges. Meta's Oversight

Board could be regarded as a contemporary third way to regulate social media platforms. Therefore, this study fills the gap in analysing the current regulatory methods for moderating content on social media platforms.

Chapter 4: Revisiting Trust in the Misinformation Age: Collaboration, Legitimacy, and Responsible Platforms

4.1 Introduction

This chapter focuses on the implementation and effectiveness of Meta's Oversight Board in moderating misinformation and answers the following questions: what is the definition of misinformation; and who should be responsible for defining and governing misinformation? First, this chapter defines the terms 'misinformation' and 'disinformation' to respond to the questions. Second, this chapter presents a discussion of the large-scale moderation of misinformation using study participant data. This chapter concludes that becoming a responsible platform requires Meta to collaborate with the government to follow the law, or working with the government can be called 'the regulatory return'.

In light of this, this chapter aims to answer the core question: In regard to misinformation, how do social media platforms moderate content and who should be responsible for regulating social media platforms?

4.2 Misinformation and Disinformation

It is essential to have clear definitions of misinformation and disinformation to more clearly guide those given the responsibility for the removal or retention of content on social media platforms (moderators). These terms also work as guidelines for users to post their content on social media platforms. The ineffectiveness of the self-regulation

framework and the continual resistance of technology companies to government involvement have eroded public trust. There have also been public expressions of disappointment with the way these companies have handled the distribution of fake and harmful information, labour-associated issues, and other matters. (Popiel & Sang, 2021). While some scholars (Wardle, 2017) have criticised the term and argued that it should be abandoned owing to its loaded definitions and the way politicians have used it to characterise content they disagree (as 'fake news'), it remains a contentious issue.

Study participants defined misinformation as online information that misleads users on social media platforms. However, misinformation is not necessarily intentionally distributed. Some scholars believe that the release of inaccurate information is not the same as spreading false information for a mentioned purpose, which they refer to as 'disinformation' (Benkler et al., 2018). Although disinformation and misinformation are similar terms, users are often confused about their differences, and academia and industry also use these two terms interchangeably when referring to false information. Disinformation is intentionally distributed with the intention of misleading the recipients. Disinformation campaigns weaken public trust in civic norms and institutions, whether they are performed for amusement, violence, or financial gain (Vaidhyanathan, 2021), and deliberately disseminating false information on platforms will reduce the trust of users in platforms. Therefore, users will reduce information dissemination on social media platforms so as to avoid damaging their credibility due to the spread of false information. More importantly, however, the access of platforms to user information and their control over user information allow them to target users

with false information specifically tailored to their interests. In this vein, the origin of the widespread and quick distribution of disinformation on Meta, X (formerly Twitter), and YouTube could be traced to the eroding of user privacy and the access of other parties to user data. The surveillance practices of Facebook are considered to breach public trust through their technologically designed tools. Bennett and Livingston (2020) characterised disinformation as deliberate lies or distortions that are usually disseminated as news to realise political objectives. These objectives include undermining opponents, swaying voters, interfering with policy discussions, intensifying existing social conflicts, creating a general atmosphere of confusion, and informational paralysis.

Several study participants expressed their ideas about the definition of misinformation and disinformation. For example, participant 1 commented:

> About the point of disinformation and misinformation, I did think that there was value in distinguishing between information that has a kind of malicious intent and is circulated without that kind of malicious intent. They both are wrong, misleading and harmful, but I did think it was useful to distinguish between those two.

In view of this, it is valuable to distinguish the terms ‘misinformation’ and ‘disinformation’, as they represent different meanings in different contexts. In other words, distinguishing these items could help social media platforms rebuild public trust, because users can better understand the differences between them and avoid

being misled by certain misinformation or disinformation events.

Participant 3 commented:

> Information is false and harmful. So, those two elements come to bear. The standard distinction between misinformation and disinformation is that one is intentional and the other is not. For the purpose of harm, however, I think probably all we need to know is that it's false and harmful.

Participant 4 suggested that,

> Misinformation generally refers to false information that people spread. They don't know that it's false. Now, the problem with that is that you need to define what false is…. Determinations about what is true and false are always political and ethical because most of the time, there's no way they don't correspond to anything right like.

Participant 6 argued that,

> I prefer to use the term propaganda when talking about manipulative and persuasive communications, because this is kind of broader. It encompasses different kinds of manipulation that are not just about false or true.

Interview data provide more nuanced distinctions for these definitions. Participant 1 emphasised the necessity of determining whether the dissemination of such information was malicious and argued that different measures should be applicable to different types of information dissemination. Participant 3 indicated that, while the

dissemination of false information and misinformation may be intentional, the focus should be on what is false and harmful for regulatory purposes. Participant 4 introduced an epistemological perspective and noted that defining 'falsity' proves particularly challenging in political and ethical debates—such assertions cannot be verified like empirical facts. Participant 6 favoured employing the term 'propaganda' to contain a broader spectrum of manipulative dissemination practices, which may not hinge entirely on binary judgements of truth or falsity.

When reposting and posting false information, users are unaware that they are being misled by the false information. Misinformation is also less traceable (i.e., its source is more difficult to identify), which poses a challenge for moderators.

As such, misinformation can be used as a strategic asset for campaigns and certain digital media firms (Bennett & Livingston, 2020). To prevent misinformation from being misused, policymakers and governments should play a central role in moderating content by introducing relevant laws or rules. The reason is that it is difficult for corporates to moderate misinformation without exerting power over users, because corporate regulation could only remove the content or suspend users' accounts, but cannot legally punish the users. In this vein, the legitimacy of the guidelines should be established by the states to effectively moderate misinformation online.

In conclusion, misinformation on social media platforms is unintentionally distributed, while disinformation is intentionally spread. This chapter uses the term

‘misinformation’ for consistency with the Meta Oversight Board’s use of the term. Nevertheless, as pointed out by Briant and Bakir (2024), in addition to the platform’s power to distribute misinformation, different actors also play a role in shaping, designing and exploiting the power to distribute misinformation for their profits, such as the influence industry. Moreover, bots and fake accounts also spread misinformation online for propaganda or misinformation distribution. This phenomenon makes it difficult to moderate content on social media platforms because of users’ difficulty in recognising the bots and fake accounts.

4.3 Moderating Misinformation on a Large Scale

Despite enabling users worldwide to connect, the development of social media platforms has also generated enormous risks and challenges around freedom of speech and content moderation. Meanwhile, the development of social media platforms has also made it hard to moderate misinformation, as false information can also spread globally. In this vein, misinformation has been moderated on a large scale in this age.

4.3.1 Moderating Misinformation

Facebook, YouTube, Twitter (now X), and other major user-generated content platforms, have increasingly deployed automated hash-matching and predictive machine learning technologies (what users refer to as algorithmic moderation systems) to undertake large-scale content moderation (Gorwa et al., 2020). With the increase of political pressure on giant technology companies, corporations and legislators alike are looking for technical solutions to challenging platform

conundrums like misinformation. To moderate misinformation on a large scale, the keyword-based detection system developed by Meta has been adapted to local contexts and covers market-specific terms. It is necessary to consistently apply initiatives like the keyword-based detection system in all countries undergoing elections and other democratic processes.

Nevertheless, democracy is faced with multiple digital challenges, and large-scale misinformation cannot be solved by more technology (Faris & Donovan, 2021). Unlike human-moderated information, the information designed and implemented by algorithms have their own drawbacks, because human-designed algorithms inherently carry biases. In addition, real information is sometimes disseminated misleadingly (for instance, by slowing down the speed of a video or audio recording), which can mislead the audience's interpretation of the information and exert effects on the behaviour of users (Briant & Bakir, 2024). Likewise, these issues cannot be solved within a technology governance framework that establishes a clear distinction between commercial and public values and activities. Instead, they raise more general issues associated with exploitation and power (Taylor, 2021). On the other hand, the larger complex of interconnected international anti-misinformation efforts is made visible by paying close attention to the meso level of institutional logics and relationships (Bélair-Gagnon et al., 2022). It involves fact-checkers, news organisations, technology companies, government agencies, transnational institutions, non-governmental organisations, academic institutions, and charitable foundations, among others.

Although the contemporary moderation of misinformation requires large-scale regulation of content, moderating content on a large scale by technology is often criticised for lacking transparency. Meta's Oversight Board has played a key role in enhancing the transparency of decisions made by Meta. Participant 2 from a non-governmental organisation noted:

> I think probably a focus on transparency is pretty important. I also think de-platforming without remedies and the right to appeal is a problem. Hence, I think those things need to be addressed. The scrutiny of platforms needs to be open because a lack of transparency has created energy for regulation. That's not always been good regulation, but it's a response to a failure to be open about how algorithmic content moderation takes place. That's particularly true for vulnerable people.

Participant 4 took a similar view:

> They have all of these inconsistencies about how they actually apply their moderation in practice when doing content moderation. Thus, they have a multitude of terms around authenticity, coordinated inauthentic behaviour, coordinated social harm and platform manipulation.

From the perspective of moderating misinformation on a large scale, human-designed tools should be used to moderate misinformation on Meta to meet the expectations of the public. However, algorithmic decision-making or other human-designed tools have the bias of the procedure of designing the tools. More importantly, when

researchers have the right to access and study it and investigate algorithms, Facebook will also modify its algorithms (Graham, 2024). In this vein, human involvement (like Meta's Oversight Board) in moderating content, within a legal framework, to ensure human rights and user freedom of speech, fills the gaps left by technology.

4.3.2 Reflections on Regulating Misinformation on A Large Scale

In the digital age, individuals are partly forced to engage in using social media platforms. The design decisions of social media platforms define what is feasible; the content regulation of social media platforms determines what is acceptable; the personalisation algorithms of social media platforms decide what is visible (Sander, 2020). As such, content moderation on digital platforms has its own drawbacks, since the rights of the public are not respected. Users cannot decide what content to view. The information that flows to users is determined by algorithmic recommendations rather than their preferences or choices. Such an approach violates the human rights of users, and users should be treated as complete human beings rather than programs needing to be fed information. Individuals are also finding it increasingly difficult to choose not to interact with social media platforms due the data-driven power of commercial technology companies over individuals, together with the growing participation of the private sector in public governance (Taylor, 2021). Furthermore, algorithms are used for ranking content feeds, prioritising user engagement above all other considerations, and creating minimal discernible differentiation between the presentation of various forms of content. Algorithms soon level the playing field for online content and put even the roughest-hewn user-generated content in a shared

stream and on par with carefully fact-checked pieces written by experienced journalists. Under this unified circulation model, anything can immediately become viral with sufficient interaction, regardless of its authenticity or provenance (Bowers & Zittrain, 2020). When mediating political communication, commercial algorithms also inexorably affect people's perception of the significance and consequences of the forays of technology companies into the public domain (Taylor, 2021).

Users will no longer trust social media platforms and the information they disseminate after repeated experiences of discovering that they have shared false information, believing it to be true. When users have no choice but to engage in declining-trust social media platforms, refining concentrated platform power and revisiting designed technologies become part of content moderation on social media platforms. Concentrated media power is antidemocratic because it endows unelected organisations with definitional, analytical and interpretive power and makes it harder for citizens to gather and share the variety of knowledge and ideas needed to make well-informed decisions on public life. Such power concentration is also risky because it warps the logic of the media industry itself, transforming it from symbolic interaction instruments into increasingly important capital accumulation engines (Freedman, 2014).

The COVID-19 outbreak rendered taking action against misinformation more crucial than ever. COVID-19 raised concerns about the distribution of information on Meta, especially the approaches to truthful content. The period experienced an infodemic when inaccurate, misleading, and unsupported information and rumours regarding

COVID-19 began to circulate (Siapera, 2022). According to the World Health Organisation, this global pandemic was characterised by excessive information. Some information was correct, and some was not, which made it challenging for individuals to locate reliable sources and guidance (Bélair-Gagnon et al., 2022). This results from the simultaneous global scale and possibly fatal consequences of the problem, as people looked for guidance on how to reduce their chance of infection (Di Mascio et al., 2021). Nonetheless, users found it difficult to understand what content was prohibited because of the disorganised rules and policies scattered throughout Facebook, the ambiguity of important terms like ‘misinformation’, and the inconsistent criteria for determining whether a post could contribute to or genuinely cause imminent harm.

Individuals feel that their actions are restricted and that they may be compelled to behave against their own will (Hosking, 2014). As a result, content removal can foster distrust and encourage claims of cover-up and bias instead of promoting trust, particularly eroding freedom of speech. Without commenting on its fundamental ideas, for example, Meta might mark misleading content to alert users that it was created or greatly altered, which provides users with some context when they are considering the content's authenticity. However, platforms must utilise automated techniques for proactively and widely identifying illegal or otherwise harmful content because they are increasingly limited in the amount of time they have to take down content (Gorwa et al., 2020). For this reason, Meta is more inclined to delete content rather than choose other ways to moderate content, and its primary reliance on the

removal of content violating regulations could result in excessive restrictions on the right to freedom of speech. With regard to disinformation, platforms should enhance transparency by preserving details about disinformation they have identified, apart from documenting their attempts to monitor and remove such content. Given the sensitivity of this information, platforms might choose to limit access to approved, completely independent academic researchers and postpone making it available for several months (Bowers & Zittrain, 2020).

Regarding the efforts of the government to moderate content on social media platforms, Article 14 of the Digital Service European Union's Act is particularly interesting because it can be analysed from two perspectives: first, the article is used as a means of strengthening transparency requirements to ensure the informed consent of users with the terms and conditions of companies; and second, the article is used as a way of requiring social media companies to adopt practices complying with international human rights law (Fasel & Weerts, 2024). However, from the perspective of the Meta Oversight Board, it is concluded that its aim is to foster a clean online environment to win the trust of users in the content of Meta. This is because Meta's Oversight Board moderates content within a legal framework and enhances the transparency of Meta in moderating misinformation by reviewing the selected case studies. The purpose is to earn trust for Meta on the one hand. The platform complies with legal certainty requirements to ensure the freedom of speech in the global context, on the other hand.

4.3.3 Beyond 2020: Future Direction for Moderating Content: Automated Moderation

From this section, although this goes beyond the empirical scope of this study, it focuses on the current issues after 2020. Social media platforms are an integral and indispensable part of people's lives. Platforms like Facebook and X (formerly known as Twitter) make it simple for their users to access a range of information, from what a friend had for lunch to the intricate details and nuances of a specific political discourse (Tobi, 2024). In some respects, social media platforms know users better than users know themselves because platform-designed tools can predict users' behaviours. Opaque algorithmic power makes it possible for these platforms to know what users are doing and predict what users will do. To moderate content and rebuild a clean online environment, it seems that opacity is at the very heart of new concerns about algorithms among social scientists and legal scholars (Burrell, 2016). To earn more profits by selling the data of users to advertising companies, platforms advance opaque algorithmic tools to attract more users to engage in particular activities. However, users distrust an unknown system and the opaque algorithms have led to declines in levels of user trust in social media platforms.

However, government pressure on major technology companies to moderate content and reduce harmful or false information on social media platforms has been building. Both companies and legislators appear to hope that technical solutions to content governance puzzles can be found (Gorwa et al., 2020). The participants in this study indicated that moderating content or misinformation on Meta is expensive.

Participants 2 and 7 commented that the technology companies preferred to use technology-based tools to moderate misinformation. However, moderators demand low salaries so, Meta has recruited large numbers of employees to moderate content. Putting this action of large-scale recruitment of employees to moderate content into reality is expensive. In addition, Meta recruited more staff for moderating misinformation about COVID-19 during the pandemic, and those staff were laid off after that. Compared with human beings, technology-based tools can be expected to work 24 hours a day without rest. As participant 7 argued that even though human-reviews for misinformation can never achieve the same accuracy of 100% as technology-based review, technology companies still expect human review to be as accurate as machines. This is a difficult contradiction to resolve. Despite the ability to moderate content on the basis of cultural context, human reviews cannot match the workload capacity and accuracy of machines. Furthermore, technology companies are seeking automated moderation to meet the demands of users and governments for the moderation of harmful content over shorter and shorter timeframes. And, of course, using automated moderation can assist technology companies in removing illegal content faster and more efficiently. Hence, companies are investing heavily in their content moderation 'technology stack' to strengthen the accuracy of moderation by optimising their technical systems (Gorwa et al., 2020).

Study participant 2 (from a non-governmental organisation) argued that,

> Given the immense damage caused by opaque and inappropriate algorithmic models of content amplification, I think we do need greater transparency for

> people to understand how they work. At the moment, they're being gamed by all sorts of people, and it's creating huge problems.

Participant 3 offered their thoughts on this matter:

> I think the pathology of algorithms is that they became very successful tools able to maximise for commercial indicators, but less able to maximise for civic or democratic values.

According to their viewpoints, algorithms play a big role in moderating misinformation in the current era. Nevertheless, the design and operation of algorithms are opaque. Since then, in order to conduct content moderation based on democratic values, it is critical to understand and comprehend the operation of algorithms, since automated moderation of content has become a crucial approach in the future.

Participant 4 argued that,

> One of the reasons is that the problems of transparency and accountability get really complex…. If you have algorithmic moderation. We need to have an extremely clear technical understanding of the specifications of those algorithmic systems.

Participant 5 claimed that,

> You're dealing with billions of pieces of content. Every day, you surely need machines and automation to do it. That's a really important part of it….

> Then, when do you need the right to second-level appeals? Which Facebook set up the court of Facebook? It is the Facebook Appeals Board. You need those different levels at a certain stage, and that can take some human resources.

Regarding their viewpoint, automated content moderation is important for providing users with a clean online environment. However, as Participant 5 pointed out, setting up the second-level appeals is equally crucial, just like Meta's Oversight Board.

Participant 6 commented,

> I mean, I think AI can help flag content and then content should be reviewed by humans…. In my opinion, we need transparency over both what algorithms are doing. If having this kind of approach to social media, we also need to have some transparency over research and make sure that you know social media platforms are taking responsibility for what is done on them.

Participant 7 asserted that,

> Everything has changed…. It is such an exciting time because…. Historically. You would have said that you always need humans to make decisions. This is because computers are not good enough at considering context and nuance and understanding very fine distinctions. Now, I think the opposite is true.

From this perspective, human reviewers possess the ability to distinguish the nuance and understand the cultural context. However, with the advancement of technology,

machines can do better than human reviewers. More importantly, transparency and accountability of algorithms or artificial intelligence should be made public to the public or researchers, because rebuilding public trust in social media platforms requires the openness and accountability of social media platforms.

These participants expressed their concerns about the large-scale moderation of misinformation by human reviewers or algorithm tools. Algorithms are opaque, and how they work is difficult to know. With the emergence of AI, the generation of content has changed, which also poses a threat to content moderation. In this context, government policy has aimed to reduce the spread of illegal content on social media platforms. From the perspective of technology companies, specific keywords have been used as the basis for automatically deleting posts or blocking content through technological means. However, these measures are grounded in technical ways, which can give rise to the excessive removal of content and negatively impact on human rights. Furthermore, through technological means to moderate content, only deletion or suspension of accounts can be achieved, but there is no mandatory or enforceable effect, like the government's power to punish users through legislation. Therefore, government legislation is needed to prevent the spread of misinformation to achieve the goal of moderating content. In this regard, the next section discusses the moderation of misinformation within a legal framework, which means a return to government regulation of social media platforms through the introduction of laws: the regulatory return. The aim of introducing laws is to rebuild public trust in social media platforms and public confidence in government institutions.

4.4 The Regulatory Return: Moderating Misinformation within a Legal Framework after 2020

This section adopts the study participants' view that trust is a difficult concept to define and is usually regarded as legitimacy in the discipline of law to discuss the moderation of misinformation within a legal framework. This section then analyses legitimacy and concludes with Meta's Oversight Board as an intermediary of legitimacy, giving specific attention to the Board's capacities and limits.

4.4.1 Moderating Misinformation within A Legal Framework

Platforms always exercise discretion. Convincing platforms to exercise their discretionary power responsibly is a large part of making governance legitimate (Suzor, 2018). In addition, scholars suggest that one critical task ahead for studies of platform governance is to gain a better understanding of how discretionary power can and should be properly limited and made accountable, namely what regulatory scholars term throughput legitimacy (Haggart & Keller, 2021). Meta's Oversight Board could contribute to persuading Meta to exercise their discretionary power responsibly, i.e. moderating misinformation within a legal framework. In this way, the Board and its actions have the potential to contribute to rebuilding public trust in Meta, or trust can be regarded as legitimacy. Participant 7 stated that he found it difficult to define the concept of trust and that he preferred to use the concept of legitimacy instead. He defined legitimacy along two dimensions:

1. The empirical dimension: whether people believe a company has the right to

govern or operate in the manner it does. This dimension aligns closely with concepts like social trust and the ‘social licence to operate’. In essence, it is an empirical question: do people support the way the company is behaving?

2. The normative dimension: whether the actions of a company align with societal values in principle. This dimension focuses on whether the behaviour is morally or ethically accepted in light of shared norms and expectations.

From the perspective of participant 7, the moderation of content on Meta operated on two levels. First, the government could introduce laws to require technology companies to align with rules at their discretion. Second, collaboration among different countries would lay a solid foundation for governments in different countries to regulate technology companies on account of the differences in platform governance among multiple countries. According to the other study participants, things illegal in countries need to be moderated on platforms not illegal in other countries, or at least the legislation functions differently. For example, US second amendment laws are not applicable in Australia. Moreover, social media platforms have numerous inconsistencies in how they actually apply their moderation. However, platforms are not responsible for what users post due to the protection of social media platforms under the ‘safe harbour’ provision of Section 230 of the U.S. Communication Decency Act (Gillespie, 2018). Since there are no rules for moderating content on social media platforms, social media platforms like Facebook could remove or retain some online content in accordance with their terms and services. Before the Meta Oversight Board was established, Meta had a great deal of

discretion in the absence of mandatory government regulation.

While the European Union Digital Service's Act constituted a good start for government intervention, it was not enough. If the national states are able to force technology companies to delete content involving child abuse, then the government should introduce laws to prohibit the appearance of other misleading information on social media platforms. Given that different countries have different laws to regulate technology companies, it is crucial for different countries to collaborate in regulating global companies. Especially for some small social media platforms, they do not have the financial assets to establish a third-party institution like Meta's Oversight Board. Some study participants suggested that small technology companies run with their fans, but they are known to fewer people. Content moderation at these small technology companies also requires the introduction of laws. Apart from small companies, some small countries have different cultures, but their citizens are also engaging with social media. Small countries have limited capacities to regulate global social media platforms. Consequently, introducing laws from the government is also critical for these countries, especially through collaboration with multiple countries.

The commercialisation of social media platforms has reduced their legitimacy and credibility. These changes can be discussed using concepts including utopian/dystopian, the shift from pull models to push models, and surveillance capitalism. The digitisation of user data and digital colonialism play a pivotal role in this transformation from utopian to dystopian. One decade ago, users found that social media platforms met their requirements for communicating with their friends. Now,

social media platforms collect and analyse user data to generate profit. Utopia has transformed into dystopia. When social media platforms first emerged, they exhibited a pull model. Only friends were on user's list for communication. If users wanted to search for something or communicate with their friends, they needed to search or find someone by searching. Nowadays, social media platforms have shifted to a push model, as algorithms can predict users' behaviour and push the information they like to their social media platforms. Users and their data are resources for social media platforms: resources to be collect and analysed to earn profits. From one perspective, it is possible for social media platforms to predict the behaviour of users based on their data, like their 'likes' and 'shares' online. However, platforms do not focus on what users share or like. They only pay attention to the fact that users have performed these actions, which thereby enables more relevant recommendations. This kind of digital colonialism by social media platforms can also cause user fatigue. In this vein, because these social media platforms rely on advertising to make profits by selling users' data to advertising companies, digital platforms have been commercialised, even though they do not want to be so. However, the platforms' profit model leaves social media platforms with no other choice.

The commercialisation model of platforms has reduced user trust in platforms. Rebuilding trust in platforms must rely on external trusted systems, namely, government agencies. As such, rebuilding public trust means reintroducing laws or rules to moderate content on digital platforms. This action can be understood on two levels. The first level relates to news providers or true information provided by

advertisers. If the information flowing on a platform is all true, users will continue to trust the platform. The second level is to introduce laws to moderate misleading information on social media platforms. Trust could be rebuilt on social media platforms if users or news providers share true information by obeying laws. If the government cooperates and works together to formulate laws to moderate information, misinformation will disappear on social media platforms.

To sum up, legitimacy is another term for trust in the age of misinformation. This section has discussed the definition and forms of legitimacy, followed by the profits model and its impacts on social media platforms. This section has concluded that rebuilding public trust requires government intervention.

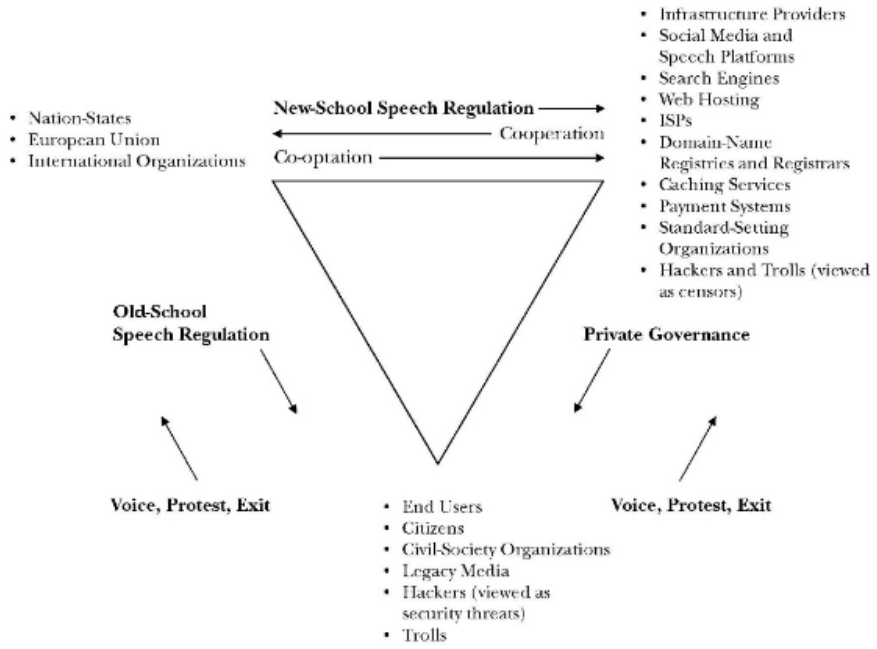
4.4.2 Balancing Freedom of Speech and Human Rights: Meta's Oversight Board

Meta's Oversight Board arose to fill a legitimacy gap. The Board, as a bridge between technology companies and users, mediate and moderates content within the legal framework. This study's analysis of the Board's reviews of moderation of misinformation on a large scale and its case studies has led to the finding that it is vital to balance human rights and freedom of speech when moderating misinformation. In other words, moderating content on social media platforms should respect users' rights and improve the platform's transparency in moderating content to users. Users are entitled to express themselves. However, the lack of transparency in platform regulation and the fact that users have no choice in terms and conditions mean that users' rights are not actually respected. Therefore, Meta's Oversight Board,

a third-party platform reviewer, needs to strike a balance between human rights and freedom of expression.

According to Balkin, 21st century freedom of speech is dependent on "a third group of players: a privately owned infrastructure of digital communications made up of firms supporting and governing the digital public sphere used by people for communication". Discussions about freedom of speech have been dualistic or dyadic, and oppose nation-states on the one hand and a range of speakers (from individuals to mass media) on the other (Balkin, 2018, p. 2012). The framework for freedom of speech develops into a free speech triangle in this context. These relationships can be seen in Figure 4.

Figure 4: Free Speech Triangle



Source: Balkin, 2018, p. 2014

Balkin saw the logical trajectory of this triangle tipping towards the further diminution of speech rights in the absence of a concerted action to expose the decision-making processes of governments and providers of private digital infrastructure to a degree of public scrutiny, accountability and transparency (Flew, 2021b).

In Balkin's framework, free speech has three interrelated elements, represented as corners of a triangle: governments, end users, and social media platforms. Unlike free speech in the 20th century, states regulate and protect free speech directly for end users. In the 21st century, states moderate content and free speech via social media companies. In this context, social media companies have turned into private bureaucratic institutions. In light of this, governments intend to collaborate with social media platforms to moderate content on social media. However, this raises three issues. First, that a third-party body like Meta's Oversight Board cannot impose harsh penalties in the same way as a bureaucratic institution can. The Board draws on international human rights law when presenting proportionality analysis as the legal standard to review interferences with human rights, specifically freedom of expression, and refers to the internal standards of Meta (Encinas Duarte, 2025). Of note, the United Nations Human Rights Council unanimously adopted the United Nations Guiding Principles on Business and Human Rights in 2011. Despite being non-binding, the United Nations Guiding Principles on Business and Human Rights have been endorsed by scores of large-scale companies around the world (San Martín, 2023), including the Meta Oversight Board. The Board can only remove or retain

certain content or suspend some accounts. Despite the existence of the Board, the original self-regulation has shifted towards procedural review. Second, Meta primarily concentrates on its own data-driven business model, which differs from the public interest focus of states. And third, in regard to government surveillance, the government will also use the data from social media companies for monitoring citizens or users. In this context, government intervention in content moderation without proper consonance with international rights may lead to the unwarranted censorship of online speech and disproportionately penalise or threaten platforms and their users (Santos et al., 2025). However, Tyler (2006) proposed that the public obeys the law not because they are afraid of punishment but because they believe that the law is fair and can be trusted. The public trusts the governments because the government respects users' rights by introducing fair laws to moderate misinformation on social media platforms. If it intends to enact laws to prohibit the spread of misinformation, these laws should serve the public interests. In this vein, government legislation on false information and the information technology used by the government is not reliant on the technology or government itself, but rather on whether these measures protect the public interest. When these measures protect the public interest and are consistent with social values, the public will trust the government or social media platforms and voluntarily comply with rules.

4.5 Looking Forward: Who should be Responsible for Moderating Misinformation?

Propaganda, conspiracy theories and politically motivated misinformation all have a

long history. None of this is new to some extent, as is the deliberate spread of falsehoods for economic or personal gain (Flew, 2021a). Robots are being used to automate the distribution of misinformation on social media platforms like X (formerly Twitter). 'Bots' are computer programs that mimic human users of Twitter, but disseminate content automatically by means of the network (Hendricks & Vestergaard, 2019). In addition to technological issues, a study found that Facebook users were more inclined to post popular fake news than popular mainstream news (Allcott & Gentzkow, 2017). Due to advancements in communications technology, misinformation can now readily and swiftly travel across the world (Morris et al., 2020). This gives rise to the concern that misinformation becomes more whitewashed as more misinformation is spread and replicated, and reaches new audiences and media outlets for both mainstream and alternative news (Hendricks & Vestergaard, 2019).

Misinformation issues on Meta raise the issue of how to moderate content effectively and who (the government or platforms) should be responsible. Self-regulation has the advantage of time, as companies can act swiftly, while laws take longer to come into force (Gorwa, 2019a). Nevertheless, nation-states persist in the face of multidimensional crises by evolving to fit the changing environment (Castells, 2013). Whenever democracies deteriorate and occasionally become autocracies, a crucial component of that decline is the use of discriminatory and offensive language in public discourse (Bennett & Kneuer, 2024).

Hence, it is argued that the government should organise self-regulation while giving

companies a great deal of discretion in making decisions (Stockmann, 2023). If the issue is framed as one of non-domination, the concept of legitimacy becomes a tool to assess the degree to which people seek to regulate technological power and how to approach such regulation (Taylor, 2021). Developing legitimacy is different from finding the substantively 'right' solution. Legitimacy is of particular importance when a wide public consensus on the right solution is lacking (Bowers & Zittrain, 2020). To avoid undermining other elements of content moderation governance, like the work of its Oversight Board, the United Nations Guiding Principles on Business and Human Rights recommendations are adopted, especially those that require the respect of the principle of legal certainty.

A number of the study participants believed that governments should play a part in moderating misinformation. Participant 1 contended that,

> There is a role for the government. If the government had moved and tried to do this six or seven years ago, it might have been seen as pre-emptive or premature. However, I think we've seen enough examples of problems now and the potential for them to be serious problems, especially with the sophistication of deepfakes and others. So, I think there is a role for the government.

Participant 2 (from a non-governmental organisation) suggested that,

> The government plays a role in regulating misinformation and disinformation, because it is a problem. A great amount of misleading content

> is on the web. We have an information ecosystem difficult to trust, and a set of business incentives for viral extremist content that may not have any reference to the truth, or there's no requirement that there's a reference to the truth, and it can still be extremely profitable.

Participant 3 argued that,

> Government plays a role in regulating false and harmful information…. In my view, it is also true that the media environment has shifted in ways that it's become quite important to hold platforms accountable for the spread of misinformation and disinformation.

Participant 5 asserted that,

> The government probably plays an important role, but you could see that working better in some contexts than others. However, I think the issue is if you're obviously using the misinformation laws to attack your critics…. I think the government has a role to play.

Participant 7 commented,

> Yes, the government plays a role. Government and private regulation pose a threat not only to freedom of speech but also to other human rights. That is to say, we have seen that plenty of governments worldwide are using the language of misinformation and disinformation to clamp down on political speech that they don't like.

In contrast, participant 4 claimed that,

> We have known for a very long time that truth and fault are rather context-dependent. So, we have to tread really carefully when it comes to state intervention. Personally, I think it's a bad idea.

Participant 6 opined,

> So, I think we fail to appreciate the full range of how complex communication can be, including when it is trying to deceive us, if we narrow ourselves to talking about mis/disinformation. Usually, if someone is trying to deceive us like a state actor, they will be doing it in many different ways.

The participants' different views indicated that the government plays a vital role in moderating misinformation but that government intervention should be careful to avoid wrong policies. If the government' one-fits-all policy is wrong, then moderating misinformation on social media platforms may become more complicated. In this regard, governments should be held accountable for moderating misinformation in this age. In addition to the effort of the government, involving other stakeholders in moderating misinformation is also crucial. Becoming a responsible platform involves the engagement of multiple stakeholders, including third parties. From the perspective of the Meta Oversight Board, interacting with third parties and receiving public comments plays a key part in making recommendations to Meta. The Oversight Board collaborates with a variety of third parties based on different case studies. For

instance, the Board commissioned independent research with the assistance of Duco Advisors, an advising company specialising in the intersection of geopolitics, trust, safety, and technology. Analysis was also supplied by Memetica, an organisation conducting open-source research on social media trends (the case study of the Australia Election Votes). Collaboration with third parties was also undertaken in other Board case studies. This collaboration has exerted positive effects on Meta's fair treatment of users and the rational decision-making of Meta's Oversight Board.

4.6 Reframing Platform Power and the Role of Meta's Oversight Board in the Misinformation Age (After 2020)

The changing structure of platform power has resulted in the gradual loss of trust in Meta. Unlike the previously open internet, social media platforms exercise extensive control over the distribution of information and the collection of user data and, consequently, they play a significant role in politics, society, and technology. The growing power of platforms has changed the narrative structure of politics and society, since there are millions or billions of users on social media platforms.

Participant 6 noted that,

> Elon Musk is so powerful. He is so incredibly powerful. His role in monopolising this technology infrastructure within political campaigns in the U.S. is very dangerous.

This trend is also visible when considering Meta. In early 2025, Meta, under the leadership of Zuckerberg, became the first major platform to follow the right-wing

turn of Musk. Shortly before the second inauguration of Trump, Zuckerberg issued a personal statement announcing that the platform would cease fact-checking and loosen its protections against hate speech (Leerssen, 2025; Booth, 2025). In the meantime, platform leaders have also played a uniquely influential role in structuring technological development and, more recently, reorganising financial systems and arrangements for public administration in particular ways, aligning with their personal beliefs. Different from traditional media, social media platforms can reach several billion users a day, and their influence has become ingrained in people's daily lives. The power of digital platforms is overwhelming in this age, which is attributable to the concentration of platform power on a few social media platforms. The impact of social media platforms has extended far beyond the power imbalance between users and platforms and the commercial structure of platform power exerts an influence in the political, economic, and social spheres. Participant 2 (from a non-governmental organisation) said,

> Platforms will be enormously tempted to double down on their general approach to content moderation and insufficient transparency and cooperate with authoritarian regimes, because they see that as the price they pay for being able to offer their service undisturbed.

The trust of users in digital platforms has shifted from a previous reliance on platforms themselves to a distrust of their commercially driven interests. Platforms may have never been truly neutral. However, their media-like functionality in shaping public discourse has steadily become more technically refined, politicised and

institutionalised as a form of speech governance (Leerssen, 2025). The binding character of external recommendations (like those from Meta's Oversight Board) may play a limited but concrete role in shaping the hierarchical configuration of content governance (Magalhães et al., 2025). In view of this, social media platforms fulfil multiple roles. They are market entities, infrastructure providers, and quasi-public spaces. The decisions made by social media platforms are restricted by various factors, including capital logic, national regulation, and public opinion. Relying on Meta's Oversight Board may help Meta improve transparency, but it is not enough to rebuild public trust in social media platforms like Meta.

Furthermore, the Board is funded via a Meta-endowed trust despite being framed as third-party oversight. This raises a governance paradox: can a mechanism be substantively independent when resourced by the entity it reviews? To avoid binary judgements, the study treats 'independence' as multi-dimensional and comprising: financial independence (endowment design and non-discretionary disbursement); organisational independence (appointment/tenure and case-selection autonomy); procedural independence (reason-giving and precedent); and epistemic independence (evidence standards and expert inputs). The analysis tested observable implications that would be costly for Meta if the Board lacked autonomy: patterns of decisions adverse to short-term platform incentives, and unedited publication of reasoning and post-decision policy shifts with measurable enforcement consequences. Thus, the standard is qualified independence, namely sufficient insulation to generate real accountability gains beyond symbolic self-legitimation.

Meta's Oversight Board blurs traditional self-regulation and co-regulation. While it is not a wholly external regulatory body, neither is it merely an internal compliance department. Meta's Oversight Board represents an institutional innovation born out of the content moderation dilemma. As a third party, the Board has established a channel for receiving user appeals. Concurrently, Meta has delegated a portion of its content moderation authority to the Board. In adjudicating cases, the Board publishes its rulings, offers detailed reasoning and thereby demonstrates transparency in the decision-making process of Meta. Participant 1 noted,

> I've looked at lots of different adjudication decisions and investigation reports from different self-regulatory agencies and government agencies. From my perspective, the decisions made by the Oversight Board stand up to any form of self-regulation and certainly could sit well as the results of some forms of co-regulation because they clearly set out what the complaint was and look at the rules that apply.

Nevertheless, the Board has its own limits, as participant 4 suggested:

> When you look at what the Board actually does, however, it actually only looks at like a handful of posts. It looks at individual posts. It's very slow. It requires a lot of slow deliberation, and it's done by a small board.

At the time of writing, Meta's Oversight Board has 21 members. These structural limits are further exposed by the scale and nature of platform governance. With around 100 million enforcement actions on content daily, even a 99% accuracy rate would still

result in approximately one million moderation errors per day (Oversight Board, 2022).

The Oversight Board plays a role in self-regulation and co-regulation. It represents an institutional response to the legitimacy deficit, strengthens accountability, and enhances transparency. However, it cannot replace co-regulation. It remains merely an experimental form within the content moderation dilemma and arises from the absence of unified international oversight. Its existence serves only to fragment the excessive concentration of platform power instead of constituting a structural reconfiguration of platform authority and regulatory frameworks.

4.7 Conclusion

This chapter has provided a definition of misinformation and discussed the large-scale moderation of misinformation by comparing and analysing study participant interview data. The operations of the Meta Oversight Board include balancing human rights and freedom of speech in decision-making. This chapter has also sought to answer the question of who should be responsible for moderating misinformation. The analysis of the study interview data has led to the finding that there are strong reasons for the government to play a big role in moderating misinformation. However, third parties like Meta's Oversight Board should assist or collaborate with other stakeholders to moderate misinformation in this age.

Chapter 5: Conclusion

5.1 Summary

Content moderation has drawn enormous attention at present. This thesis has sought to focus on the effectiveness and the operation of Meta's Oversight Board in moderating content, and whether the Board could be a solution used for regulating global platforms. This thesis aims to address the background issues regarding the establishment of Meta's Oversight Board from 2010 to 2020 in Chapter 3, and what " misinformation " is, who should be responsible for moderating misinformation in Chapter 4. This thesis also discusses the results of document analysis carried out to explore and understand the gradual loss of users' trust in Facebook between 2010 and 2020, as discussed in Chapter 3. The regulation of social media platforms is influenced by many factors. Specifically, the rise of platform power, with the increasing number of users on Facebook, as well as the large scale of user-generated content, has brought challenges to the regulation of social media platforms. The concentrated platform power, with the breach of users' privacy, has led to a decline in trust of users in Facebook. In this context, misinformation is the current issue on social media platforms, as misinformation can spread all over the globe, while it is difficult to trace. Under this background, this thesis has taken into account who should be responsible for defining and regulating misinformation, as discussed in Chapter 4. This thesis focuses on the moderation of content by Meta and the operation of Meta's Oversight Board with regard to content regulation to analyse such influences. This chapter makes an integrated analysis of the primary research findings

based on the two main discussion chapters and provides countermeasures to the sub-research questions proposed in Chapter 1. In this thesis, Meta's Oversight Board, established by Meta, serves as a regulatory-pressure-deflection mechanism that it has evolved, almost from the start, into an institution that does genuine if somewhat structurally constrained governance work. There were also many independent operation highlights during this process. In addition, in April 2025, the Oversight Board rebuked Meta for its January 2025 policy overhaul, where, as Trump was coming in, they announced they were cutting all fact-checking work, relaxing rules on immigration and gender identity content. It is worth noting that the Board described those changes as "hasty" and did not conduct appropriate human-rights due diligence or public disclosures. In fact, what is particularly illustrative is that Meta pretty much ignored them entirely and has not reversed the changes. Additionally, it is also worth noting that Kate Klonick has also soured on the Oversight Board, she said "Because even for a lot of the people on the board, it's just a very nice paycheck, and they'd rather not give up that paycheck (Newton, 2025)." And Kate Klonick become more critical about the post Jan 2025 operations of Facebook and its leadership's decisions, which seemed to be mainly aimed at placating the Trump administration.

First, chapter 3 discussed the regulatory history of Facebook during the period between 2010 and 2020. Based on an analysis of 40 events which took place on Facebook over the period, this thesis focused on the different periods of regulation, such as the thin self-regulation regime (2010-2012), the strengthened self-regulation regime (2012-2018), and the proposal put forward to establish the Meta Oversight

Board (2018-2020). Nevertheless, the regulatory history of Facebook evolved under the influences of the political and economic environment instead of taking place in a vacuum. Second, Chapter 3 discussed the imbalances in the power relationship between users and Facebook, while a broader context was provided for self-regulation, including the political and economic environment. Both can influence the regulatory effects. From an economic perspective, the discussion focused on the question of how the changing media landscape – including the transformation from media power to platform power, data-driven business and platformisation – has formed the discourse, bringing many profits to the industry players. Concerning the political environment, the ACCC report (2019) was used as a case study to discuss the co-regulation between technology companies and the national governments. In the end, Chapter 3 focused on Meta's Oversight Board, including the declining trust in Facebook (the technology and the entity), insufficient legitimacy in the regulation of Facebook, and the regulation of Facebook within a legal framework. The analysis put forward a conclusion: to regain public trust, it is essential to regulate Facebook within a legal framework. This is a feasible and effective manner to be adopted in this background.

Second, chapter 4 focused on who should be responsible for defining and moderating contemporary misinformation, and analysed the issue of trust in the misinformation age. First, the chapter started with the definition of misinformation. In view of the absence of a universally agreed-upon definition for misinformation, it is particularly crucial to build a clear definition. This chapter also applied interview data about

misinformation to explore the issue of moderating misinformation on a large scale. The points were made that millions of users are active on Meta and may post or repost misinformation every day due to the large influence, while the moderation of misinformation requires handling a large amount of information. Chapter 4 reflected on the issue of regulating misinformation in addition to exploring the future directions of the large-scale moderation of misinformation, such as automated regulation, which can be applied in later studies. The thesis aims to clearly illustrate the effect of the large-scale moderation of misinformation based on the implementation of interviews. However, self-regulation of misinformation moderation, via, for example, the Meta Oversight Board, is not enough to moderate content on social media platforms in an effective manner. It is crucial to cooperate with national governments to moderate misinformation under a legal framework, as it is hard to trace the spread of misinformation.

On the whole, the concentrated platform power and the unequal relationship between Facebook and users have led to declining trust in Facebook. The declining trust in Facebook is also influenced by the changing economic and political environment. The economic and political settings are changing due to the implementation of policies or the occurrence of certain events, which cannot be ignored. Like current issues on social media platforms, the spread of misinformation online leads to significant public concern. As a third-party which plays a partially independent role in moderating misinformation online, the Meta Oversight Board cannot pursue the goal of rebuilding public trust in social media platforms. The organisation of Meta's Oversight Board is

small. The Meta's Oversight Board cannot easily moderate content in a large number of cases. By carrying out semi-structured interviews, this study discovers that it is essential for the government to intervene in the moderation of misinformation. However, with the development of technology, generative AI also puts forward questions about the future directions of platform regulation. Specifically, the AI-generated content is changing cultural production and content moderation. Because of the fast generation speed of content and high association of existing knowledge, AI is powerful in generating content in a very fast manner and building connections of different knowledge from various fields, which may appear to be effective and encyclopaedic superficially. In this regard, the next section discusses AI-generated content, which can be taken as a way to take into account the future directions that platform regulation might take.

5.2 Future Directions for Platform Regulation Research

Pate explained that "platform manipulation refers to the activity of malicious actors using social media platforms for deceiving users" (2025, p. 874). Platform manipulation can be classified on two levels: the activity of malicious actors, and the actions of platforms themselves. In the past, social media platforms crowdsourced content creation from users. In this regard, social media platforms were unable to control the dissemination of content, although they were able to design algorithmic mechanisms to make some content visible or invisible to the public or set the agenda. Nevertheless, with the emergence of AI-generated content, social media platforms do not depend on users for content production. Instead, content is generated by

themselves 24 hours a day without interruption. As a result, social media platforms can turn into machine-manipulated media in the age of AI. This change can also lead to other issues, like deepfakes.

Generated from the application of AI or machine learning techniques, deepfake videos make a combination of images, videos, and audio. Deepfakes can display a lot of behaviours, activities not conducted by a person in fact, or something not said by him in deed (Maras & Alexandrou, 2019; Murphy & Flynn, 2022). The emergence and widespread use of deepfakes have brought new and difficult challenges to both individuals and society (Groh et al., 2022). People find it difficult to distinguish between what is real and fake when they watch deepfake videos. As social media platforms generate certain information for users, the dissemination of deepfake videos on social media platforms reduces user trust in these platforms. They do not trust what is pushed to them on these platforms. Future research could focus on AI-generated content and feasible solutions for AI to moderate content online.

This thesis has discussed Meta's Oversight Board as a third-party moderating content on Facebook, Instagram, and Threads. The Meta's Oversight Board can be deemed as a self-enhancing regulation. This serves as a response and supplementation to the platform governance triangulation of Gorwa (2019b). In addition, this thesis has situated platform governance at the intersection between technological infrastructure and data-driven business models. It means the excessive concentration has been drawn to platform power and its origins. The winner-take-all business model of social media platforms, which is coupled with the sheer scale of content, generates

formidable challenges to regulation. In this context, integration of considerations of platform governance in view of shifts in the technological infrastructure and the structure of international capitalism can further explain platform power and content moderation. This reinterpretation is based on facts of political economy. Concurrently, the selection of the contentious topic of misinformation as a litmus test shows the trade-offs between freedom and human rights when decisions are made by the Meta Oversight Board.

This thesis has discussed the current content moderation and put forward future directions of moderating information online. Much attention is paid to the regulatory history of Facebook and current issues online, in particular, the misinformation. This thesis has analysed the role played by Meta's Oversight Board in the regulation of content on Facebook, Instagram and Threads. Future research should make further investigation into other digital platforms, like X and TikTok. Not all social media platforms are covered due to the limitation of words; this thesis offers a meaningful way for searching on the platform regulation, as Meta's Oversight Board is a new type of content moderation. Future studies could further investigate other types of content moderation to facilitate platform regulation in the future.

References

Ananny, M., & Gillespie, T. (2016). Public platforms: Beyond the cycle of shocks and exceptions. *The internet, policy & politics conference*. http://blogs.oii.ox.ac.uk/ipp-conference/sites/ipp/files/documents/anannyGillespie-publicPlatforms-oii-submittedSept8.pdf

Abbott, K. W., & Snidal, D. (2009). CHAPTER TWO. The Governance Triangle: Regulatory Standards Institutions and the Shadow of the State. In W. Mattli & N. Woods (Eds.), *The Politics of Global Regulation* (pp. 44–88). Princeton University Press. https://doi.org/10.1515/9781400830732.44

Ahn, S., Baik, J. (Sophia), & Krause, C. S. (2022). Splintering and centralizing platform governance: How Facebook adapted its content moderation practices to the political and legal contexts in the United States, Germany, and South Korea. *Information, Communication & Society, 1*–20. https://doi.org/10.1080/1369118X.2022.2113817

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *The Journal of Economic Perspectives*, *31*(2), 211–235. https://doi.org/10.1257/jep.31.2.211

Andersson Schwarz, J. (2017). Platform Logic: An Interdisciplinary Approach to the Platform-Based Economy: Platform Logic. *Policy & Internet*, *9*(4), 374–394. https://doi.org/10.1002/poi3.159

Andrews, L. (2019). *Facebook, the Media and Democracy: Big Tech, Small State?*

(1st ed.). Routledge. https://doi.org/10.4324/9780429466410

Arun, Chinmayi. (2020). “The Facebook Oversight Board: An Experiment in Self-Regulation.” Just Security(blog). May 6, 2020. https://www.justsecurity.org/70021/the-facebook-oversight-board-an-experiment-in-self-regulation/

Atal, M. R. (2021). The Janus faces of Silicon Valley. *Review of International Political Economy, 28*(2), 336–350. https://doi.org/10.1080/09692290.2020.1830830

Australian Competition and Consumer Commission (ACCC) (2019) *Digital Platforms Inquiry: Final Report*. Canberra, ACT, Australia: ACCC.

Balkin, J. M. (2018). FREE SPEECH IS A TRIANGLE. *Columbia Law Review*, 118(7), 2011–2056.

Bélair-Gagnon, V., Graves, L., Kalsnes, B., Steensen, S., & Westlund, O. (2022). Considering Interinstitutional Visibilities in Combating Misinformation. *Digital Journalism*, *10*(5), 669–678. https://doi.org/10.1080/21670811.2022.2072923

Bellanova, R., Carrapico, H., & Duez, D. (2022). Digital/sovereignty and European security integration: An introduction. *European Security*, *31*(3), 337–355. https://doi.org/10.1080/09662839.2022.2101887

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Bennett, W. L., & Livingston, S. (Eds.). (2020). *The Disinformation Age: Politics, Technology, and Disruptive Communication in the United States* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108914628

Bennett, W. L., & Kneuer, M. (2024). Communication and democratic erosion: The rise of illiberal public spheres. *European Journal of Communication (London)*, *39*(2), 177–196. https://doi.org/10.1177/02673231231217378

Bickert, M. (2019, October 14). *European Court Ruling Raises Questions About Policing Speech*. Meta. https://about.fb.com/news/2019/10/european-court-ruling-raises-questions-about-policing-speech/

Bickert, M. (2020, January 6). *Enforcing Against Manipulated Media.* Meta. https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/

Black, J. (2005). What is Regulatory Innovation? In J. Black, M. Lodge, & M. Thatcher, *Regulatory Innovation* (p. 3769). Edward Elgar Publishing. https://doi.org/10.4337/9781845427979.00007

Bodó, B. (2021). Mediated trust: A theoretical framework to address the trustworthiness of technological trust mediators. *New Media & Society, 23*(9), 2668–2690. https://doi.org/10.1177/1461444820939922

Booth,R. (2025, January 8). *Meta to get rid of factcheckers and recommend more political content.* The Guardian. https://www.theguardian.com/technology/2025/jan/07/meta-facebook-instagram-threads-mark-zuckerberg-remove-fact-checkers-recommend-political-content

Bossio, D., Flew, T., Meese, J., Leaver, T., & Barnet, B. (2022). Australia's News Media Bargaining Code and the global turn towards platform regulation. *Policy & Internet*, *14*(1), 136–150. https://doi.org/10.1002/poi3.284

Botsman, R. (2017). *Who can you trust?: how technology brought us together : and why it could drive us apart.* Portfolio/Penguin.

Boudana, S., & Segev, E. (2024). Fake News Makes the News: Definitions and Framing of Fake News in Mainstream Media. *Journalism Practice*, 1–20. https://doi.org/10.1080/17512786.2024.2379898

Bowen, G. A. (2009). Document Analysis as a Qualitative Research Method. *Qualitative Research Journal, 9*(2), 27–40. https://doi.org/10.3316/QRJ0902027

Bowers, J., & Zittrain, J. (2020). Answering Impossible Questions: Content Governance in an Age of Disinformation. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-005

Briant, E. L., & Bakir, V. (2024). *Routledge Handbook of the Influence Industry (1st edn).* Routledge. https://doi.org/10.4324/9781003256878

Bucher, T. (2021). *Facebook*. Polity Press.

Burrell J (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society 3*(1): 2053951715622512.

Calderaro, A., & Blumfelde, S. (2022). Artificial intelligence and EU security: The false promise of digital sovereignty. *European Security*, *31*(3), 415–434. https://doi.org/10.1080/09662839.2022.2101885

Castells, M. (2013). *Communication power.* (2nd ed.). OUP Oxford.

Cathcart, W. (2017, January 25). *Continuing Our Updates to Trending*. Meta. https://about.fb.com/news/2017/01/continuing-our-updates-to-trending/

Calif, M. P. (2012, May 17). *Facebook Announces Pricing of Initial Public Offering*. Meta. https://about.fb.com/news/2012/05/facebook-announces-pricing-of-initial-public-offering/

Castells, M. (1997). An introduction to the information age. *City, 2*(7), 6–16. https://doi.org/10.1080/13604819708900050

Cherubini, F., & Nielsen, R. K. (2016). Editorial Analytics: How News Media are Developing and Using Audience Data and Metrics. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.2739328

Chin, Y. C., Park, A., & Li, K. (2022). A comparative study on false information governance in Chinese and American social media platforms. *Policy & Internet*, *14*(2), 263–283. https://doi.org/10.1002/poi3.301

Clegg, N. (2019, August 1). Smart regulation can deliver a better Internet for all Australians. *Sydney Morning Herald.* https://www.smh.com.au/business/companies/smart-regulation-can-deliver-a-betterinternet-for-all-australians-20190731-p52cm5.html

Clegg, N. (2020, May 6). *Welcoming the Oversight Board,* Meta. https://about.fb.com/news/2020/05/welcoming-the-oversight-board/

Coglianese, C., & Mendelson, E. (2010). Meta-Regulation and Self-Regulation. In M. Cave, M. Lodge, & R. Baldwin (Eds.), *The Oxford Handbook of Regulation*.

Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199560219.003.0008

Coleman, S. (2012). Believing the news: From sinking trust to atrophied efficacy. *European Journal of Communication, 27*(1), 35–45. https://doi.org/10.1177/0267323112438806

Cusumano, M. A., Gawer, A., & Yoffie, D. B. (2021). Can self-regulation save digital platforms? *Industrial and Corporate Change*, *30*(5), 1259–1285. https://doi.org/10.1093/icc/dtab052

Das, T. K., & Teng, B. S. (2004). *The risk-based view of trust: A conceptual framework. Journal of Business and Psychology, 19*(1), 85-116.

Declining Trust in Social Media Giants Affecting Consumer Purchase Decisions, New Data Released By Outbrain Ahead of Cyber Weekend Shows: Retail spend on digital ads set to surge 21% this year as marketers adapt to consumer behavior shift. (2020, Sep 02). *PR Newswire* https://www.proquest.com/wire-feeds/declining-trust-social-media-giants-affecting/docview/2439229454/se-2

Dijck, J. van (2013). *The Culture of Connectivity: A Critical History of Social Media.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199970773.001.0001

Dijck, J. van, Poell, T., & Waal, M. de. (2018). *The platform society.* Oxford University Press. http://dx.doi.org/10.1093/oso/9780190889760.001.0001

Dijck, J. van, Nieborg, D., & Poell, T. (2019). Reframing platform power. *Internet Policy Review*, *8*(2). https://doi.org/10.14763/2019.2.1414

Easton, W. (2020, August 31). *An Update About Changes to Facebook's Services in Australia*. Meta. https://about.fb.com/news/2020/08/changes-to-facebooks-services-in-australia/

Encinas Duarte, G. A. (2025). Interlegal argumentation in the UK Drill Music decision of Meta's Oversight Board. *Journal of Argumentation in Context, 14*(1), 3–39. https://doi.org/10.1075/jaic.24004.enc

Evans, D.S., A. Hagiu, and R. Schmalensee. 2006. *Invisible Engines: How Software Platforms Drive Innovation and Transform Industries.* Cambridge, MA: MIT Press.

Evens, T., & Donders, K. (2020). Regulating digital platform power. *Journal of Digital Media & Policy*, *11*(3), 235–239. https://doi.org/10.1386/jdmp_00024_2

Everett, C. M. (2018). FREE SPEECH ON PRIVATELY-OWNED FORA: A DISCUSSION ON SPEECH FREEDOMS AND POLICY FOR SOCIAL MEDIA. *The Kansas Journal of Law & Public Policy*, *28*(1), 113-145.

Fabrizio Di Mascio, Barbieri, M., Natalini, A., & Selva, D. (2021). Covid-19 and the Information Crisis of Liberal Democracies: Insights from Anti-Disinformation Action in Italy and EU. *Partecipazione e Conflitto*, *14*(1), 221–240. https://doi.org/10.1285/i20356609v14i1p221

Facebook. (2012). *Facebook Annual Report 2012.* https://s21.q4cdn.com/399680738/files/doc_financials/annual_reports/FB_2012_10K.pdf

Faris, R., & Donovan, J. (2021). The Future of Platform Power: Quarantining Misinformation. *Journal of Democracy*, *32*(3), 152–156. https://doi.org/10.1353/jod.2021.0040

Fasel, M., & Weerts, S. (2024). Can Facebook's community standards keep up with legal certainty? Content moderation governance under the pressure of the Digital Services Act. *Policy and Internet*. https://doi.org/10.1002/poi3.391

Feick, J., & Werle, R. (2010). Regulation of Cyberspace. In R. Baldwin, M. Cave, & M. Lodge (Eds.), *The Oxford Handbook of Regulation* (pp. 522–547). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199560219.003.0021

Filippi, P. D., & Hassan, S. (2016). Blockchain technology as a regulatory technology: From code is law to law is code. *First Monday*. https://doi.org/2013

Flew, T. (2018a). *Understanding global media* (Second edition.). Palgrave.

Flew, T. (2018b). Platforms on Trial. *Intermedia. 46* (2): 24–29.

Flew, T. (2019), 'The platformized internet: Issues for internet law and policy', *Journal of Internet Law, 22*:11, pp. 4–16.

Flew, T. (2021a). *Regulating platforms*. Polity Press.

Flew, T. (2021b). Fake News, Trust and Behaviour in a Digital World. In Gary D. Rawnsley, Yiben Ma and Kruakae Pothong (Eds.), *Research Handbook on Political Propaganda*, (pp. 28-40). Cheltenham: Edward Elgar Publishing.

Flew, T. (2022). The Challenge of Trust in Digital Societies: Digital Platforms and New Public Spheres. *SSRN Electronic Journal*.

https://doi.org/10.2139/ssrn.4151098

Flew, T. (2023). Global Internet Governance in a Post-Global Age. In T. Flew, J. Thomas, & J. Holt (Eds.), *The SAGE Handbook of the Digital Media Economy* (pp. 3–28). SAGE Publications Ltd. https://doi.org/10.4135/9781529757170.n3

Flew, T. (2024). Mediated trust, the internet and artificial intelligence: Ideas, interests, institutions and futures. *Policy & Internet, 16*(2), 443–457. https://doi.org/10.1002/poi3.390

Flew, T., & Gillett, R. (2021). Platform policy: Evaluating different responses to the challenges of platform power. *Journal of Digital Media & Policy*, *12*(2), 231–246. https://doi.org/10.1386/jdmp_00061_1

Flew, T., Gillett, R., Martin, F., & Sunman, L. (2021). Return of the regulatory state: A stakeholder analysis of Australia's Digital Platforms Inquiry and online news policy. *The Information Society*, *37*(2), 128–145. https://doi.org/10.1080/01972243.2020.1870597

Flew, T., & Jiang, Y. (2021). Trust and Communication: Looking Back, Looking Forward. *Global Perspectives, 2*(1), 25395. https://doi.org/10.1525/gp.2021.25395

Flew, T., & Lin, F. (2022). The third way of global Internet governance: A dialogue with Terry Flew. *Communication and the Public*, *7*(3), 121–129. https://doi.org/10.1177/20570473221123150

Flew, T., & Wilding, D. (2021). The turn to regulation in digital communication: the

ACCC's digital platforms inquiry and Australian media policy. *Media, Culture & Society, 43*(1), 48–65. https://doi.org/10.1177/0163443720926044

Freedman, D. (2008). *The Politics of Media Policy.* Polity Press.

Freedman, D. (2010). Media Policy Silences: The Hidden Face of Communications Decision Making. *The International Journal of Press/Politics*, *15*(3), 344–361. https://doi.org/10.1177/1940161210368292

Freedman, D. (2013). *The politics of media policy*. Wiley.

Freedman, D. (2014). *The Contradictions of Media Power* (1st ed.). Bloomsbury Publishing (UK). https://doi.org/10.5040/9781849661089

Galletta, A. (2013). *Mastering the Semi-Structured Interview and Beyond: From Research Design to Analysis and Publication* (1st ed., Vol. 18). NYU Press. https://doi.org/10.18574/9780814732953

Gallup. (2020, March). *Techlash? America's Growing Concern With Major Technology Companies*. https://knightfoundation.org/wp-content/uploads/2020/03/Gallup-Knight-Report-Techlash-Americas-Growing-Concern-with-Major-Tech-Companies-Final.pdf

Gallup. (2021, November 18). *Young People Rely on Social Media, but Don't Trust It*. https://news.gallup.com/opinion/gallup/357446/young-people-rely-social-media-don-trust.aspx

Gerlitz, C., & Helmond, A. (2013). The like economy: Social buttons and the data-intensive web. *New Media & Society, 15*(8), 1348–1365. https://doi.org/10.1177/1461444812472322

Gleicher, N. (2019, December 20). *Removing Coordinated Inauthentic Behavior From Georgia, Vietnam and the US.* Meta. https://about.fb.com/news/2019/12/removing-coordinated-inauthentic-behavior-from-georgia-vietnam-and-the-us/

Gillespie, T. (2010). The politics of 'platforms'. *New Media & Society, 12*(3), 347–364. https://doi.org/10.1177/1461444809342738

Gillespie, T. (2014). The Relevance of Algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media Technologies* (pp. 167–194). The MIT Press. https://doi.org/10.7551/mitpress/9780262525374.003.0009

Gillespie, T. (2016). Algorithms, clickworkers, and the befuddled fury around Facebook Trends. *Social Media Collective.* May 18. Retrieved from https://socialmediacollective.org/2016/05/18/facebook-trends/

Gillespie, T. (2018). Regulation of and by Platforms. In J. Burgess, A. Marwick, & T. Poell, *The SAGE Handbook of Social Media* (pp. 254–278). SAGE Publications Ltd. https://doi.org/10.4135/9781473984066.n15

Gillespie, T. (2019). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press. https://doi.org/10.12987/9780300235029

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society, 7*(2), 205395172094323. https://doi.org/10.1177/2053951720943234

Gillespie, T. (2022). Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society, 8*(3), 205630512211175.

https://doi.org/10.1177/20563051221117552

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

Godding, C. (2020), 'May market pulse: How will the economy change after lockdown relaxation?', *Tilney Market News*, 26 May, https://www.tilney.co.uk/news/may-market-pulse-how-will-the-economy-changeafter-lockdown-relaxation

Goldsmith, J. L., & Wu, T. (2006). *Who Controls the Internet? Illusions of a Borderless World.* New York: Oxford University Press.

Gorwa, R. (2019a). What is platform governance? *Information, Communication & Society*, 22(6), 854–871. https://doi.org/10. 1080/1369118X.2019.1573914

Gorwa, R. (2019b). The platform governance triangle: conceptualising the informal regulation of online content. *IDEAS Working Paper Series from RePEc*. https://doi.org/10.31219/osf.io/tgnrj

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society, 7*(1), 205395171989794. https://doi.org/10.1177/2053951719897945

Graham, T. (2024, October 4). Is big tech harming society? To find out, we need

research – but it's being manipulated by big tech itself. *The Conversation*. https://theconversation.com/is-big-tech-harming-society-to-find-out-we-need-research-but-its-being-manipulated-by-big-tech-itself-240110

Grewal, P. (2018, March 16). *Suspending Cambridge Analytica and SCL Group From Facebook*. Meta. https://about.fb.com/news/2018/03/suspending-cambridge-analytica/

Gritsenko, D., & Wood, M. (2022). Algorithmic governance: A modes of governance approach. *Regulation & Governance, 16*(1), 45–62. https://doi.org/10.1111/rego.12367

Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, *119*(1), e2110013119. https://doi.org/10.1073/pnas.2110013119

Haggart, B. (2020). Global platform governance and the internet-governance impossibility theorem. *Journal of Digital Media & Policy*, 11(3), 321–339. https://doi.org/10.1386/jdmp_00028_1

Haggart, B., & Keller, C. I. (2021). Democratic Legitimacy in Global Platform Governance. Telecommunications Policy, 45(6). Norm Entrepreneurship in Internet Governance (July 1), 102152.

Halpern, S. (2020, October 15). The ad-hoc group of activists and academics convening a 'real Facebook oversight board'. *The New Yorker.* https://www.newyorker.com/tech/annals-of-technology/the-ad-hoc-group-of-

activists-and-academics-convening-a-real-facebook-oversight-board

Harris, B. (2019a, December 12). *An Update on Building a Global Oversight Board*. Meta. https://about.fb.com/news/2019/12/oversight-board-update/

Harris, B. (2019b, September 17). *Establishing Structure and Governance for an Independent Oversight Board,* Meta. https://about.fb.com/news/2019/09/oversight-board-structure/

Harris, B. (2020a, October 22). *Oversight Board to Start Hearing Cases*, Meta. https://about.fb.com/news/2020/10/oversight-board-to-start-hearing-cases/

Harris, B. (2020b, December 1). *Oversight Board Selects First Cases to Review*, Meta. https://about.fb.com/news/2020/12/oversight-board-selects-first-cases-to-review/

Hasenfeld, Y., & Brock, T. (1991). Implementation of social policy revisited. *Administration & Society, 22*(4), 451-479.

Hegeman, J. (2018, May 23). *Facing Facts: Facebook's Fight Against Misinformation.* Meta. https://about.fb.com/news/2018/05/facing-facts-facebooks-fight-against-misinformation/

Helmond, A. (2015). The Platformization of the Web: Making Web Data Platform Ready. *Social Media + Society, 1*(2). https://doi.org/10.1177/2056305115603080

Helmond, A., Nieborg, D. B., & Van Der Vlist, F. N. (2019). Facebook's evolution: Development of a platform-as-infrastructure. *Internet Histories, 3*(2), 123–146. https://doi.org/10.1080/24701475.2019.1593667

Helberger, N. (2020). **“**The Political Power of Platforms: How Current Attempts to Regulate Platforms Fail to Address the Risks to Democracy**”** *Philosophy & Technology, 33*(4), 1–8.

Helberger, N., Pierson, J., & Poell, T. (2018). Governing online platforms: From contested to cooperative responsibility. *The Information Society, 34*(1), 1–14. https://doi.org/10.1080/01972243.2017.1391913

Hendricks, V. F., & Vestergaard, M. (2019). Alternative Facts, Misinformation, and Fake News. In V. F. Hendricks & M. Vestergaard, *Reality Lost* (pp. 49–77). Springer International Publishing. https://doi.org/10.1007/978-3-030-00813-0_4

Hofmann, J., Katzenbach, C., & Gollatz, K. (2017). Between coordination and regulation: Finding the governance in Internet governance. *New Media & Society*, *19*(9), 1406–1423. https://doi.org/10.1177/1461444816639975

Hosking, G. (2014). *The Coconut Tree: The Ups and Downs of Trust*. In Trust. Oxford University Press, Incorporated.

Hurni, T., Huber, T. L., & Dibbern, J. (2022). Power dynamics in software platform ecosystems. *Information Systems Journal (Oxford, England), 32*(2), 310–343. https://doi.org/10.1111/isj.12356

Issar, S., & Aneesh, A. (2022). What is algorithmic governance? *Sociology Compass*, *16*(1). https://doi.org/10.1111/soc4.12955

Just, N., & Latzer, M. (2017). Governance by algorithms: Reality construction by algorithmic selection on the Internet. *Media, Culture & Society, 39*(2), 238–

258. https://doi.org/10.1177/0163443716643157

Kaloudis, M. (2022). Sovereignty in the Digital Age – How Can We Measure Digital Sovereignty and Support the EU's Action Plan? *New Global Studies*, *16*(3), 275–299. https://doi.org/10.1515/ngs-2021-0015

Kacholia, V. (2013, August 23). *Showing More High Quality Content,* Meta. https://about.fb.com/news/2013/08/news-feed-fyi-showing-more-high-quality-content/

Kacperska, E., Łukasiewicz, K., & Horin, N. (2022). Building digital trust in social media during crisis. In J. Paliszkiewicz, J. L. Guerrero Cusumano, & J. Gołuchowski, *Trust, Digital Business and Technology* (1st edn, pp. 49–60). Routledge. https://doi.org/10.4324/9781003266495-5

Kaplan, J. & Osofsky, J. (2016, October 21). *Input From Community and Partners On Our Community Standards.* Meta. https://about.fb.com/news/2016/10/input-from-community-and-partners-on-our-community-standards/

Kaplan, J. (2017, October 2). *Improving Enforcement and Transparency of Ads on Facebook*. Meta. https://about.fb.com/news/2017/10/improving-enforcement-and-transparency/

Kelly, H. & Guskin, E. (2021, December 22). Americans widely distrust Facebook, TikTok and Instagram with their data, poll finds, *The Washington Post.* https://www.washingtonpost.com/technology/2021/12/22/tech-trust-survey/

Kim, G., & Lee, J. Y. (2020). Digital Trust Gap: The differences in perceptions of trust between experienced and inexperienced users. *Journal of*

Telecommunications and the Digital Economy, *8*(2), 94–109. https://doi.org/10.18080/jtde.v8n2.237

Klonick, K. (2018). The new governors: The people, rules, and processes governing online speech. Harvard Law Review, 131(6), 1598--1670.

Kvale, S. (2008). Introduction to interview research. In *Doing Interview*s (pp. 1–10). SAGE Publications. https://doi.org/10.4135/9781849208963.n1

Kvale, S., & Brinkmann, S. (2015). *InterViews : learning the craft of qualitative research interviewing (Third edition.).* Sage Publications.

Langley, P., & Leyshon, A. (2017). Platform capitalism: The intermediation and capitalization of digital economic circulation. *Finance and Society, 3*(1), 11–31. https://doi.org/10.2218/finsoc.v3i1.1936

Langlois G, Elmer G, McKelvey F, et al. (2009) Networked publics: the double articulation of code and politics on Facebook. *Canadian Journal of Communication 34*: 415–434.

Leerssen, P. (2025). From Murdoch to Musk: Platform ownership and the political economy of online content governance. *Platforms & Society*, *2*, 29768624251386260. https://doi.org/10.1177/29768624251386260

Lessig, L. (2006). *Code : version 2.0* (2nd ed.). BasicBooks.

Liu, B. F., & Mehta, A. M. (2024). *Routledge Handbook of Risk, Crisis, and Disaster Communication (1st edn).* Routledge. https://doi.org/10.4324/9781003363330

Luhmann, N. (1979). *Trust and power*. Chichester: John Wiley & Co.

Lunt, P. K. , & Livingstone, S. M. (2012). *Media regulation : governance and the*

interest of citizens and consumers. Sage.

Lyons, T. (2018, May 23). *Hard Questions: What's Facebook's Strategy for Stopping False News?*, Meta. https://about.fb.com/news/2018/05/hard-questions-false-news/

Magalhães J, Keller C and Gorwa R (2025) *The Great Sysop: Elon Musk, X, and the Emergence of Platform Illiberalism.* OSF. https://doi.org/10.31235/osf.io/6grbc_v2

Makavy, R. (2013, July 22). *Feature Phone Milestone: Facebook for Every Phone Reaches 100 Million.* Meta. https://about.fb.com/news/2013/07/feature-phone-milestone-facebook-for-every-phone-reaches-100-million/

Mansell, R. (2015). The public's interest in intermediaries. *Info*, *17*(6), 8–18. https://doi.org/10.1108/info-05-2015-0035

Maras, M.-H., & Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. *The International Journal of Evidence & Proof*, *23*(3), 255–262. https://doi.org/10.1177/1365712718807226

Martens, B. (2016). An Economic Policy Perspective on Online Platforms. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2783656

Mayntz, R. (2004). Governance im modernen Staat. In A. Benz (Ed.), *Governance—Regieren in komplexen Regelsystemen.* Eine Einf¨ uhrung (pp. 65–76). Wiesbaden, Germany: VS Verlag.

McAfee, A.P. (2009), "Shattering the myths about enterprise 2.0", *Harvard Business*

Review, Vol. 87 No. 11, pp. 1-6.

McCay-Peet, L., & Quan-Haase, A. (2016). What is Social Media and What Questions Can Social Media Research Help Us Answer? In L. Sloan & A. Quan-Haase, *The SAGE Handbook of Social Media Research Methods* (pp. 13–26). SAGE Publications Ltd. https://doi.org/10.4135/9781473983847.n2

McKnight, D. H., & Chervany, N. L. (2002). What trust means in e-commerce customer relationships: An interdisciplinary conceptual typology. *International Journal of Electronic Commerce, 6*(2), 35-59.

Medzini, R. (2022). Enhanced self-regulation: The case of Facebook's content governance. *New Media & Society*, *24*(10), 2227–2251. https://doi.org/10.1177/1461444821989352

Meta. (2012a, April 9). *Facebook to Acquire Instagram*. https://about.fb.com/news/2012/04/facebook-to-acquire-instagram/

Meta. (2012b, August 2). *Introducing Facebook Stories.* https://about.fb.com/news/2012/08/introducing-facebook-stories/

Meta. (2013, August 21). *Technology Leaders Launch Partnership to Make Internet Access Available to All*. https://about.fb.com/news/2013/08/technology-leaders-launch-partnership-to-make-internet-access-available-to-all/

Meta. (2014a, February 19). *Facebook to Acquire WhatsApp*. https://about.fb.com/news/2014/02/facebook-to-acquire-whatsapp/

Meta. (2014b, March 25). *Facebook to Acquire Oculus*. https://about.fb.com/news/2014/03/facebook-to-acquire-oculus/

Meta. (2017a, June 26). *Facebook, Microsoft, Twitter and YouTube Announce Formation of the Global Internet Forum to Counter Terrorism.* https://about.fb.com/news/2017/06/global-internet-forum-to-counter-terrorism/

Meta. (2017b, November 22). *Continuing Transparency on Russian Activity*. https://about.fb.com/news/2017/11/continuing-transparency-on-russian-activity/

Meta. (2017c, December 4). *Update on the Global Internet Forum to Counter Terrorism*. https://about.fb.com/news/2017/12/update-on-the-global-internet-forum-to-counter-terrorism/

Meta. (2018a, March 21). *Cracking Down on Platform Abuse.* https://about.fb.com/news/2018/03/cracking-down-on-platform-abuse/

Meta. (2018b, June 21). *How People Help Fight False News*. https://about.fb.com/news/2018/06/inside-feed-how-people-help-fight-false-news/

Meta. (2019a, January 28). *Draft Charter: An Oversight Board for Content Decisions*. file:///C:/Users/Rumen/Desktop/draft-charter-oversight-board-for-content-decisions-2-1.pdf

Meta. (2019b, December 12). *Ready for California's New Privacy Law*. https://about.fb.com/news/2019/12/californias-new-privacy-law/

Meta. (2019c, November 7). *How Facebook Has Prepared for the 2019 UK General Election*. https://about.fb.com/news/2019/11/how-facebook-is-prepared-for-the-2019-uk-general-election/

Meta. (2020a, January 15). *Facebook Disaster Maps Help Those Affected by Australia's Bushfires.* https://about.fb.com/news/2020/01/facebook-disaster-maps-help-those-affected-by-australias-bushfires/

Meta. (2020b, August 20). *Facebook Files Official Comments on Data Portability with Federal Trade Commission.* https://about.fb.com/news/2020/08/comments-on-data-portability/

Mitchell, A. (2014, April 24). *Announcing FB Newswire, Powered by Storyful.* Meta. https://about.fb.com/news/2014/04/announcing-fb-newswire-powered-by-storyful/

Moore, M., & Tambini, D. (2018). *Digital dominance: the power of Google, Amazon, Facebook, and Apple.* Oxford University Press.

Morgan, B. (2014, March 7). *Announcing the Public Content Solutions Program*. Meta. https://about.fb.com/news/2014/03/announcing-the-public-content-solutions-program/

Morris, D. S., Morris, J. S., & Francia, P. L. (2020). A fake news inoculation? Fact checkers, partisan identification, and the power of misinformation. *Politics, Groups & Identities*, *8*(5), 986–1005. https://doi.org/10.1080/21565503.2020.1803935

Mosseri, A. (2017, April 6). *A New Educational Tool Against Misinformation*, Meta. https://about.fb.com/news/2017/04/a-new-educational-tool-against-misinformation/

Mueller, M. L., & Farhat, K. (2022). Regulation of platform market access by the

United States and China: Neo-mercantilism in digital services. *Policy & Internet, 14*(2), 348–367. https://doi.org/10.1002/poi3.305

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, 37(11), 1905–1922. https://doi.org/10.1080/00140139408964957

Murphy, G., & Flynn, E. (2022). Deepfake false memories. *Memory*, *30*(4), 480–492. https://doi.org/10.1080/09658211.2021.1919715

Napoli, P. M. (2015). Social media and the public interest: Governance of news platforms in the realm of individual and algorithmic gatekeepers. *Telecommunications Policy, 39*(9), 751–760. https://doi.org/10.1016/j.telpol.2014.12.003

Napoli, P. M. (2019). *Reviving the Public Interest.* In Social Media and the Public Interest (pp. 163–198). Columbia University Press. https://doi.org/10.7312/napo18454-008

Nash, V., Bright, J., Margetts, H., & Lehdonvirta, V. (2017). Public Policy in the Platform Society. *Policy & Internet, 9*(4), 368–373. https://doi.org/10.1002/poi3.165

Newton, C. (2025, December 4). Where Meta's biggest experiment in governance went wrong. *Platformer.* https://www.platformer.news/meta-oversight-board-5-years/

Nieborg, D. B., & Helmond, A. (2019). The political economy of Facebook's platformization in the mobile ecosystem: Facebook Messenger as a platform

instance. *Media, Culture & Society*, *41*(2), 196–218. https://doi.org/10.1177/0163443718818384

Nieborg, D. B., & Poell, T. (2018). The platformization of cultural production: Theorizing the contingent cultural commodity. *New Media & Society*, *20*(11), 4275–4292. https://doi.org/10.1177/1461444818769694

Nielsen, R. K., & Ganter, S. A. (2022). Platform Power. In R. K. Nielsen & S. A. Ganter, *The Power of Platforms* (pp. 157–188). Oxford University Press. https://doi.org/10.1093/oso/9780190908850.003.0005

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism.* New York University Press.

Nuccio, M., & Guerzoni, M. (2019). Big data: Hell or heaven? Digital platforms and market power in the data-driven economy. *Competition & Change*, *23*(3), 312–328. https://doi.org/10.1177/1024529418816525

Ogus, A., 'Rethinking Self-Regulation', in Robert Baldwin, Colin Scott, and Christopher Hood (eds), A Reader on Regulation, *Oxford Readings in Socio-Legal Studies* (Oxford, 1998)

O'Hara, K. (2018). *Four internets: the geopolitics of digital governance* (Vol. 63/2019). Centre for International Governance Innovation.

O'Neill, O. (2014). Trust, Trustworthiness, and Accountability. In *Capital Failure*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198712220.003.0008

O'Neill, O. (2020). Trust and Accountability in a Digital Age. *Philosophy*, *95*(1), 3–

17. https://doi.org/10.1017/S0031819119000457

Osofsky, J. (2016, May 12). *Information About Trending Topics*, Meta. https://about.fb.com/news/2016/05/information-about-trending-topics/

Oversight Board. (2022 December 6), *META'S CROSS-CHECK PROGRAM,* https://www.oversightboard.com/decision/pao-nr730ofi/

Oversight Board. (2023, February). *Oversight Board charter*. file:///C:/Users/Rumen/Downloads/charter_february_2023.pdf

Parker, Geoffrey, Alstyne, Marshall Van and Paul Choudary, Sangeet (2016), *Platform Revolution: How Networked Markets Are Transforming the Economy*, New York: W. W. Norton & Co.

Pate, S. (2025). Platform liability for platform manipulation. *Columbia Law Review*, 125(4), 873–924.

Poell, T., Nieborg, D. B., & Duffy, B. E. (2023). Spaces of Negotiation: Analyzing Platform Power in the News Industry. *Digital Journalism*, *11*(8), 1391–1409. https://doi.org/10.1080/21670811.2022.2103011

Poell, T., Nieborg, D., & Dijck, J. van (2019). Platformisation. *Internet Policy Review*, *8*(4). https://doi.org/10.14763/2019.4.1425

Popiel, P. (2020). Addressing platform power: The politics of competition policy. *Journal of Digital Media & Policy*, *11*(3), 341–360. https://doi.org/10.1386/jdmp_00029_1

Popiel, P. (2022). Regulating datafication and platformization: Policy silos and tradeoffs in international platform inquiries. *Policy & Internet, 14*(1), 28–46.

https://doi.org/10.1002/poi3.283

Popiel, P., & Sang, Y. (2021). Platforms' Governance: Analyzing Digital Platforms' Policy Preferences. *Global Perspectives*, *2*(1), 19094. https://doi.org/10.1525/gp.2021.19094

Puppis, M. (2008). National Media Regulation in the Era of Free Trade: The Role of Global Media Governance. *European Journal of Communication*, *23*(4), 405–424. https://doi.org/10.1177/0267323108096992

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

Purtill J. (2024, March 7). Facebook ate and then ignored the news industry. It's hard, but we should leave it be. *ABC News.* https://amp.abc.net.au/article/103554982

Riemer, K., & Peter, S. (2021). Algorithmic audiencing: Why we need to rethink free speech on social media. *Journal of Information Technology, 36*(4), 409–426. https://doi.org/10.1177/02683962211013358

Romero, J. (2019, December 5). *Taking Action Against Ad Fraud*. Meta. https://about.fb.com/news/2019/12/taking-action-against-ad-fraud/

Sandberg, S. & Goler, L. (2017, December 8). *Sharing Facebook's Policy on Sexual Harassment.* Meta. https://about.fb.com/news/2017/12/sharing-facebooks-policy-on-sexual-harassment/

Sandberg, S. (2019, June 30). *A Second Update on Our Civil Rights Audit.* Meta.

https://about.fb.com/news/2019/06/second-update-civil-rights-audit/

Sander, B. (2020). Freedom of expression in the age of online platforms—The promise and pitfalls of a human rights-based approach to content moderation. Fordham International Law Journal, 43(4), 939–1006. https://ir.lawnet.fordham.edu/ilj/vol43/iss4/3

Santos, A., Cazzamatta, R., & Napolitano, C. J. (2025). Holding platforms accountable in the fight against misinformation: A cross-national analysis of state-established content moderation regulations. *International Communication Gazette*, 17480485251348550. https://doi.org/10.1177/17480485251348550

San Martín, P. (2023). Meta's Oversight Board: Challenges of Content Moderation on the Internet. *Erasmus Law Review, 16*(2), 124–139. https://doi.org/10.5553/ELR.000253

Schlesinger, P. (2020). After the post-public sphere. *Media, Culture & Society*, *42*(7–8), 1545–1563. https://doi.org/10.1177/0163443720948003

Schrage, E. & Ginsberg, D. (2018, April 9). *Facebook Launches New Initiative to Help Scholars Assess Social Media's Impact on Elections.* Meta. https://about.fb.com/news/2018/04/new-elections-initiative/

Schroepfer, M. (2015, November 3). *New Milestones in Artificial Intelligence Research.* Meta. https://about.fb.com/news/2015/11/new-milestones-in-artificial-intelligence-research/

Schroepfer, M. (2018, April 4). *An Update on Our Plans to Restrict Data Access on*

Facebook. Meta. https://about.fb.com/news/2018/04/restricting-data-access/

Schuilenburg, M., & Peeters, R. (2021). *The algorithmic society: Technology, power, and knowledge*. Routledge.

Schwarz, O. (2019). Facebook Rules: Structures of Governance in Digital Capitalism and the Control of Generalized Social Capital. *Theory, Culture & Society*, *36*(4), 117–141. https://doi.org/10.1177/0263276419826249

Segreti, G. (2016, August 30). *Facebook CEO says group will not become a media company*. Reuters. https://www.reuters.com/article/technology/facebook-ceo-says-group-will-not-become-a-media-company-idUSKCN1141WM

Sethuraman, R. (2019, March 31). *Why Am I Seeing This? We Have an Answer for You*, Meta. https://about.fb.com/news/2019/03/why-am-i-seeing-this/

Shapiro, Carl (2018), 'Antitrust in a time of populism', *International Journal of Industrial Organization, 61*, pp. 714–48.

Shapiro, E. (2023, May 1). MEDIA UNIVERSE MAPS 2020-2024, *Substack*. https://eshap.substack.com/p/media-universe-maps-2020-2023

Shapiro, Jesse M., Natalia Rigol, and Benjamin N. Roth. "Independent Governance of Meta's Social Spaces: The Oversight Board." *Harvard Business School,* Teaching Note 823-126, May 2023.

Siapera, E. (2022). Platform Governance and the "Infodemic". Javnost - *The Public, 29*(2), 197–214. https://doi.org/10.1080/13183222.2022.2042791

Silverman, H. (2019a, April 10). T*he Next Phase in Fighting Misinformation?*, Meta. https://about.fb.com/news/2019/04/tackling-more-false-news-more-quickly/

Silverman, H. (2019b, December 17). *Helping Fact-Checkers Identify False Claims Faster*. Meta. https://about.fb.com/news/2019/12/helping-fact-checkers/

Smyrnaios, N., & Baisnée, O. (2023). Critically understanding the platformization of the public sphere. *European Journal of Communication*, *38*(5), 435–445. https://doi.org/10.1177/02673231231189046

Söllner, M., Hoffmann, A., & Leimeister, J. M. (2016). Why different trust relationships matter for information systems users. *European Journal of Information Systems*, *25*(3), 274–287. https://doi.org/10.1057/ejis.2015.17

Sonderby, C. (2017, December 18). *Reinforcing Our Commitment to Transparency*. Meta. https://about.fb.com/news/2017/12/reinforcing-our-commitment-to-transparency/

Sonderby, C. (2018, May 15). *Reinforcing Our Commitment to Transparency*. Meta. https://about.fb.com/news/2018/05/transparency-report-h2-2017/

Spencer, M. (2024, April 19). *Meta is Surprisingly Relevant in Generative AI*, AI Supremacy. https://aisupremacy.substack.com/p/meta-is-surprisingly-relevant-in

Srnicek, N.(2017a). *Platform capitalism*. Polity.

Srnicek, N. (2017b). The challenges of platform capitalism: Understanding the logic of a new business model. *Juncture*, 23(4), 254–257. https://doi.org/10.1111/newe.12023

Stamos, A. (2017, September 6). *An Update On Information Operations On Facebook*. Meta. https://about.fb.com/news/2017/09/information-operations-

update/

Stamos, A. (2017, October 3). *Promoting October Cyber Security Awareness Month*. Meta. https://about.fb.com/news/2017/10/promoting-october-cyber-security-awareness-month/

Staub, N., Haki, K., Aier, S., & Winter, R. (2021). *Taxonomy of Digital Platforms: A Business Model Perspective*. Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2021.744

Stockmann, D. (2023). Tech companies and the public interest: the role of the state in governing social media platforms. *Information, Communication & Society, 26*(1), 1–15. https://doi.org/10.1080/1369118X.2022.2032796

Stretch, C. (2013, August 27). *Global Government Requests Report*. Meta. https://about.fb.com/news/2013/08/global-government-requests-report/

Stretch, C. (2016, May 23). *Response to Chairman John Thune's letter on Trending Topics,* Meta. https://about.fb.com/news/2016/05/response-to-chairman-john-thunes-letter-on-trending-topics/

Stretch, C. (2017, September 21). *Facebook to Provide Congress With Ads Linked to Internet Research Agency*. Meta. https://about.fb.com/news/2017/09/providing-congress-with-ads-linked-to-internet-research-agency/

Suzor, N. P. (2018). Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms. Social Media + Society, 4(3) (July 1), 1–11

Suzor, N. P. (2019). *Lawless : the secret rules that govern our digital lives.* Cambridge University Press. https://doi.org/10.1017/9781108666428

Suzor, N. (2024, March 1). *Judging Meta with Dr Nick Suzor*, Per Capita. https://percapita.org.au/blog/our_podcasts/judging-meta-with-dr-nick-suzor/

Sztompka, P. (1999). *Trust : a sociological theory* (1st ed.). Cambridge University Press.

Taherdoost, H. (2023). Enhancing Social Media Platforms with Machine Learning Algorithms and Neural Networks. *Algorithms*, *16*(6), 271. https://doi.org/10.3390/a16060271

Taylor, L. (2021). Public actors without public values: Legitimacy, domination and the regulation of the technology sector. *Philosophy & Technology, 34*, 897–922. https://doi.org/10.1007/s13347-020-00441-4

Taylor S. (2024, March 27). Fears grow Meta will block news on Facebook and Instagram as Australian government faces pressure to act. *The Guardian.* https://amp-theguardian-com.cdn.ampproject.org/c/s/amp.theguardian.com/australia-news/2024/mar/28/meta-facebook-news-tab-instagram-fears-blocked-australian-government-regulation

Timberg, C., Shaban, H., & Dwoskin, E. (2017). Fiery exchanges on Capitol Hill as lawmakers scold Facebook, Google and Twitter. *Washingtonpost*.Com.

Tyler, T. R. (2006). *Why People Obey the Law*. Princeton University Press. https://doi.org/10.1515/9781400828609

The European Commission, *Tackling online disinformation*, https://digital-strategy.ec.europa.eu/en/policies/online-disinformation

Tobi, A. (2024). Towards an Epistemic Compass for Online Content Moderation. Philosophy & Technology, 37(3), 109. https://doi.org/10.1007/s13347-024-00791-3

Ulbricht, L., & Yeung, K. (2022). Algorithmic regulation: A maturing concept for investigating regulation of and through algorithms. *Regulation & Governance, 16*(1), 3–22. https://doi.org/10.1111/rego.12437

Vaidhyanathan, S. (2021). Antisocial media: how Facebook disconnects us and undermines democracy (Updated edition.). Oxford University Press.

Van Der Vlist, F. N., & Helmond, A. (2021). How partners mediate platform power: Mapping business and data partnerships in the social media ecosystem. *Big Data & Society*, *8*(1), 205395172110250. https://doi.org/10.1177/20539517211025061

Watts, J. (2006). Backlash as Google shores up great firewall of China: US search engine agrees to government restrictions; Firm admits inconsistency with its corporate ethics. *The Guardian (1959-2009),* 3.

Wang, Y., & Gray, J. E. (2022). China's evolving stance against tech monopolies: A moment of international alignment in an era of digital sovereignty. *Media International Australia, 185*(1), 79–92. https://doi.org/10.1177/1329878X221105124

Wardle, C. (2017, February 16). *Fake news. It's complicated*. First Draft.

https://firstdraftnews.org/articles/fake-news-complicated

Yeung, K. (2018). Algorithmic regulation: A critical interrogation: Algorithmic Regulation. *Regulation & Governance, 12*(4), 505–523. https://doi.org/10.1111/rego.12158

Zingales, L. (2017). Towards a Political Theory of the Firm. *Journal of Economic Perspectives*, *31*(3), 113–130. https://doi.org/10.1257/jep.31.3.113

Zuboff, S. (2019). *The age of surveillance capitalism: the fight for a human future at the new frontier of power (First edition.)*. PublicAffairs.

Zuckerberg, M. (2012, October 4). *One Billion People on Facebook.* Meta. https://about.fb.com/news/2012/10/one-billion-people-on-facebook/

Zuckerberg, M. (2018, November 15). *A Blueprint for Content Governance and Enforcement.* Meta. https://www.facebook.com/notes/751449002072082/?hc_location=ufi

Zuckerberg, M., & Senate Commerce, Science and Transportation Committee. (2018, April 11). Transcript of Mark Zuckerberg's Senate hearing. *Washington Post.* https://www.washingtonpost.com/news/the-switch/wp/2018/04/10/transcript-of-mark-zuckerbergs-senate-hearing/

Zuckerberg, M. (2019, March 30). Mark Zuckerberg: The Internet needs new rules. Let's start in these four areas. *Washington Post.* https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b752 5a8d5f_story.html

Zuckerberg, M. (2024, April 19). *Meta AI?*, Facebook. https://www.facebook.com/4/videos/377361005296904/

Zuckerman, E. (2021). *Mistrust: why losing faith in institutions provides the tools to transform them (First edition.)*. W. W. Norton & Company, Inc.

Appendix I: List of Participants

Interview number	Affiliation	Position	Date of interview
Participant 1	University of Technology Sydney	Professor	28 August 2024
Participant 2	Digital Rights Watch	Chair	3 December 2024
Participant 3	Monash University	Professor	14 October 2024
Participant 4	Queensland University of Technology	Associate Professor	21 October 2024
Participant 5	Western Sydney University	Professor	11 November 2024
Participant 6	Monash University	Associate Professor	22 October 2024
Participant 7	Queensland University of Technology	Professor; member of Meta's Oversight Board	13 November 2024

Appendix II: Interview Questions

1. I would like to begin by asking you to say something about yourself and your role in [add organization]? What initially interested you in digital platforms? Note: The organisation will be primarily of interest for industry and non-governmental organizations participants (the role someone has at a university is not so important).

2. Over the time that you have been working in a field related to digital platforms, what have been the major changes that you have identified?

3. From your perspective, what have been the positive and negative impacts of these changes?

4. As for negative impacts, what actions do the industry or other stakeholders need to take? And do you have some expectations for content moderation on digital platforms?

5. Do you see content moderation as an issue primarily for governments to resolve, or one for digital platforms themselves to address?

6. What do you think about the role of algorithms in moderating content? And how do you think artificial intelligence influences content moderation?

7. What do you understand by terms such as misinformation and “fake news”? Is there a role for governments in regulating content for misinformation, or does this present dangers to freedom of expression?

8. In terms of these negative effects, what are the difficulties of digital content moderation, and what kind of governance mode is needed to rebuild digital trust? (for industry participants only)

9. How do you define trust in the digital age from an economic and societal perspective?

10. What do you think will be the most important issues facing digital platforms in the next five years?

11. What achievements has the board made in rebuilding trust, particularly in the areas of accountability, transparency, and content moderation (for Meta's Oversight Board and Meta only)?

12. What's the particular case in your opinion? And what experiences have you had in making decisions about specific case studies (for Meta's Oversight Board only)?

Appendix III: Members of Meta's Oversight Board

Name	Country of Origin	Background
Afia Asantewaa Asare-Kyei	Ghana	Human Rights Advocate
Evelyn Aswad	United States	Professor and Chair, University of Oklahoma College of Law
Endy Bayuni	Indonesia	Senior Editor and Board Member, The Jakarta Post
Paolo Carozza	United States	Professor, University of Notre Dame
Katherine Chen	Taiwan	Professor, National Chengchi University
Nighat Dad	Pakistan	Executive Director, Digital Rights Foundation, and Lawyer
Tawakkol Karman	Yemen	Journalist, human rights activist and Nobel Peace Prize Laureate
Sudhir Krishnaswamy	India	Vice Chancellor and Professor of Law, National Law School of India University
Ronaldo Lemos	Brazil	Professor, Rio de Janeiro State University's Law School
Khaled Mansour	Egypt	Journalism and Communications Humanitarian Affairs Human Rights
Michael W. McConnell	United States	Professor and Director of the Constitutional Law Center, Stanford Law School
Suzanne Nossel	United States	Former Chief Executive Officer, PEN America
Julie Owono	Cameroon France	Executive Director, Internet Sans Frontières
Emi Palmor	Israel	Advocate and Lecturer, Interdisciplinary Center

		Herzliya
Alan Charles Rusbridger	United Kingdom	Editor, Prospect magazine
András Sajó	Hungary	University Professor, Central European University
John Samples	United States	Vice President, Cato Institute
Pamela San Martín	Mexico	Former Electoral Councilor at the National Electoral Institute (INE) in Mexico
Nicolas Suzor	Australia	Professor, School of Law at Queensland University of Technology
Helle Thorning-Schmidt	Denmark	Former Prime Minister, Denmark
Kenji Yoshino	United States	Chief Justice Earl Warren Professor of Constitutional Law and Faculty Director of the Meltzer Center for Diversity, Inclusion and Belonging