Governing Hate: Facebook and Hate Speech

By Paloma Viejo Otero. MPhil

Thesis submitted for the Degree of Doctor of Philosophy

Dublin City University

School of Communication

Supervised by Dr. Debbie Ging

School of Communications

Dublin City University

January 2022

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: (Candidate) ID No.: 14211934 Date: 12 January 2022

Acknowledgements

I wish to express my sincere gratitude and admiration to my supervisor Dr Eugenia Siapera from whom I have learnt how to express my ideas and formulate them with rigour. Eugenia is truly an inspiration whose compromise with justice and ideas is an example to follow. Without Eugenia's persistent help, the goal of this project would not have been realised. Dr Debbie Ging deserves particular mention too. For both her important work at the transfer stage and stepping in as supervisor for my last year of writing and guiding my thesis through the final stages. Thank you Dr Ging, I am immensely grateful.

I want to give special mention to Ronit Lentin who encouraged me to do this thesis, Andrew Finlay whose way of thinking I truly admire, and James O'Higgins Norman for offering me the possibility to work in academia and finish this thesis.

Thank you Gavan Titley (church) for your friendship, immense support, commitment against racism and advice. You have taught me how the academic world works, and it is thanks to you that I want to be part of it.

I could not have done this without my mother Margarita Otero Palacios. It is thanks to her that I understand how important and complex it is to fight against oppression – she laid the foundations that were fundamental to my inspiration to explore this topic. I own to her my work ethic and dedication in achieving my goals. Gracias Mama. Esta tesis es tuya. Y todo lo que venga después. Alongside my mother I want to mention my grandmother Covadonga Palacios Diaz, my grandfather David Otero Canto, my sister Olga Viejo Otero and my love Francois Kervarec - my emotional anchors, I dedicate this thesis to all of you.

I also must thank my friends Victor, Aisling, Caroline, Christine, Giordano, Niamh and Shannon. You are my Dublin family. Candela (my Cande), Marina, Alicia (the one and only), Silvia y Elena (the 1st of May), Tete y Mila (sois amor), Gela, Carlos, Sara y Nani, los Emilio's, los Natalio's, los Jorge's and all my friends in Leon and elsewhere who taught me everything I know about punk, rock and roll, and soul. The real seed of this dissertation.

Table of Contents

Contents

List of Figures

List of Tables

List of Appendices

Abstract

Governing Hate. Facebook and Hate Speech.

By Paloma Viejo Otero

This research investigates the relationship between Facebook and hate speech. In doing so, it deconstructs the governance structures of Facebook and analyses the principles and values that underpin Facebook's schemes of intervention. This thesis argues that, from the perspective of Facebook, hate speech is approached less in terms of its substance and more in terms of a practical problem that needs to be resolved in an operational manner. They therefore conceive of relevant policy in terms of its fit within Facebook's overall structure of governance, by which we mean the techniques or mechanisms used to internally order the various actors and actions. Theoretically, this thesis frames different approaches to hate speech regulation and adopts an understanding of governance as the means by which to regulate and order behaviours and actions, by using the work of Foucault and Miller and Rose to study governing systems.

The research question concerns the ways (including the mechanisms, instruments, features and action sequences and above all the discourses) by which Facebook orders and regulates the creation and circulation of content when it comes to hate speech. The empirical materials upon which the thesis relied in order to identify the parameters of the governmentality of digital hate include: Mark Zuckerberg's publications (May 2016 to November 2020), Facebook Principles and Values (2009 -2021), Facebook Community Standards (2016 2021), Content Standards (2018 2021), user's settings (2016-2021), and the Oversight Board (2019-2021). To supplement these materials, the thesis makes use of three in depth interviews with Facebook's Director of Public Policy, Campaigns and Programs (EMEA) Siobahn Cummiskey in 2016 and 2019 and with Facebook's Public Policy officer Aibhinn Kelleher in 2017. Finally, the thesis makes use of secondary data that includes internal Facebook training materials for content moderators leaked to Pro-Republica and The Guardian.

The key findings show that, while Facebook articulates its hate speech policies following a traditional liberal approach to hate speech, they operationationalise hateful content as a question of user safety. This results in an overall approach that is far removed from questions of social justice and emancipation. Instead, the focus is on procedural enforcement that produces more and more data. This emphasis on data, in turn, feeds into techno solutions relying on artificial intelligence and machine learning tools that are currently used moderately but which are planned as the preferred solution to what is constructed as a technical problem of content regulation.

Chapter 1. Introduction

1.1 Introduction

The purpose of my research is to understand how Facebook governs and regulates hate speech. Specifically, I aim to identify the series of principles and values that inform Facebook's hate speech policies and identify and analyse the techniques that Facebook employs to enforce these values on its platform. The overall purpose is to sketch Facebook's apparatus developed in order to govern hate speech and to determine whether Facebook contributes to fighting discrimination or whether it manages hateful content under a different rationale. If the latter, what is the rationale for this?

Specifically, this dissertation seeks to answer the following research questions.

How does Facebook regulate and govern hate speech? What are Facebook's principles and values around hate speech definition? What techniques has Facebook developed to enforce its policies? What values underpin Facebook's operational system?

The empirical research consists of a qualitative analysis of a diverse set of data. Specifically, these include: Facebook Principles (2009 version and 2019 version), Mark Zuckerberg's public posts (from 2016 to 2020), four interviews with Facebook Policy Actors (2015, 2016, 2017 and 2019), Facebook Terms of Service (2016 to 2020), Facebook Community Standards and Facebook Content Standards (2019 to 2020), and a range of non-textual settings, eg. visibility settings and newsfeed settings, reporting mechanisms that include buttons, drop-down menus or clickable links. Additionally, the specific analysis of Public Policy interviews, Mark Zuckerberg posts, Community Standards, About FB documents and literature lead by Roberts (2016, 2019) and Gillespie (2019) reveal the existence of Oversight Board, automatic detection and human reviewer teams, which I included in the list of techniques and, therefore, to the analysis of the series of Facebook techniques. In addition, I have used Facebook Transparency Reports such as About FB documents which inform on Facebook new initiatives, and Facebook investor, a Facebook-run site dedicated to informing on Facebook financial benefits.

These research questions were formulated on the basis of a genealogical analysis (now Chapters Two and Three), which showed the importance of looking at the values of equality and freedom and how they shape hate speech regulation. This analysis was threefold: firstly, I looked at the political actors who regulated hate speech; secondly, I looked at the historical context in which they occurred; and thirdly, I identified the values that different actors embed in hate speech regulation.

In drawing the genealogy of hate speech regulation, I identified three meaningful decades: 1940s, 1960s and 1980s. I also identified three important political actors associated with them: The Soviet Union, American Universities and the European Union. Next, I identified two fundamental principles that underpin any hate speech regulation, which are equality and freedom. In my analysis, I show that not all actors valued freedom and equality in the same manner. On the basis of this analysis, I deduced a total of four approaches to regulate hate speech, namely: Social Justice, Neutral Viewpoint, Freedom Absolutism and European Approach. All four differ in their understandings of equality and freedom. Following this analysis, I set out to understand to what extent Facebook can be thought of as the new hate speech governor, offering a new perspective on these four approaches – or, conversely, if Facebook reproduces one of these approaches. Why is it important to understand how Facebook governs hate speech? I outline my answer in the paragraphs below.

My understanding is that a well-lived life is one in which an individual's social value is realised in her capacity to contribute to society to the best of her ability (Kayes, 1998).

However, research points out that not all individuals enjoy the same opportunities to realise their potential. For instance, Roma and homeless people continue being the subject of violence and securitisation by local groups of vigilantes who organise online (Vasiuc, 2019; Viejo Otero and Siapera, 2015), migratory background continues to negatively affect the life trajectories of migrants' descendants (Gabrielli and Impicciatore, 2021), and individuals with diverse functionality continually encounter 'hard' and 'soft' barriers that ultimately determine long-term inequality gaps (Venturiello et al., 2020). In order to facilitate a good life, Kayes argues that it is society's duty to facilitate an individual's ability to contribute to society by guaranteeing or enabling access to social equality, freedom or opportunity (Adler and Seligman, 2016).

This dissertation focuses on equality and its relationship with hate speech. I understand equality as twofold: a) the right to not be discriminated against, and b) the creation of conditions that enable equal participation. Equality has an intrinsic relationship with hate speech regulation, since hate speech is a form of discrimination. Fundamentally, hate speech regulation is a legal resource, the purpose of which is to regulate discrimination and therefore achieve equality. However, discrimination is understood in two different ways: as discrimination against individuals, with no consideration of any historical or other factors, or discrimination against minorities or 'vulnerable groups' (European Commission, 2015), whose participation in society often encounters structural barriers. For hate speech to operate effectively as a legal resource, it is evident that it needs to take into account discrimination that targets minorities, and to facilitate the elimination of barriers that undermine the efforts of minorities to participate equally in society. The question that arises, therefore, is do social media fight discrimination in this sense?

I began this study in 2015, when hate speech activity was associated with the rise of the far right and with organised groups exploiting social media structures to spread hate. The focus, however, changed after the US elections in 2016, and in 2017 when the guidelines and training material that Facebook gave to its human moderators were leaked to Pro-Republica and The Guardian. These events shifted the general attention from the activities of groups to the activities of social media platforms. In 2017, I began to look at social media as an actor with a focus on how social media regulates hate speech. Among all platforms, Facebook was particularly at the centre of attention.

Indeed, it matters how Facebook governs hate speech. Its enormous concentration of power, its number of users, the success of its business model, and the ease with which it transmits and, to some extent, imposes its policies make Facebook an important subject of research.

Facebook is a private enterprise with the capacity to self-regulate matters that were previously in the public domain. The concern around their status as private companies and capacity to self-define and self-regulate hate speech is shared by many authors in the field (Hestres, 2013; De Nardis & Hackl, 2015; Gillespie, 2018; Milosevic, 2018) who look at how the majority of social media platforms are protected by 'safe harbour' provisions, such as those in the US 230 Decency Act. These provisions establish that social media platforms are not legally liable for illegal content that they host, because they are not responsible for creating these contents. As such, any actions that they take to control or limit hate speech and other kinds of problematic content are voluntary.

Studies that look at Facebook have mostly focused on its moderation activity and specifically on the role of human reviewers (Gillespie, 2015, 2018; Roberts, 2016). To the best of my knowledge, no study has looked at how Facebook governs hate speech, what principles and values drive the governance of hateful content, and how it operationalises hate speech. This lack means that we don't yet know if Facebook contributes to the fight against discrimination of minorities and, in this manner, contributing to equality.

In sum, all the above justifies the importance of researching Facebook, based on the understanding that Facebook is not as a neutral technological device or platform, but an active player, whose influence on how hate speech is conceptualised is comparable, for example the influence of the Soviet bloc in the 1940s, as Chapters Two and Three will illustrate.

1.2 Location of my research within literature.

As Chapter Two, Three and Four will outline in depth , this dissertation is influenced by three bodies of work: 1) hate speech literature from a legal perspective (Morsink, 1999; Matsuda, 1993; Walker, 1991; Altman, 1993; White, 1996; Shiell, 2009; Mchangama, 2011; Berg, 2012; Banks, 2011; Heinz, 2016); 2) studies on governance (Foucault, 1978, 2007; Miller and Rose, 2008), and 3) studies on platform governance (Gillespie, 2010, 2015, 2018, 2020; van Dijck, 2013, 2015; Terranova, 2015, 2017; Roberts, 2016, 2019; Owen, 2018; Gorwa, 2019; Suzor, 2018; Klonick, 2018). While hate speech literature is concerned with equality and regulation of discrimination, studies on governance are responsible for understanding the means by which social media governs, and how its mechanisms of governance affect the social fabric. Specifically, I have drawn from the legal literature the two core concepts or principles that underlie the creation of hate speech regulation, which are equality and freedom. Those two principles lie at the centre of all discussions on hate speech regulation, and applying them to the study of social media platforms is a novel contribution. From studies on governance, I have gathered the importance of studying the formation and functioning of several techniques and the principles that underpin those techniques (Miller and Rose, 2008). Finally, from platform governance literature, I have acquired detailed knowledge about the main technological and regulatory techniques involved in governing content, and the function of those techniques around the platform. Platform governance is a new approach that I understand as an emerging field.

Platform governance is influenced by the philosophy of technology studies (e.g. Feenberg, 2010) and by studies on governance (Foucault, 1978, 2007). The field draws from early studies on technology that look at platforms from an architectural viewpoint. My position in the field aligns with the work developed by Gorwa (2019), who understands platform governance in its broadest dimension, as involving the various 'layers of governance relationships structuring interactions between key parties' (Gorwa: 2019 pag 2) and as focusing 'on the "external" processes that influence platform governance' (Siapera and Viejo Otero, 2021). My main influences from the platform governance field are: Jose van Djick (2013), whose systematic study of platforms' terms of service illuminated part of the empirical research, and Tarleton Gillespie (2010; 2015; 2018), who paid attention to the actual means by which social media moderate content. Similarly, this research owes part of its reasoning to De Nardis & Hackl (2015), specifically their critique on how platforms privately regulate, and Terranova (2015; 2017), whose work examined social media platform studies from a Foucauldian perspective, providing an overarching sociological approach to how social media governs and operates. In addition, I draw inspiration from Fuchs (2015), whose account of social media points out that, despite the empowering rhetoric used by platforms, they have actually commodified users for their own interests.

Early research on platform governance has focused on systematically and critically analysing all the techniques and elements that include the governance of content, namely moderation. It seemed evident to me that there was a need to understand how all those tools are designed, what the overarching purpose of these techniques is, and what are the actual governing tendencies behind digital platforms, not at the technical level but in relation to the intention and direction that the platform aims to take as Miller and Rose (2008) suggest. By intention and direction, I mean a theoretical approach that seeks to identify the set of ideas, beliefs and values that platforms produce or reproduce and which can be 'read' through analysing their policy documents, design architectures, norms and practices. The analysis of how social media companies intend to direct and influence what happens on their platforms are understood as principles in this dissertation. Studying and understanding these principles and the values is necessary in order to provide depth and nuance to descriptive analyses of technological means to govern hate speech. Looking at the principles underlying the responses and techniques will offer an explanation and, in this manner, enable the integration of all research fragments and their relation to each other.

This dissertation is making use of a specific terminology that, for the most part, emerged from and is grounded in the above-mentioned strands of literature. The main terms used are outlined below.

1.3 Key terminology

Hate Speech: The term 'hate speech' is used interchangeably to refer to "regulation of hate speech" and "the act of speaking with hate". The difference is that hate speech is a term that comes from the legal and cultural sphere and encompasses all attempts to regulate hate speech within society (Shiell, 2009), whereas "the act of speaking with hate" is, as its name indicates, the action of reproducing or creating expressions that seek to subordinate and discriminate (Matsuda, 1993). On the basis of this differentiation, this dissertation will focus solely on the regulation of hate speech, which this dissertation understands as a regulatory tool whose purpose is to address discrimination and is a measure to achieve equality (Viejo Otero, 2020). Among those discussing hate speech, there is a given argument, often repeated in seminars and professional environments, that, uncritically used, functions as unquestionable truth: 'there is no universally accepted definition of what constitutes hate speech'. This argument is founded on two reasons. First, there is no common agreement among legal theorists whether hate should be regulated and, if so, which categories of people should be protected. Second, there is no common universal legislation on hate speech that equally applies to all societies and, instead, different countries have implemented hate speech laws, with different limitations in expression and different protected categories. In addition, this thesis argues for a third reason as to why there is no universal definition of hate speech, which is that there are ideological differences between actors responsible for regulating hatred and discrimination, and, more precisely, differences concerning how these political actors conceptualise inequality and seek to achieve equality.

Finally, because the focus here is on hate speech as a regulatory tool, it is important to note that this thesis is not examining the different forms of hateful content that is uploaded daily to the platform. Rather, this thesis focuses on unveiling how social media platforms in general, and Facebook in particular, conceptualise equality and freedom and apply it to the governance of hateful content. To this aim, this dissertation has adopted a research perspective that embraces a Foucauldian perspective on governance (Foucault, 1979; Miller and Rose, 2008). As such, I seek to examine Facebook's governance by looking at its principles and values around equality and freedom and how these are transposed into governance processes and enforcement policies.

Ideology: Platforms both perpetuate and re-create principles and beliefs. Whereas some principles are adopted from previous models, i.e. the notion of freedom from liberal ideology, others are re-created, i.e. equality understood as equality of Facebook profiles/accounts. However, neither freedom of equality are ideas created by Facebook. Rather, the platform adopted them, reproduced them and re-created them (see Sessen, 2002).1 In considering this, I specifically refer to ideology as those ideas that Facebook re-creates or perpetuates and that function as principles that the platform uses to guide its policies.

According to Marx, ideology refers to the series of ideas that serve the social function of masking and legitimating the dominance of the ruling class. Gramsci argued that ideologies

[1] Saskia Sessen (2002) argues that the tendency is to interpret technological as purely technological and sociological as purely sociological. However, and as the authors points out, understanding the place of these new technologies from a sociological perspective requires avoiding a purely technological interpretation and recognising the influence of the different previous social orders that have led to technological beings. Its result may condusively constitute new social dynamics, but they can also be derived or reproduce older conditions (2002)

are distorted philosophical theories that become a doctrine. Monasta stated that idologies are a set of moral principles for orienting practical actions and human behaviour (Monasta, 1993). Marxists' 'super structural' interpretation of ideology emerged in the 1970s and sought to integrate the notion of ideology into the material reality of the state holding that ideology is enforced and enabled via an apparatus (see Althusser, 1971). In considering this, for the structuralisms, ideology could no longer be regarded as a matter of spiritual ideas or false ideas, but as a matter of assemblages, apparatuses, and complex relations with their own conditions. As such, the task of governance studies was to understand the formation and functioning of those apparatuses (Miller and Rose, 2008). Foucault excelled in this task by tracing with precision the small and disperse events that brought a concept into existence. Following this line of thought, the present study understands ideology in terms of what it does, and how it exerts power through generating and operationalising certain ideas and concepts. Facebook's ideology, therefore, consists of the ideas and concepts that Facebook mobilises and relies upon when developing its hate speech policies.

Governing techniques: This refers to the series of mechanisms and strategies by which users' content is rendered controllable in Facebook's platform. Governing techniques are diverse and range from the infinitesimal, such as a Mark Zuckerberg post or a setting, to more common aspects, such as community guidelines or terms of service.

Dispositive: The present research uses the term dispositive to describe the ensemble of techniques that govern the directing, controlling and disposing of something in a particular manner. These techniques steer towards a particular direction, precluding or obstructing others. For instance, in Chapter Six the term 'dispositives of ideology' is used to examine the set of techniques that, assembled together, legitimate the superiority of Facebook leadership in all matters that relate to the values, beliefs and general direction of the platform. Chapter Seven focuses on the dispositive of discipline which is used to determine the set of techniques that, together, ensure that all values that emanate from Facebook leadership are translated into how the platform operates. Finally, the dispositive of security is used in Chapters Seven and Eight to determine the set of techniques that together control how hateful content circulates.

Dispositive is a term that Foucault first used during the courses at the College de France, later published in the collection 'Security, Territory and Population' (1977 and 1978):

> What I am trying to pick out with this term [the dispositive; le dispositif] is... a thoroughly heterogeneous ensemble, consisting of discourses, institutions, architectural planning, regulatory decisions, laws, administrative measures, scientific statements, philosophical, moral and philanthropic proportions – in short, the said as much as the unsaid. Such are the elements of the dispositive. The dispositive itself is the network that can be established between these elements. Foucault (2007: 11)

This definition is, however, not satisfactory for many Foucauldian scholars who consider the term vague, dispersed and easily interchangeable with the term apparatus (Dreyfus and Rabinow 1982; Miller and Rose, 2008). In light of this controversy, Raffnsøe et al. (2014) point to the use of the term by the German school of critical discourse theory and analysis (see Woodak and Meyer, 2009), which adopts the term dispositive from a 'research perspective' and refer to it as 'the ensemble of techniques with includes discourses, practices, institutions, objects, and subjects' (Raffnsøe et al., 2014: 8) and how they interrelate (Jager and Maier, 2008). As such, this dissertation uses the word dispositive to refer to the ensemble of techniques that steer towards a particular direction, or are in charge of directing, controlling and disposing of something in a particular manner.

Apparatus: The Critical Discourse Analysis school acknowledges that 'there is possibility for several dispositives to work together' (Raffnsøe et al., 2014: 21). Similarly, Foucault's lectures in 1977 referred to history, which shows the different dispositives that evolve and interact as 'complex edifices' (Foucault on Raffnsøe et al., 2014:3) or 'systems of correlation' (Foucault, 2007: 8). As such, and in light of all the above, apparatus is for this dissertation the sum of dispositives that work together.

1.4 Structure of this thesis

Chapter Two gather the background research conducted for this thesis and begins by exploring the history of hate speech and identifying key moments where regulating the act of speaking hatefully was on the agenda. Based on literature on the field, this research identifies four key moments:

1. 1946 to 1948 Drafting of the Universal Declaration of Human Rights,
2. 1964-1969 Convention on the Elimination of All Forms of Racial Discrimination and establishment of the High Committee on National Minorities,
3. 1980s Campus Hate Speech, and
4. 1989 to 2010s and the rise and establishment of the European Union.

Chapter Two highlights three arguments. First, that the origins of hate speech regulation are to be found in the Soviet Union and all different regulations around hate speech are the result of a compromise between the Soviet Union and the Western Bloc in 1948. Second, that the regulation of discrimination occurred in a period framed by the Cold War, and all discussion around hate speech heavily influenced by the ideological tensions of the time. And third, that hate speech regulation has a life history, meaning that it goes beyond the simple zero-sum that represents the debate between regulations of hate speech versus freedom of expression.

Chapter Three Complements Chapter Two, in that it aims to identify the logics, imaginaries, and key values around equality and freedom that underpin different regulatory approaches to hate speech. The analysis, organization, and systematization of the positions held by actors involved in hate speech regulation reveal that there are four different approaches to hate speech regulation: the Social Justice approach, Neutral Approach, Freedom of expression absolutist approach, and European approach. The differences between the approaches lie in how each political actor understands freedom, how each actors understands equality and, specifically, the path to achieve equality. The following briefly explains the main characteristics of these approaches.

- The Social Justice Approach considers that oppressing minorities is an obstacle to achieving equality; since these minorities are historically specific, the approach aims to name oppressed groups and minorities and actively protect them from discrimination. For this approach the right to freedom of expression does not justify to perpetuate inequality.
- The Neutral viewpoint approach considers that all individuals are born equal and they are equal before the law regardless the history of the groups. As such, it seeks to protect all individuals against discrimination by legislating around any ‘wrongdoing’ (Altman 1993), on the grounds of generic categories such as gender or race. Freedom of expression is protected and the approach aims to balance the right to freedom of expression with the right to be protected against acts that discriminate.
- The Freedom of Expression Absolutist Approach considers that all individuals are born equal and they all are equal before the law. Those who champion this approach consider individuals to be free and capable of commanding their own destiny. As such, instead of regulating against hate speech, the Freedom of Expression Absolutist approach argues that it is only by using the right to freedom of expression that minorities can defend themselves against oppression and discrimination.
- Finally, the European Approach holds that all European citizens are equal and no one should be discriminated against. This approach seeks the equality of all Europeans, limits freedom of expression, protects domestic minorities, and seeks the integration

of all European citizens around a narrowly conceived European history that recognises the history of fascism, but which denies the history of colonialism.

With this chapter, this dissertation finishes the background analysis. This analysis serves to frame the research questions to explore Facebook. However, there is another kind of analysis that needs to be undertaken before focusing on the empirical part. This analysis, which is presented in Chapter Four, critically explores what current literature has to say on the relationship between hate speech and social media.

Chapter Four is concerned with social media platforms and reviews existing literature that explores how platforms govern hate speech. Social media have added to the conversation a technological dimension that has added a layer of complexity to hate speech. Due to this technological dimension hate speech in social media is no longer only a philosophical discussion, but a conversation that involves understanding how technology serves ideology. When compared with previous actors, Social media emerged decades away from the tension between the two blocks of the Cold War period; however, the heritage of the tension and discourses that emerged during this period casts a long shadow. Social media is the fruit of the Californian ideology (Barbrook & Cameron, 1996) which, according to the authors, emerged as a movement to overcome the perceived rigidity that the Cold War had imposed on the political and social world by creating an alternative space online where individuals could overcome traditional social boundaries. In considering this influence, it is therefore important to explore how social media platforms conceive hate speech regulation and the techniques it has developed to govern hateful content in the digital space. Chapter Four, therefore, looks at the problem of hateful content on social media platforms, their responses to hateful content, and looks in particular at two research areas: socio-technical dynamics research and platform governance research. Socio-technical dynamics refers to the study of user's behaviour on the platform, taking into account both social dimensions and technological features, whereas platform governance looks at the 'political effects of digital platforms, as well as the complex challenges that the governance of platform company's presents' (Gorwa, 2019, p. 855). To embrace platform governance involves delving into what constitutes governance systems. In this regard, studies of governance are, according to Miller and Rose (2008), the study of all techniques that intervene in the life of individuals and the knowledge and ideology that underpins this scheme of intervention. Platform governance studies have thoroughly examined different techniques, namely: Terms of Service (Van Dijck, 2013; Suzor, 2018) human reviewers, flagging technique or artificial intelligence (Gillespie, 2018, 2020; Roberts, 2016, 2019). However, the field lacks a systematic analysis that explains social media ideological settings and the values, principles, and knowledge that underpin the techniques that social media employ to govern hate speech. To cover this gap in literature, I have chosen Facebook as a field of study due to the number of users that Facebook enjoys and the possible influence over politics its users have. I argue that we can identify from what perspective or approach Facebook regulates hate speech and the particularities of Facebook's contribution to hate speech regulation by looking at Facebook Principles and Values around equality and freedom and how they affect their governance techniques.

Chapter Five is concerned with the Research Design and Methodology of the present dissertation. Overall, the chapter explains how I identified extracted and analyse data to explore Facebook. The data I have analysed is Facebook Principles (2009 version and 2019 version), Mark Zuckerberg public posts (from 2016 to 2020), four interviews with Facebook Policy Actors (2015, 2016, 2017 and 2019), Facebook Terms of User (2016 to 2020), Facebook Community Standards and Facebook Content Standards (2019 to 2020) a range of non-textual settings: i.e. visibility settings or Newsfeed settings. Additionally, the specific analysis of Public Policy interviews, Mark Zuckerberg posts, Community Standards, About FB documents and literature lead by Roberts (2016, 2019) and Gillespie (2019) reveal the existence of Oversight Board, Automatic Detention and Human reviewer's teams, which I included in the list of techniques and therefore to the analysis of the series of Facebook techniques.

The analysis conducted has been threefold. First, the set of data is analysed independently, understanding the set of data as documents that contain information. Specifically, the analysis seeks to identify Facebook set of Principles, understand how Facebook interpret each principle and how these flows and permeates across the different set of data or techniques. The second analysis consisted in the identification of techniques that Facebook employees to govern hate speech and identifying hierarchy among techniques. The third analysis amiss to reorganise data into techniques and into dispositives to draw Facebook Apparatus of Governance.

The different set of data was analysed in three different manners . Analysis 1, *Analysing the data as carriers of information* looks after the analysis of documents as standalone forms of information. This this analysis I used Sociology of Knowledge Approach to Discourse or SKAD (Keller, 2011; Keller Hornidgerand Schunemann, 2018) for textual analyses and dispositive analysis (Jagger and Maier, 2009) for non-textual analysis. In addition, I complemented SKDA analysis with word count coding (Dickingson and Pool, 200) and Dispositive analysis with walkthrough method (Burgess and Duguay). Analysis 2, *Analysing the data as techniques and in terms of their location* looks at the form of data as techniques of governance. For this I used Document location analysis (Prior 2018). Finally, the chapter explains Analysis 3, *Analysing the relationship between techniques*, delves on how I rearranged all the techniques, grouping them into categories which I refer throughout this dissertation as dispositive. According to Foucault, dispositive represent a network of power relations (2007). In considering this, I grouped all techniques according to their roles to understand how they relate to each other in forming a more comprehensive apparatus. The three categories I used were: ideology (Miller and Rose, 2008), discipline (Foucault, 2007), and enforcement (Milosevic, 2018).

Chapter Six is the first of the empirical analysis chapters and it seeks to exemplify the principle and values that underpin Facebook hate speech governing apparatus. With this aim the chapter analyses Facebook Principles, Mark Zuckerberg posts, and their influence on Facebook's definition of hate speech. This chapter argues that those two techniques drive Facebook ideologically, influencing its governance of hate speech. The chapter argues that the most relevant principle that influences Facebook's definition of Hate speech is Fundamental Equality or Article 4 of Facebook Principles. The chapter points out how Facebook's hate speech policies are not an end in themselves but the consequences of Facebook's mission of Connectivity and Community. Whereas Connectivity and Community imply that all users are on the same platform, Article 4 of Facebook's principles on "Fundamental Equality" refers to the idea that on Facebook, all user profiles are treated equally, whether they are a page of a public institution or a personal profile with a handful of friends. The fundamental finding is that Facebook makes distinctions between users, nor takes into account their cultural capital or background. Therefore, taking into account that Facebook considers all its users arithmetically equal and that it makes an effort to limit hateful content, this thesis argues that Facebook adopts and adapts the Neutral Viewpoint to hate speech as seen in Chapter Three.

To expand into this finding, this chapter next analyses how Facebook deals with the notion of freedom and what role it plays on the platform. In particular, the chapter points out that Facebook's explanation of freedom has changed in the last four years: whereas freedom was originally understood as freedom of expression and freedom of information, successive changes on Facebook Principle documents reveal that Facebook revisited its notion of freedom and decided to favour freedom of expression at the expense of freedom of information. It is important to point out the differences between the two types of freedom. Whereas freedom of expression refers to the act of creating thoughts, ideas, works, and opinions; freedom of information specifically refers to the act of receiving and issuing information through media (United Nations 59.I, 1949). In considering this, I argue that Facebook does not limit freedom of expression but specifically freedom of information. That is, Facebook does not control what users say, but the information that circulates across the platform. In considering this finding, the question remaining is how Facebook controls the flow of hate speech, what values justify the control of the flow of content and the techniques that Facebook uses to enforce this control. These issues are covered in Chapter Seven.

Chapter Seven is the second and last chapter dedicated to empirical analysis, and it is concerned with how Facebook operationalises and enforces its policies-and values- on hateful content. This chapter argues that Facebook enforces its hate speech regulation not by limiting freedom of expression but freedom of information. That is, Facebook does not limit what users say, but how content circulates. In Facebook, operationalising content refers to all the activities and strategies that the company has put in place to detect potential hateful content, and how it analyses it and categorises it as acceptable or unacceptable. Facebook operational activity is handled by policy officers and their task is to determine the type of content that goes against Facebook policies. To know what constitutes acceptable or unacceptable content, policy officers follow a diverse range of activities such as engaging with NGOs, academics, and different experts in the field to define the content standards and to work to enforce them. A policy team translates this work into product solutions or tools, guidelines for human reviewers and also facilitates the work of the Oversight Board. Policy officers work across the board according to the three values of ‘voices, equity and safety’ (Cummiskey, 2019). The principle of Voice is Facebook equivalent to freedom of expression and it is the most protected principle, to such extent that it is possible to argue that Facebook operates to protect the Principle of voice while maximising the possibility of safety for the users, and it does so by controlling the volume and characteristics of content that flows across the platform. To explore all the above, the chapter first examines the disciplinarian role of the policy team and looks at the process by which Facebook decides Content standards. Next, the chapter analyses the role of the principle of 'Voice' and the meaning of Facebook's mission of Safety, concluding that, in order to provide safety Facebook has created a disciplinary and a security dispositive whose aim is to protect the principle of Voice, to creating a safe environment for users to upload content, and to control what kind of content circulates across the platform, which substantially differs from protecting its users or the individual behind each profile or from fighting discrimination or oppression.

Chapter Eight is the conclusion of this thesis and it is concerned with the Facebook apparatus to govern hate speech and the dynamics between the governing techniques. In the previous chapter I treat the different sets of data that Facebook uses, understanding them as techniques of governance. I treated them as disjointed pieces of a puzzle, analysed them in isolation to understand their role and design. In each chapter, I point out how each technique, according to its role, falls into categories of ideology, discipline, or security, conforming to what I have referred to as dispositives. In considering this, this chapter aims to explain how each dispositive relation relates to the other and how, together, they conform to Facebook's apparatus of governance for hate speech. To be specific, I refer to the following dispositives:

- Dispositive of ideology - composed by Facebook principles and Mark Zuckerberg.
- Disciplinarian dispositive - includes policy officers, external stakeholders, and trust flaggers and
- Dispositive of security - commonly known as moderation (Gillespie, 2018), and which includes user settings, artificial systems, human moderators, and the oversight board.

In analysing the relation among dispositives, the chapter illustrates that Facebook is ideologically driven.

Facebook Principles and values permeates all Facebook governing techniques, determining how Facebook approaches hate speech regulation and content enforcement. The principles that influence how Facebook governs hate speech are Voice, or freedom of expression, and Fundamental Equality, by which all users are equally treated under the same set of regulations. In considering this, the chapter concludes that Facebook approaches hate speech governance under the Neutral viewpoint approach, as laid out in Chapter Three. Whereas platform governance critiques the lack of coherence between rules and enforcement (Gillespie, 2018), this thesis argues that Facebook acts coherently. By focusing on the notion of safety, Facebook implements a coherent set of rules and enforcement, but, in the absence of courts, judges, or superior bodies, it has, as a result, displaced the intention held by previous actors. If for previous actors Hate Speech regulation was intended to be a resource to achieve equality, Facebook operationalised hateful content as a matter of security. Hence, this is how Facebook contributes and transforms the notion of hate speech, by no longer associating hate speech with equality but, instead, to a safe environment that little or nothing relates with the fight against inequality that previous actors aimed to achieve.

As a final note, the final chapter outlines how this dissertation contributes to existing literature, revitalising Mornsik (1999) arguments around the role of the Soviet Bloc in the fight against discrimination, which planted the seed for hate speech regulation. Additionally, this dissertation contributed to the literature on hate speech by identifying and organising four different approaches to hate speech regulation, countering the domination of American perspectives with two distinct approaches: The Social Justice Approach and the European approach. Finally, the empirical chapters contribute to platform governance by systematically analysing principles and values behind Facebook’s operational techniques and mechanisms, using Miller and Rose’s (2008) suggestions for the study of governance. The chapter then outlines the limitations of the research and implications for future research.

Chapter 2. History of Hate Speech Regulation. From 1946 to the 2010s

2.1 Introduction.

The purpose of this chapter is to identify and explore the history of hate speech regulation. This chapter functions as the thesis background, and nurtures the grounds for researching Facebook. Indeed, when Facebook got involved in the governance of hate speech, the topic acquired popularity. However, hate speech and hate speech regulation pre-exist the creation of social media platforms. The exercise, therefore, has consisted in identifying the actors, and the historical context in which hate speech regulation was on the agenda. The efforts have provided both perspective and knowledge in relation to hate speech and has contributed to demystified the notion of hate speech and Facebook praxis on hateful content.

Specifically, this chapter identifies four historical periods: (i) 1946 to 1948 Drafting of the Universal Declaration of Human Rights; (ii) 1964-1969 Convention on the Elimination of All Forms of Racial Discrimination and establishment of the High Committee on National Minorities (Convention on minorities); (iii) the 1980s US Campus Hate Speech which refers to the series of debates that emerged around the implementation of Hate Speech Codes at American University campuses; and (iv) 1989 to 2010s, which focuses on the creation of the European Union and how they regulate hate speech. These four moments were identified based on previous literature that considers them to be pivotal in the evolution of hate speech regulation. Morsink (1999) Mchangama (2012) and Berg (2012)[2] point out that the seeds of hate speech regulation were sown during the Drafting of the Universal Declaration of Human Rights, while Matsuda (1993), identifies the 1964-1969 Convention on the Elimination of

[2] The author of this chapter acknowledges the existence of other relevant documents ie:The International Covenant on Civil and Political Rights (ICCPR) of 1966, the Convention on the Rights of Persons with Disabilities (CRPD) of 2006, or 2018 UN Strategy and Plan of Action on Hate Speech however they have been discarded since they do not advance or alter any previous understanding of hate speech.

All Forms of Racial Discrimination and establishment of the High Committee on National Minorities. 1980's Campus hate speech debate is analysed by several authors Delgado (1993), Walker (1991), Altman (1993), Matsuda (1993), White (1996), Shiell (2009), Banks (2011), and Bleich (2014). Finally, Heinz's (2010, 2016) work on Freedom of Expression and Europe led this dissertation to take into account how Europe conceptualised and regulates hate speech.

One of the main points this chapter seeks to support is that, previous to Facebook, the history of hate speech regulation evolved to the beat of the Cold War with the ideological tensions of the time. Indeed, one of the big actors of this periods was the Soviet bloc, whose contribution to the development of regulation against discrimination was crucial, and which seems to have been forgotten in recent history. By bringing back to the table the input offered by the Soviet bloc, this chapter offers a perspective and understanding of hate speech regulation that has not yet been explored in depth and which highlights the need to fight discriminatory speech and the regulation of privilege and class instead of limiting freedom of expression. In considering this, the chapter contends that the regulation of hate speech has a life, history, and meaning that go beyond the simple zero-sum that represents the debate between regulations of hate speech versus freedom of expression.

The following chapter is organised on chronological order and both highlights the peridos on which hate speech was on the agenda, the main actors involved and the series of tensions and contradictions they encountered.

The two main arguments that this chapter raises are that 1) hate speech regulation had a previous life and further meaning than limitation to freedom of expression; and 2) that the evolution of hate speech regulation has been shaped by the ideological tensions that characterized the Cold War period.

2.2 Hate speech in the 1940s. Declaration of Human Rights

The notion of 'hate speech' is closely linked to the fight against discrimination and the Declaration of Human Rights. After the crimes committed by fascism, the task of avoiding a repetition of such racial horrors was entrusted to the United Nations, who prepared the Declaration of Human Rights as a means of protecting human dignity and preventing discrimination. The Declaration was proclaimed by the United Nations General Assembly in Paris on 10 December 1948 and was born as a 'common ideal for all peoples and nations'. However, far from harmonious, its drafting was marked by ongoing ideological tensions and divisions that ultimately shaped the final edition of the Charter.

Thanks to the thorough task of United Nations' administrative staff, all discussions, agreements, proposals, and corrections to proposals that took place between 1946 and 1948 were transcribed. Some of those notes, even when scattered (See list in Appendix I), have been reviewed for the present research, with the help of the work carried on by Johannes Morsink (1991). These documents indeed reflect the rising tensions at the beginning of the Cold War period, revealing the ideological tendencies and agendas of the members of the Commission with great clarity. Among them, Morsink (1999) highlights the contributions made by the Soviet bloc, which constantly referred in their speeches to the treatment of African-Americans and Native Americans in the United States, apartheid in South Africa, and the treatment of peoples of the British colonies of Gold Coast, Nigeria, and Rhodesia. Vladimir Koretsky was the main commissioner on behalf of the Union of Soviet Socialist Republics and advocated that 'one of the first principles to be adopted in the formulation of an International Bill of Rights must be the destruction of discrimination and inequality' (Koretsky, E/CN.4/AC.1/SR.5 See Appendix II; see also Morsink, 1999, p. 93). As such, soviet commissioners like Vladimir Koretsky, on behalf of the communist bloc, made a tremendous push for non-discrimination on the basis of race and sex (Morsink, 1999).

For the Soviet bloc, the Declaration of Human Rights was, not only a legal, but a moral document, one that should recognize the damage of colonialism and repair the impact of European fascism, guaranteeing the representation of the female gender in public spheres. Alongside Koretsky, Borisov, Bogomolov, and Pavlov were the names of the commissioners representing the Soviet Union (Morsink, 1991) and their view was that historical and systematic discrimination were important obstacles to overcome in order to acquire universal equality. In order to overcome such obstacles, the Soviet bloc presented the first draft, of what Morsink argues was the seed of hate speech regulation (1991). The proposal was submitted to the Subcommittee on Prevention against Discrimination and The Protection of Minorities, with the intention of this formula becoming Article 7 of the Declaration. For the Soviets, the intention of article 7 was to protect individuals against the privilege that leads to discrimination and to have equal rights before the law. It reads as follows:

> any defence of national, racial and religious hostility or national exclusivity or hatred and contempt, as well as any action establishing **privilege** or discrimination based on distinctions of race, nationality or religion, would constitute a crime and should be punishable under the law of the state' (Morsink, 1999, p.71) (emphasis added)

This proposal was not accepted by the Western Bloc. The Soviet idea, to its full extent, required the removal of the privileges of some social groups and also sought to empower the state and governments to define which groups could be considered minorities. Some Western representatives interpreted the communist bloc's proposal as a way to identify and suppress their opponents. However, we must also take into account that the problem of removing privileges, as suggested within the Soviet proposal, was a problem for most Western countries who engaged in the above privileges.

The communist bloc understood privilege to be the benefits that are systematically granted to members of certain social groups or persons recognized or perceived as members of those groups, such as the historical privilege of the white man. Needless to say, this was a problem for the Western world of the 1940s, where there was a de facto racial segregation in the United States, colonies were ruled by the metropolis, apartheid reigned in South Africa, fascism in Europe had not been eliminated, and where, in most member countries of the United Nations, women could not vote. As a result, for the soviet commissioners, the concern underlying the Western Bloc lied in conveying an argument against discrimination and equality when most western societies had accepted and standardized discrimination within their social systems. It is, therefore, not surprising that the Western Bloc based their argument against the soviet proposal on the notion of Freedom of Expression, which was heavily criticized by the soviet bloc, as the following quote illustrates:

> It cannot be said that prohibiting the promotion of ethnic, national, or religious hatred constitutes a violation of press freedom or freedom of expression. Between Hitler's racial propaganda and any other propaganda designed to provoke hatred and racial, national, or religious incitement to war, there was only one small step. Freedom of the press and freedom of expression cannot serve as a pretext for spreading opinions that poison public opinion. Propaganda in favour of racial or national exclusivity or superiority served as an ideological mask for imperialist aggression. This is how the German imperialists tried to justify their plan of destruction and plunder in Europe and Asia (AC.1/SR.7/po.9) (Bogomolov, cited in Morsink, 1999, p. 70)

Mitigating the discussion, representatives of the Western Bloc proposed an amended version of the Soviet proposal. The Western proposal attempt to redraft the Soviet proposal was based on the Rights of Man (sic). As such, instead of protecting specific groups such as women or the population of the colonies, the Western Bloc proposed to protect each and every single human being individually from discrimination. By protecting all individuals, the West avoided the problem of regulating privileges and the problem of empowering the state over the definition of minorities. The amendment to the Soviet bloc proposal was accepted and voted on by a majority and it is known as Article 7 of Human Rights: 'Everyone has the right to equal protection against all discrimination' (UN, 1998). In this definition, the use of the word 'everyone' implies that all humans have the right to be protected by the same hate speech laws, regardless of the historical process or the groups they belong to. In addition, when Article 7 was drafted, Morsink notes that the General Commission did not examine what it had written in Articles 18: Freedom of thought, conscience and religion, Article 19: Freedom of expression and opinion, and Article 20: Freedom of assembly and association (Morsink, 1999). As a result of this, the word "everyone" included the Nazis and fascists, who held the rights to freedom of thought, expression, and assembly. Finally, Article 7, does not name specific groups for their protection against discrimination, as was the desire of the Soviet bloc, instead states that all individuals are given equal protection against hatred, regardless of whether the individual belongs to a historically oppressed group or oppressor group. Remarkably, this crucial change ended with the inability to define what constitutes fascism, as article 7 offers the same protection against hatred for both fascists and non-fascists. Indeed, the Soviets point out on different occasions that what brought together different nations to write the Declaration of Human Rights was the fight against fascism. The soviet commissioners noticed that words like ‘fascist’ or ‘fascism’ were less referred to, as the discussions progressed. The notions of fascism and its implications were fading away from the agenda, even when ironically, it was the fight against fascism and the will to fascism never happen again that brought all together. Finally, the Declaration of Human Rights was written without naming fascism or offering a definition of special protection to oppressed groups, which led the Soviet bloc to abstain in the final rounds of votes (Morsink, 1999).

Despite the above, the Soviet Bloc has not been historically presented in an attractive light when it comes to its role on hate speech regulation. I particularly refer to two articles: Th*e Sordid Origins of Hate Speech by* Mchangama (2011) and *The Soviet Origins of Hate Speech* by Berg (2012). These articles define the regulation of hate speech as the evilest of all possible limitations of freedom because they emanated from the heart of a communist dictatorship. However, strangely, both articles take Morsink as a source (1999), who rather than presenting the Soviets as the ultimate evil argues that the anti-discriminatory tone of the human rights chart has a Soviet signature (Morsink, 1991).

In sum, the present section has outlined the origins of hate speech regulation which are pinpointed to between the end of the Second World War and the beginning of the Cold World. The analysis of the Soviet Bloc proposal has offered a new angle to understand hate speech, which consisted of giving protection to historically oppressed groups and ensure fascism was not repeated.

According to Morsink (1999), a remarkable weakness of the Declaration of Human Rights was the unwillingness of commissioners to define historically oppressed groups, a question that was reopened in 1964-1969 and which constitutes the second period of analysis of the present chapter.

2.3 1964-1969. Convention on the Elimination of All Forms of Racial Discrimination and the Establishment of the High Committee on National Minorities

Just a few years later, after the final vote of the Declaration of Human Rights, in the 1950s, attacks on synagogues across the United States showed that, generally speaking, some groups are more targeted than others. In the period between 1959 and 1960, there was a resurgence of anti-Semitic incidents that author Howard J. Ehrlich (1962) referred to as the '*Swastika epidemic of 1959-60'*. According to Howard, on Christmas morning 1959, a synagogue in Cologne, Germany, was 'desecrated' (1962). This event created an unprecedented wave of anti-Semitic incidents that took place in a total of thirty-four countries. Taking advantage of this moment, and perhaps encouraged by civil rights movements in the United States, the United Nations reopened the issue of minorities. The question on this occasion was whether minorities could be defined and protected and, if they were, who would have the power to define them as minorities.

Similar to the Drafting of the Universal Declaration of Human Rights, the USSR with Poland was confronted to the Western Bloc (Matsuda, 1993). During this convention, there were two aspects that are relevant for the present case. The definition of minorities and the question of superiority. As such, considering that the focus is how to approach the definition of historically oppressed groups, to explore what constitutes a minority and what is the significance of the question of superiority emerge as crucial. These are discussed below.

2.3.1 Minorities

One of the main points of the Soviet Union, during the Drafting of the Universal Declaration of Human Rights, concerned the power of the state to define and accommodate minorities into the national regulation. Lenin supported national self-determination since 1917, to recognise and accommodate ethnic minorities and national groups was part of a larger strategy to sustain the Soviet Union unity (Kreindler, 1977). The Soviet Union created institutions that allowed them to access the modern state and, by 1923, a policy on self - determined minorities was established, known as Lenin's policy of nationalities, which guaranteed cultural and linguistic rights. This was of particular benefit for ethnic minorities who faced discrimination (Martin, 2001). This policy made the Soviet Union the first major state power in the world that promoted the national consciousness of ethnic minorities, understanding ethnic minorities as both native and non-native groups (Kotlarchuk and Sundström, 2017). However, the Soviet Union changed radically in the 1930s under the Stalin era. The Bolsheviks argued that linguistic isolation, and particularly religious beliefs, were preventing ethnic groups from contributing to the progress of industrial cities. In 1937, Stalin ordered the arrest of Germans working in electric, military, and chemical factories. It followed the arrest of Polish, Korean, Chinese, Afghan, Iranian, Greek, Bulgarian-

Macedonian, Finish, Estonian, Latvian, Greek, Japanese, Lithuanians, Norwegians, Swedes and also Sami, Shamans, and other minoritarian ethnic groups. According to official data, a total of 335,513 people were arrested in the course of the national operations, and 247,157 of those arrested were shot (Werth, 2003, p. 232), making up '34 per cent of all the murdered victims of the Great Terror' (Kotljarchuk and Sundström, 2017, p.16).

Scholars of the Great Terror focus on the aspect of security in the border and argue that 'Western' minorities were killed, not because of their ethnicity, but because of their possible connections to countries hostile to the USSR (Kotljarchuk and Sundström, 2017). As such, and based on this particular historical precedent, the Western Bloc considered that the Soviet proposal, of providing the state the power to name, protect, and accommodate minorities, was an excuse to eliminate those who would present opposition to the communist party.

Based on this precedent from the Drafting of Human Rights Chart, in 1964, the Western Bloc shifted towards the argument of self-determination. According to Thornberry (1989), the power of self-identification, or self-determination, as a minority must reside with the affected groups and should not be determined or accommodated unilaterally by the state. Minority group members can seek recognition from the state based on claims of a shared history, culture, language, religion, and so on. Often, this takes years of struggle. As an example, we could put forward the case of Travellers in Ireland, who have fought for their definition as an ethnic minority and who was recognised in 2017 after years of struggle. In some cases, recognition and self-determination is fought for, as in the case of Travellers in Ireland, but in other cases it is deeply political. For example, France adopts a republican position that accepts only the citizen /non-citizen position. In some cases, as we saw in the Balkans in the 1990s, claims of ethnic minority status were linked to state formation and the redrawing of borders and lead to genocide. Yet, in other cases, minority status is linked to certain benefits, as, for example, with native Americans in the US, creating material incentives for people to self-identify.

To make matters even more complex, questions of existing groups of designated minorities also arise. Can individuals choose to leave a group designated as a minority? Can individuals leave a group and the bureaucracy that fixes the group? The paradox is that, while it is claimed that a minority cannot be designated by the state, the state does designate culture and bureaucratic boundaries, via birth certificate, national identification cards, and passports. Therefore, the issue of who determines what constitutes a minority cannot unilaterally be taken, either by the state or by the groups themselves, and must allow individuals to pursue the ability to speak within the group and the ability to leave the group (Hirschmann, 1978). However, the state has the power to define the majority and minority groups and the rights and responsibilities they enjoy.

2.3.2 Superiority

Returning to the Convention and, given the difficulties of defining a minority, a concept that gained momentum was that of superiority and, in particular, the relationship between superiority and propaganda, as reflected in Article 4 of the Convention, which puts in black and white the idea of 'theories of superiority':

> States Parties condemn all propaganda and all organizations which are based on ideas or **theories of superiority** of one race or group of persons of one colour or ethnic origin, or which attempt to justify or promote racial hatred and discrimination in any form, and undertake to adopt immediate and positive measures designed to eradicate all incitement to, or acts of, such discrimination and, to this end, with due regard to the principles embodied in the Universal Declaration of Human Rights and the rights expressly set forth in article 5 of this Convention […](Article 4 UN General Assembly 21 December 1965) (emphasis added)

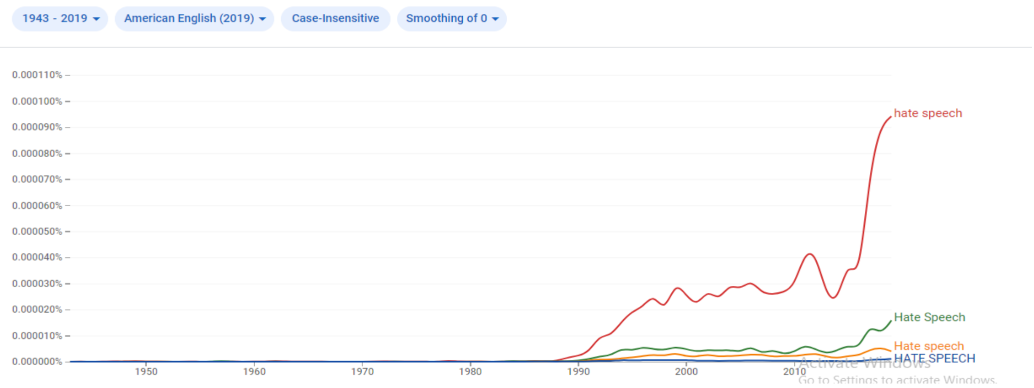
The idea of 'superiority' also enjoys a place on the preamble to the convention, where it explicitly states that 'any doctrine of superiority based on racial differentiation is scientifically false, morally condemnable, socially in just and dangerous, and that there is no justification for racial discrimination' (Racial Discrimination Convention, 214). By including the words 'theories of superiority', the commissioners referred to the six million Jews murdered during Nazism because of 'theories of superiority'. Indeed, this article paved the way for regulation against holocaust denial and dealt with the problems of the attacks over the synagogues that inspired the convention. However, it can also be observed that the idea of 'superiority' presented in this Convention displaced the idea of 'privilege' that the Soviet bloc proposed during the Drafting of the Universal Declaration of Human Rights and does not fully cover the scope intended by the Soviet bloc. That is, the notion of privilege proposed by the Soviets directly alluded to the material and structural effects of oppression, whereas 'theories of superiority' explicitly referred to Fascist ideology, the narratives and symbols constructed around the movement that lead to the Holocaust. As such, and by removing the notion of privilege, the focus remained on ideologies and not so much on the material benefits these ideologies confer on some people (i.e. the privileges). In defining racism and anti-Semitism as theories of superiority, the battleground was moved from struggles over equal access to resources to matters of ideas, discourses and narratives, paving the way to what is known as 'battleground of ideas' (Heinz, 2016).

In sum, this section and its two sub-sections have illustrated a second period of the evolution of hate speech regulation. The section has explored The 1964-69 Convention on the Elimination of All Forms of Racial Discrimination and the Establishment of the High Committee on National Minorities and has raised the question of minorities and the power of the state in naming minorities, as framed by the tensions between the Western Bloc and the soviet bloc. The section ultimately argues that the states have the power to name minorities. The section has also explored the question of superiority, as included in the convention, where there is explicit regulation against 'theories of superiority'. In the section, I argue that 'theories of superiority' benefited the fight against anti-Semitic narratives and ideological constructions of superiority. However, by focusing on superiority and not privilege, the emphasis was moved from notions of material oppression and confined racism to the terrain of ideas, narratives and opinions. In the end, the 1964-69 Convention of Minorities did not resolve the matter of what groups are to be protected in hate speech regulation.

2.4 1980s US Campus Hate Speech

With the gradual incorporation of African Americans into American campus classrooms, aggressive racist actions challenged the assumption that educated people were immune from prejudice and discrimination (Walker, 1991). Among many others, there was a particular incident that occurred at Arizona State University that attracted' national attention. Individuals hung a racist poster with the title "Job Request" on the door of one of the students, listing several 'ostensibly employment-related issues that foster hostile and degrading racial stereotypes for students African Americans and Mexican-Americans' (Calleros, 1995, p.1259). The reaction to this type of racist attack by many of the major American colleges and universities was to adopt student codes of conduct, i.e., to regulate and punish any form of discrimination based on race, gender, religion, marital status, physical, and capacity preferences. Regulation of hate speech began in 1987 and was led by Stanford University and the University of Michigan, and by 1992, most of the U.S. campuses had adopted hate speech codes and included them in their internal regulations. In the end, more than three hundred universities had implemented systems, with variations, in order to coexist with different state standards; as such, certain universities adopted hate codes, other hostile environment codes, or determinations of areas such as areas of free speech or restricted areas. These codes were primarily aimed at students and occasionally applied to regulate faculty and staff as well (Shiell, 2009). To the best of my knowledge, there are no records of measures taken; however, the assumption is that breaking the codes would incur some form of punishment, from a warning to expulsion (Ehrlich and Scimecca, 1991, Gibbs 1992).

The proliferation of policies attracted concern from the American Civil Liberties Union (ACLU), which viewed these codes as a significant threat to freedom of expression and academic freedom. These events sparked a large number of academic publications that imported the legal lexicon of hate speech into academia (Walker, 1991; Shiell, 2009) and framed a debate that, mostly, presented hate speech as a limitation to freedom of expression. Such debates came to prominence in the public eye from 1989 to mid-late 1990s, with hundreds of articles appearing in national newspapers and magazines, as well as in academic journals, publications by professional organizations, revisions of laws, and alumni newsletters (Shiell, 2009). As such, and what is particularly important in this period, is that the majority of hate speech literature emerged here, bound to a particular American context, bound to freedom of expression, and bound to American legislation, notoriously influencing all posterior hate speech studies worldwide.



1943 - 2019
American English (2019)
Case-Insensitive
Smoothing of 0

Figure 1 Extracted from Google Books Ngram Viewer. Whereas literature emerged in the 1940s, the figure, with no smoothing curve, illustrates the 1980s and 1990s as the moment where a major number of references to hate speech appear. Retrieved April 2021.

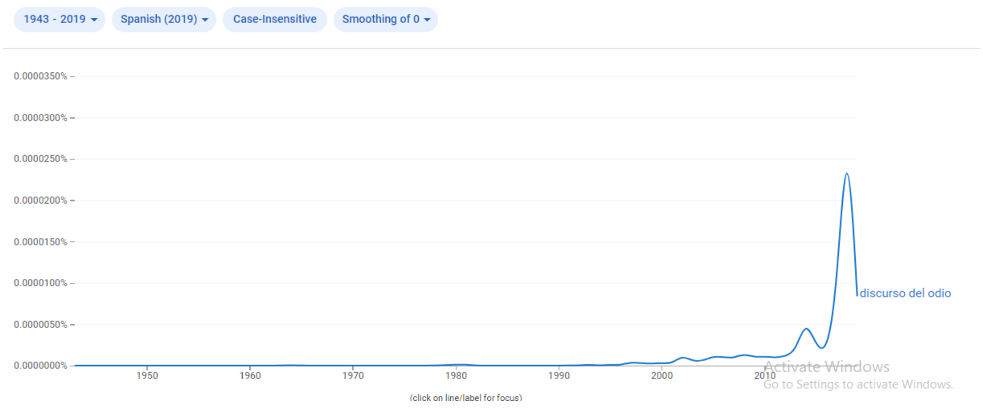


Figure 2 Extracted from Google Books Ngram Viewer. The figure, with no smoothing curve, illustrates a major number of references to hate speech appearing in Spanish. Spanish literature has been heavily influenced by American Literature on Hate Speech from the 1980s and 1990s. Retrieved April 2021.

Indeed, the vast majority of publications that emerge in this period had a tendency to present hate speech as a limitation to freedom of expression. Even academic circles, who did not intend to frame hate speech codes as a limitation of freedom of expression, were dragged into this debate. This was the case of the Critical Race theory group, whose views hold historical material perspectives, but who develop their arguments in a rather hostile territory.

2.4.1 Critical Race Theory Group

The most prominent critics of the liberal perspective were Matsuda, Delgado, Lawrence, and Crenshaw (1993), who were members of what is known as the Critical Race Theory group. The Critical Race Theory Group emerged in the late 1970s in legal circles and was established in the early 1980s. The origin of this group is narrated by Kimberlé Crenshaw, who became concerned that, in the mid-1980s, the subject of racism and American law was about to disappear from Harvard law school curricula. The author and creator of this subject, Derrick Bell, was about to leave for a different university and students were looking for an adequate replacement (Matsuda, 1993). The lack of an adequate response by the university management triggered a series of meetings and associations, where Harvard law students, not only discussed the need to continue with this subject, but the need to persuade the administration to bring in more teachers from communities of colour. While the university administration did not agree with the demands of the protesters, the group of protesters stepped aside and, independently, commenced to organize seminars, workshops, and conferences at Harvard and various other universities that helped to forge, from the margins, a well-known body of work which aims to alter the power relations around race, and which is collectively known as Critical Race Theory. Indeed, the main feature of the Critical Race Theory (CRT), in comparison with the Race Critical Group (RCT), is that it has its origins in the legal field. It is no wonder, then, that, when they developed the body of work that relates to hate speech codes, they did so from a regulatory perspective that primarily criticized the First Amendment of the American Constitution.

The Critical Race Theory group developed a critique of the First Amendment from a legal and a historical perspective. According to this group, in the right to freedom of expression, resides the right for oppressive groups to perpetuate situations of discrimination. For Critical Race theorists, the first amendment is a racist and liberal weapon alike, which grants, based on the so called First Amendment Neutrality, the institutional right to be racist. That is, content neutrality assumed by the First Amendment disguises the harm caused to historically oppressed groups who are disproportionately targeted by hate speech. By not condemning hate speech against these groups in explicit and unequivocal terms, and by not condemning the privilege that certain groups held and perpetuate when uttering hate speech slurs, the First Amendment ends up, almost by default, enabling racist hate speech to circulate unchecked. It, therefore, actively discriminates against racialised groups. Cases like the Burning Crosses in 1990 exemplify the point that CRT wants to advance. This refers to an incident where a Black family, the Jones, moved to a white neighbourhood in St. Paul, Minnesota, and were targeted with different racist attacks. The last of the attacks was a cross burned in their backyard. The case was first identified and prosecuted under the local hate crime ordinance, where the perpetrator claimed that his actions were protected under the

First Amendment. However, the supreme court of Minnesota rejected the perpetrator's arguments, as it considered that the burning cross is an undeniable symbol of violence and racial hatred, with explicit references to the recent history of the United States. The case garnered the attention of the right-wing in the United States who claimed that a 'politically correct movement' was running to the detriment of a white man, whereupon it was brought to the Supreme Court of the United States in June 1991. It was not a surprise that the associate justice of the Supreme Court, Antonin Scalia, opined 'the anti-cross burning ordinance was unconstitutional, and while littering people's property is prohibited, the directive should not single out racially motivated acts' (Matsuda and Lawrence, 1993, p.135). In sum, this case illustrates the point that the Critical Race Theory Group wants to advance: that the First Amendment of American constitution historically shields the privileges of white men and permits them to justify any action under the notion of freedom of expression.

The Critical Race Theory group also explored the effects that speech has on oppressed groups. Indeed, emphasis on the notion of 'speech' is an important contribution in this period. Though this does not mean that it is a novelty or a new contribution, since, for instance, the term 'Hate Speech and Racist Hate speech can be found in volumes dating to the early 1960s of American Jurisprudence (Lawyers Cooperative et al., 1962, p.205). The objective of the Critical Race Theory group was to define words and terms so bound to racial oppression that they should be seen as constituting criminal offenses. While racial slurs can, and have been, reclaimed by historically oppressed groups, this, according to the Critical Race Theory group, does not diminish their ability to contribute to oppression. In these terms, it is necessary to enforce controls over their use in public discourse, not in order to police freedom of expression, but in order to counter racial discrimination. This notion was, in turn, criticised by Butler (1997), who agrees with the fact that words harm, however the author argues that 'to account for such speech acts, one must understand language not as a static and closed system whose utterance are functionally secured in advance by the 'social positions' (Butler 1997, p145) as Matsuda argues. Instead, for Butler words are not determined by prior context or 'positions' but might gain its force 'precisely by virtue of the break with context that it performs'(Idem) and are therefore not an easy subject of regulation as the Critical Race Theory Group proposes.

The Critical Race Theory Group, and universities which adopted the codes of regulation, had countless detractors who specifically constructed their arguments around freedom of expression. According to Andrew Altman (1993), those opposed to Campus hate speech codes divided the liberal between conservatives and the liberal left. Conservatives and Neoliberal authors such as Le Marche (1996) or Sunstein (2000) opposed the implementation of Hate speech codes for considering freedom of expression an unqualified absolutism right that cannot be balanced or compromised against any other right. Also, authors like Calleros saw in the right to freedom of expression a fundamental tool for those targeted to fight oppression and racism (1995). On the other side, liberal left groups understood freedom of expression as a qualified right that requires limitations; although those limitations do not mean that words should be limited. Instead, the limitation should rely on and be determined by the intention of the perpetrator or 'wrongdoing', a position defended by authors such as Altman (1993).

The academic and non-academic sphere echoed the arguments from both liberal and absolutists that opposed hate speech codes. The majority of the articles published adopted the free speech argument, giving it greater visibility and significance than the arguments forged by Matsuda, Delgado, Lawrence, and Crenshaw, who developed their arguments in hostile (or liberal) territory. This is also pointed out by Walker (1991), an expert on freedom of expression, who argued that the liberals won the debate because they constitute the majority, rather than because of the weight of their reasons.

In sum, there are three basic ideas that we can learn from 1980s Campus Hate Speech debates. Firstly, that literature that emerged during this time influenced the field and subsequent research. Secondly, this period emphasises the role of speech and language, seeking to understand the specific words that should be seen as constituting criminal offenses. Thirdly, this period emphasizes the idea that hate speech regulation constitutes a limitation of freedom of expression.

Often, literature in Hate Speech drafts the differences between American and European perspectives (Boyle, 2001; Heinze, 2016). The following section explores the role and position of the European Union in the definition of hate speech.

2.5 1990-2010. The European Union

The solidification of the European Union took place at a time of political upheaval in Europe, which included the reunification of Germany and the dissolution of the Soviet Union in 1991. By 1986, the Single European Act was signed. The act stipulated a six year programme whose goal was to set European standards across member states and to set up the basis of a single market. Two pivotal moments propelled the consolidation of the European Union: the fall of the Berlin Wall in Germany and the creation of the syndicate Solidarność in Poland, which together paved the end of communism in Europe. The 1990s witnessed the signature of Maastricht (1993) and Treaty of Amsterdam (1999), by which European Union member states enjoyed freedom of movement of goods, people, service, and money. The guarantee of the four freedoms was already in operation throughout the 1990s and it was followed in the 2000s by the establishment of the monetary union and the enlargement and consolidation of European institutions, including the Commission, the Council, and the Parliament .

The Council of Europe’s roots can be found in 1974. The Maastricht Treaty was particularly relevant, as it gave to the Council of Europe legal powers to influence member states. Ever since then, the main role of the Council of Europe has been to protect human rights and promote European unity by 'fostering cooperation on legal, cultural and social issues' (Council of Europe, 2021). Under the light of unity, and considering the European history of xenophobia, anti-Semitism and other forms of racism, The Council of Europe does limit freedom of expression and regulates hate and discrimination. Evidence of this can be seen from the signing of Recommendation No (97) 20 on Hate speech, in 1997, which condemn 'all forms of expression which incite racial hatred, xenophobia, anti-Semitism and all forms of intolerance, since they undermine democratic security, cultural cohesion and pluralism' (Council of Europe, 1997). Ten years later, the CoE published Recommendation 1805 (2007) on blasphemy, religious insults and hate speech against persons on grounds of their religion. This was followed by the 2008 Framework decision on combating certain forms of racism and xenophobia by means of criminal law. Again, it does not define specific minorities. At the time of writing, the most updated and detailed recommendation in terms of hate speech is the 2015 ECRI General Policy Recommendation N15 on Combating Hate Speech (September 2015). In its preamble, this recommendation refers to the limits of freedom of expression: 'freedom of expression and opinion is not an unqualified right and that it must not be exercised in a manner inconsistent with the rights of others' (15th Recommendation September, 2015). 'Unqualified right', in turn, refers to 'rights which cannot be balanced against the needs of other individuals or against any general public interest. They may be subject to specific exceptions, e.g. the right not to be deprived of liberty, Article 5; or to none at all, when they are called absolute rights, e.g. freedom from torture, Article 3' (Council of Europe, 2021). Of particular interest is that the 15th Recommendation protects specific groups and minorities such as Gitanos, Roma, Muslims, and Jews and it does so with clear allusions to Europe's historical past.

The European position on hate speech is criticized by authors such as Heinze (2016) who considers that hate speech only harms if it leads to phisical violence. This, of course, does not take into account the harm that it causes as speech alone. In addition, Heinz argues that hate speech prohibitions in Europe have not stopped racial attacks. Instead, according to the author, since racially motivated attacks are less probable in Europe than in the United States, Europe should promote freedom of expression and counter narratives. However, the rise of nationalism and the far right, Brexit, widespread xenophobic and anti-migrant attacks negates the argument about the probability of racial attacks in Europe. Instead, the aforementioned developments provide evidence that Europe needs to do more to counter discrimination.

In sum, this section has outlined the end of the Soviet Union and the rise of freedom of expression in the United States as a defining 'Western' cultural value. Overall, Europeans' position is the recognition that freedom of expression is not an absolute right and, secondly must be conceded to historically oppressed groups particularly those who were targeted by fascist. However, it does not go as far as The Soviet Union would go because it does not take into account histories of oppression beyond the Holocaust and possibly the Roma people. It therefore constitutes a compromise between liberalism and the Soviet position, but which skews toward the liberal view in its limitations.

2.6 Conclusion

The overall purpose of this chapter has been to demystified the notion of hate speech while providing perspective in relation to how Facebook governs hate speech. In particular, this chapter identify the history of regulatory approaches to hate speech and has focused on four periods that the literature refers to as pivotal moments. The chapter has shown that the concept of hate speech emerged after World War II and was shaped by the tensions that defined the Cold War. The chapter first reviewed the role of the Soviet bloc in defining what now constitutes Article 7 against discrimination of the Declaration of Human Rights Chart. Morsink (1999) indicates that the Soviet bloc planted the seed of hate speech regulation by advocating, at all times, for an anti-discriminatory agenda and aimed to set regulation that would alter the conditions that perpetuate inequality and fuel discrimination. At the core of the Soviet proposal is the notion of privilege, which the Soviet bloc saw as an obstacle to acquiring equality. For its part, the Western Bloc diluted the notion of privilege and advocated for an individualistic approach to Article 7, to protect all individuals equally from discrimination. Indeed, while all the commissioners accepted and sought equality, the Declaration does not accept the notion of equality as a cultural and historical right. The same is true of the forthcoming Convention on Minorities, which takes a clear position against propaganda, and replaces privilege with superiority, but which does not define what constitutes a minority or, more specifically, does not empower any public body to determine what a minority is. Consequently, no national or international hate speech regulation could use the Convention on Minorities to introduce specific protected categories such as, for instance, 'people of African descent' , 'people in systemic poverty' or 'women'.

In the 1980s, in the American liberal context, the Critical Race Theory paradigm raised concerns about racial discrimination on college campuses. The group supported and advocated for the stipulation, implementation, and application of hate speech codes at all American universities. Authors such as Matsuda (1993) argued that hate speech in all forms of expression perpetuates social subordination. Matsuda, Delgado, Lawrenece and Crenshaw (1993) rooted the problem of verbal racial discrimination in the First Amendment and its cultural project, which intrinsically protects white privilege and freedom of expression or, in their terms, protects the right to be racist. The arguments of Critical Race theorists were met with concern by civil liberties groups and with scepticism by Conservatives and Liberal left authors whose criticisms centred around defending freedom of expression as an absolute right. It thus highlights the difficulty of determining which specific words wound and argues that the problem lies in the intention to harm and not in the words themselves. Campus Hate Speech debates emerged as an important and divisive issue in American universities, creating an avalanche of publications in which historical materialist arguments on the role and legacies of historical oppression were vocally disputed by liberal positions focusing on the potential losses involved in limits to freedom of expression. The most abundant and rich literature on hate speech emerged during this period, so much so that, from a researcher point of view, it is difficult to find literature that explores hate speech without defining it as a limitation of freedom of expression. As such, and paraphrasing Walker (1991), the debate on hate speech was won by liberal groups whose presence and volume of publications did not leave much room for the development of different understandings of hate speech.

The late 1980s-early 1990s marked the end of the Cold War and the dismantling of the Soviet Union. With the end of the communist bloc, hate speech regulation lost any support for the idea of tackling privilege and fighting oppression in order to acquire equality. In its place, the debate was skewed towards liberal positions and overlooked or disregarded arguments on the legacy of historical oppression and the specificity of the experiences of certain historically oppressed groups.

In sum, this chapter set out to show the history and genealogy of the notion of hate speech and has identified four moments in recent history that have shaped the debate. It has also identified a set of actors (i.e. political, academic) whose ideological influences have led to different interpretations of what constitutes speaking hatefully and of hate speech regulation. In doing so, the chapter provides context and generates perspective to undertake the study of Facebook. Specifically, it indicates that when empirically researching Facebook, it is important to understand Facebook as an actor with agency, historically situated in a context that influences how to understand and govern hateful content.

The next chapter seeks to identify the logics, imaginaries, and key values that underpin regulatory approaches in the present.

Chapter 3. Approaches to the Regulation of Hate Speech

3.1 Introduction

> The notion of rights, which was launched into the world in 1789, has proved unable, because of their intrinsic ineptitude to fulfil the role assigned to it.
>
> (Weil, 2005, p.71)

The present chapter builds upon the discussion of Chapter Two and aims to identify the logics, imaginaries, and key values that underpin regulatory approaches to hate speech. The analysis, organization, and systematization of the existing literature has lead me to inductively identify and construct four different approaches to hate speech regulation: Social Justice Approach, Neutral Viewpoint , Freedom of Expression Absolutism approach, and European approach. The first approach, which is labelled here as Social Justice approach , refers to the position that understands discrimination as structural, and aims to tackle the historical, political, cultural, and economic conditions that sustains discrimination. In contrast, the freedom of expression absolutism approach emerged as a term in the US during the 1980s and 1990s to differentiate those who oppose any hate speech regulation from those who struggle to reconcile freedom of expression with regulating against racism and misogyny. The latter are referred to by Altman (1993) as the liberal left which the author identifies as people concerned with discrimination but who defends freedom of expression[3]. and that for this dissertation, we will refer to this approach as 'neutral viewpoint approach . Finally, the European approach refers to the rationale used in Europe to regulate hate speech, which limits freedom of expression and favours the integration of European citizens, and to

[3] For Altman the main distinction lies in that liberal left aims to regulate speech that discrimination, whereas conservatives opposed to any limitation of speech. I use Altman and rename the sides as 'neutral viewpoint approach and 'Freedom of expression absolutism'

avoid the repetition of historical horrors occurred within the borders of Europe i.e. Nazism and the Holocaust.

In this chapter, I argue that the difference between each approach lies in how each socio-political vision conceptualizes the term equality and, in particular, how such equality is achieved. To elaborate on this analysis, this chapter is based on the main discussion held by authors and the idea that hate speech is a 'compositional concept' (Brown, 2017). The notion of a compositional concept indicates a formulation where ‘the meaning of the whole concept is understood to be a function of the meaning of the parts that make up the whole’ (Brown, 2017, p. 433). To say that ‘hate speech’ is a compositional concept is to say that it is informed by other concepts such as discrimination , inequality and freedom of expression. In considering this, we then have to look at the components which make this up and the related notions they conjure up. For example, social hate conjures up inferiority, superiority, or Hate speech as limitations of freedom of expression. Examples can be found in authors like Shiell (2009) who discuss freedom and equality when analysing hate speech, as well as Walker (1991) whose book about hate speech explores freedom of expression. As such, and for the present case, I have based the analysis on how political actors aim to achieve equality through hate speech, and also what version of equality they aim to achieve and how freedom is conceptualised in relation to hate speech regulation. Balibar and Ingram (2014) argue that within modern political imaginaries equality implies and carries within the notions of liberty and freedom. For the author, Equality and Freedom, or ‘egaliberte’ (2014) are locked in a dialectic where the definition of one defines the other. As such, the ways in which they are conceived and the various inflections or flavours given to one or the other lead to different outcomes when it comes to positions on hate speech regulation given as results four different approaches to hate speech regulation.

This chapter therefore presents the four approaches to hate speech regulation in the following order: Social Justice Approach. Neutral viewpoint approach Freedom of expression absolutism approach and European approach. The discussion explores the popularity of Neutral viewpoint approach and critically engaged with such popularity, arguing that it does not contributes to fight against discrimination. The chapter concludes with a reflection on how social media platforms relate to these four approaches.

3.2 Social Justice Approach

The main concern of the Social Justice approach is to deal with structural hate in order to achieve equality. Structural discrimination differs from interpersonal or institutional discrimination. According to Pincus (1994), individual discrimination refers to individual behaviours whereby individuals of one group intend to have a differential and harmful effect on members of another group. Institutional discrimination alludes to the policies of institutions and those involved in the elaboration of polices who intend to have harmful effect on members of minorities, for example, schools who do not accept lower income students or Muslim girls with veils. On the other hand, structural discrimination refers to forms of social organization that do not target minorities per se, but where their policies negatively affect them by neglecting their needs or by systematically keeping minority groups in subordinate positions (Pincus, 1994, p.122).

In these terms, the Social Justice approach is based on the idea that subordination and oppression of groups is part of the broader social economy, where values and symbols circulate and become attached to certain groups and individuals, operating at a psychological and material level (Ahmed, 2004). As a result, liberal approaches to hate speech contribute to structural discrimination which, both perpetuates and creates hatred towards groups historically marginalised from dominant narratives, namely woman, black people, and people from the colonies (Essed, 1999).

As noted in the previous chapter, the two main actors who have argued against structural forms of discrimination were the Soviet Bloc, during the drafting process of the Declaration of Human Rights and during the Convention of Minorities, as well as the Critical Race Theory group during the Campus Hate Speech debate in the 1980s. The following subsection focuses on these two actors, exploring their rationale and their socio-political imaginary, putting special emphasis on how this approach understands Equality and how it is accessed. According to Morsink, during the 1947 General Assembly of Human Rights, Soviet delegates declared in front of outraged Western delegates that fascism was 'the bloody dictatorship of the most reactionary section of capitalism' (1991, p.68). This political link between fascism and capitalism is, I argue, rooted in the Marxist notion of primitive accumulation.

By primitive accumulation, Marx refers to series of violent dispossessions i.e. enclosing public spaces, such as forests, depriving people from a source of food, appropriating the labour of women, slaves and peasants; before they were needed in the factories (Federici, 2004). Authors such as Adam Smith would argue that capital emerges as a result of deregulated trade and market supply and demand, whereas for Marx (1909), the discovery of gold and silver in America, the removal, slavery and burial in the mines of the aboriginal population [of America], the beginning of the conquest and looting of the East Indies, the change of Africa in a reserve for commercial hunting of black skins, are, together with the confinement of women to domestic roles (Mies, 1983; Federici, 2004), the central moment of primitive accumulation and, consequently, the origins of capitalism (Marx, 1909). The origins of capitalism are, therefore, created through the violent subjugation of certain groups of people, which then had to be justified on the basis that these people deserved this treatment because they were ‘inferior’ or biologically destined for specific roles (e.g. women) (Maies, 1983). Exploitation was often facilitated by overt violence, expropriation and oppression, understanding oppression as ‘a state of asymmetrical relations of power characterized by domination, subordination and resistance, in which the controller person or group exercises its power through processes of political exclusion and violence and through psychological dynamics of depreciation' (Prilleltensky, 2003, p.195). Hatred is then for those who oppressed understood, not as a sentiment, but as a set of ideas which serve to construct and maintain the treatment of whole categories of humans as inferior. Following this line of thought, I argue that the Soviets followed Marxist views in its actions of defending women and people from the colonies. Therefore, the notion of primitive accumulation, as well as the reliance of capitalism on the continued subjugation and exploitation of whole categories of people, provides an insight of what led the Soviet arguments.

> capitalism entails the exploitation of certain groups of people, generally the proletariat, but also women and colonised people such as 'black people in the United States, of Indians in South Africa, and of the peoples in the British colonies of the Gold Coast, Nigeria and Rhodesia (Pavlov Alexander, in Morsink, 1991, p.131).

The fundamental arguments of the Soviet bloc centred on the representation and protection of the population of the soon to be former colonies, the inclusion of women, as well as a feminist language within the declaration of human rights, and the role of the state in the protection of minorities. As Morsink (1999) points out, all the anti-discrimination language set out in the Human Rights Charter was exclusively introduced by the Communist bloc.

> 'More than any other voting pad, the Communist pushed from the very start form the inclusion of clear antidiscrimination language in the declaration. This non-discrimination stamp is their mark on the document' (Morsink, 1999, p.93).

The ability and perseverance on the part of the communist bloc to introduce anti-discrimination language suggests that, for the Soviet bloc, humans are not born under equal conditions. Rather, their lives are determined by socio-economic and political structures. For the historical materialist perspective, therefore, equality requires the recognition and consideration of the historical oppression of certain groups of people and seeks to do them justice. Consequently, the action of controlling hatred under this rationale would aim to transform the relationship between race (and/or another exploited subject position), law and power, and repair, prevent and/or condemn the effects that hatred has had on a material level for racialised and other oppressed groups and in this manner contribute to their emancipation.

In the 1990s, Matsuda considered that the First Amendment coexisted with slavery and it was used to defend colonizer privileges and preserve the status quo of structural discrimination that only becomes relevant or visible when oppressed minorities challenged the status quo with the aim of altering the ongoing power relation.

> The first amendment coexisted with slavery, and we still are not sure it will protect us to the same extent that it protects whites. It often is argued that minorities have great protection from the First Amendment and therefore should guard it with jealousy. We are aware that the struggle for racial equality has relied heavily on the persuasion of peaceful protest protected by the first amendment, but experience also teaches us that our petitions often go unanswered until protest disrupts business as usual and requires the self-interested attention of those persons in power (Matsuda, 1993, p.76)

In relation to equality, Matsuda argues that hate speech is eloquent and seeks to put the subject in a position of subordination. For Matsuda, the subject is challenged by a voice that aims to confine the subject into a social position of subordination (1993). As such, and as it occurred with the Soviet Bloc, the Critical Race Theory group considered that, in order to acquire equality, we need to name historically oppressed groups, repair the damage, and then establish measures that alter the current power relations to eliminate the capacity of the perpetrator to use and abuse their historical privileges. These arguments are echoed in more recent trends such as #blacklivesmatter.

The Social Justice approached aim to regulate expressions that perpetuate subordination while seeking to establish regulations that contribute to alter the power relations by recognizing, protecting, and repairing the material effects that historical social damage has created over certain particular groups. Only then, they would say, would equality be universal. Both the Soviet Union and Critical Race Theory group rely on the authority of the state or the university governments, respectively, to govern hate speech and define protected communities. Overall, this approach remains marginalised and does not have extensive literature, as opposed to the neutral viewpoint approach , which traditionally emerges from the Liberal Left.

3.3. Neutral Viewpoint approach

The neutral viewpoint approach is the most widespread and known approach to tackle hate speech. Its main characteristics are: to protect freedom of expression, to protect all individuals equally against discrimination, to protect general categories that reach everyone, such as race or gender. Neutral Viewpoint is founded on the notion of viewpoint neutrality and it regulates discrimination based on factual findings of who acts wrong and who is actively committing the action of discrimination. This approach emerged from the Western Bloc during the Drafting of Article 7 of the Universal Declaration of Human Rights, but it was not until the Campus Hate Speech debate in the 1980s that it was fully defined and it gained popularity.

This section starts by looking at how this approach understands equality. Morsink (1999) suggests that during the Drafting of Article 7 of the Universal Declaration of Human Rights, the Western Bloc was mostly influenced by the Right of Man and other liberal proclamations. Under this, the notion of equality relates to equality at birth and equality before the law, leading to the understanding that all individuals have to be protected and subject to the same regulation. Next, the section looks at how national legislations have adopted generic formulations to protect categories against discrimination and to protect all individuals equally. Next, the section looks at the notion of ‘Wrongdoing’ or ‘Act wrong’ (Altman, 1993, p.312). Finally, it looks at how this approach does not challenge structural hatred by exploring the notion of ‘reverse enforcement’ (Delgado, 1993, 2018).

It can be argued that, ideologically speaking, Neutral Viewpoint comes from the 1689 Declaration of Rights and the Rights of Man and of the Citizen (1789), which says in its preamble that, 'all men are born, and always continue, free and equal in respect of their rights'. Given the weight these documents have in Western political theory and philosophy, it comes as no surprise that they are also found in the 1948 Human Rights Chart, whose preamble states that 'all beings are born free and equal in dignity and rights' (Morsink, 1991). Is it not surprising neither that due to the weight of the Western Bloc on the Drafting of the Universal Declaration of Human Rights, the Chart systematically and insistently reminds us that all Articles are for equally protecting every single individual.

> Preamble, 'all people and nations'; Article 1, 'all human beings'; Article 2 'Everyone' ; Article 3, 'Everyone' ; Article 4, 'no one'; Article 5, 'No one' ; Article 6, 'Everyone' ; Article 7, 'All' ; Article 8, 'Everyone' ; Article 9, 'No one' ; Article 10, 'Everyone' ; Article 11, 'Everyone' and 'no one'; Article 12, 'No one' ; Article 13, 'Everyone' ; Article 14, 'Everyone' ; Article 15, 'Everyone' ; Article 16, 'All man and women'; Article 17, 'Everyone'; Article 18, 'No one'; Article 19, 'Everyone'; Article 20, 'No one'; Article 21, 'Everyone'; Article 22; 'Everyone'; Article 23, 'Everyone'; Article 24, 'Everyone'; Article 25, 'Everyone' ; Article 26, 'Everyone'; Article 27, 'Everyone'; Article 28 'Everyone'; Article 29, 'Everyone' , and finally article 30 which includes 'In no case' and 'Nothing'. (Morsink 1999, 129).

Article 7 of the Human Rights Chart is particularly relevant for the present case. As it was illustrated in previous chapter, Article 7 ‘Against Discrimination’ was the centre of intense debates between the Soviets and the Western Bloc. The draft accepted was the Western amendment to the Soviet proposal, which eliminated notions such as ‘privilege’ and which extended the right to all individuals regarding their place and role in history. The outcome of this process is that the compromise of the western bloc to the soviet bloc proposal is encoded in the majority of hate speech regulations and has been influential for all national legislations. The following paragraphs aim to unpack this statement:

As it can be observed, the final version for Article 7 against discrimination protects all individuals against discrimination and it situates the individual as equal before the law.

> '**All are equal before the law** and are entitled without any discrimination to equal protection of the law. All are entitled to equal protection against any discrimination in violation of this Declaration and against any incitement to such discrimination'. (Article 7 Human Rights Declaration, emphasis added).

Due to transposition of the Declaration of Human Rights to multiple national laws, a sample of the various regulations on hate speech tells us that most countries adopt the approach that seeks to protect all individuals equally against discrimination. For instance, regulation of hate speech in Spanish law is found in article 510 of the Criminal Code and prohibits the use of discriminatory expression against generic categories (2021). Incitement to hatred in Northern Ireland regulates what is known as hate crime and is limited to the act of discrimination or the act of inciting hatred. In any case, the protection offered in Northern Ireland is universal to 'all individuals or groups'. As a non-European example, South Africa established in 2016 the prevention of hate crimes and hate speech through a relevant bill. Due to its recent date, it is possibly one of the most detailed and updated laws. Despite this, and the South African history of apartheid, hate speech regulation continues to be generic in terms of protection, however. Indeed, it could be argued that the South African agenda is that of integration between black and whites, and, because of this, no group should be protected over the other. In any case, it is another example of neutrality and compromise, as it occurred during the Drafting of the Universal Declaration of Human Rights.

> **Any person** who intentionally publishes, propagates or advocates anything or communicates to one or more persons in a manner that could reasonably be construed to demonstrate a clear intention to be harmful or to incite harm; or promote or propagate hatred, based on one or more of the following grounds: age; albinism; birth, **color**, culture; disability; ethnic or social origin; gender or gender identity; HIV status; language; nationality, migrant or refugee status; **race** religion; sex, which includes intersex; or sexual orientation (South African n Government Gazette No. 41543 of 29 March 2018. (emphasis added).

Another example of universal protection can be found in the International Covenant on Civil and Political Rights (ICCPR), which enshrines protection from incitement to hatred in Article 20: 2.

> 'Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law'. (Article 20, International Covenant on Civil and Political Rights).

All the aforementioned examples illustrate that anti-discrimination law and hate speech law have a tendency to use generic categories or do not specify particular oppressed or affected groups but focuses on the 'wrong doing' as the western bloc proposed. Regulating wrongdoing is rooted on the Harm Principle, argued by John Stuart Mill, according to which, 'actions only should be limited to prevent harm to other individuals' (1859). This means that between equal individuals the only possibility is to regulate the act of discriminating itself (Altman, 1993, p.312). At its core, a wrong doing approach actively avoids naming minorities. In mobilising the notion of wrongdoing as a singular act, discrimination is not seen as an historical question of oppression or evidence structural discrimination, but as a question of individual behaviour that discriminates (McKinnon, 1987). In these terms, all individuals, regardless of the group they belong to, are capable of discriminating. It is not surprising then that hate speech regulation that punishes wrongdoing is written with words that do not include particular groups, but general categories such as race, gender, disability, rather than particular minority groups or historically oppressed groups.

This approach was formulated during Campus Hate speech debates in the 1980s and has gained popularity to the extent that it dominates the field, thanks to the vast literature created at the time. It influenced common and popular definitions of hate speech such as the following by Raphael Almagor:

> hostile, malicious speech aimed at a person or a group of people because of some of their actual or perceived innate characteristics. It expresses discriminatory, intimidating, disapproving, antagonistic and/or prejudicial

> attitudes toward those characteristics which include sex, race, religion, ethnicity, colour, national origin, disability, or sexual orientation (Almagor, 1993).

The problem of this approach, however, is that, it gives origin to what Delgado and Stefancik refers to as 'reverse enforcement' (2018). Reverse enforcement means that, if minorities are not clearly specified as subjects of protection, the same rules to protect hate speech can be used against minorities. The phenomenon of 'reverse enforcement' makes evident that Neutral Viewpoint ends up perpetuating forms of structural hate (Pincus, 1994), where rights are granted to minorities but where privilege is still protected. As a result, the power relations are not altered but perpetuated and, while the oppressed are given rights, the oppressor continues to gain in power. This position, therefore, allows men to speak out on international women's day in favour of abused men on the grounds that the problem of gender-based violence is 'common to all genders', as defended for instance by Santiago Abascal, president of the Spanish Far Right political party, VOX. As a result, the idea of racism or misogyny is torn from its historical basis, and stories of racialization or misogynism are relativized or silenced (Titley, 2020). It is, therefore, argued that this approach does not address hate; instead, it creates conditions that allow hate to re-emerge.

In sum, this section has explored the rational and political imaginary of the Western Bloc UN commissioners, as well as liberal left academics who intervened on Campus Hate Speech in the 1980s. Their views embrace the notions of Neutral Viewpoint individuals and equality before the law and seeks to regulate the act of discrimination independently of the group the target belongs to. Neutral viewpoint approach are more concerned with freedom than with equality and aim to punish the act of discrimination while leaving current power relations intact. Evidence of this is the 'reverse enforcement' phenomenon by which whites can accuse blacks of racism, or where women can be classified as 'misandrists'. This section has argued that, while the Neutral Viewpoint is the most widespread approach, it does not succeed in altering power relations and combatting structural discrimination (Pincus, 1994).

The following approach that this chapter presents: Freedom of Expression Absolutism Approach.

3.4 Freedom of expression absolutism approach

Freedom of Expression absolutism emerged during the Campus Hate Speech debates in the 1980s as a radical opposition to the implementation of hate speech codes or any form of regulation that limits freedom of expression. This approach considers that all individuals are equal before the law and that freedom of expression is a fundamental absolute right for all individuals that should not have any limitations. Freedom of expression absolutism first emerged as a Western proposal during the Drafting of Article 7 of the Universal Declaration of Human Rights, but it was not until Campus Hate Speech debate in the 1980s that it reached peak popularity. Those who opposed the implementation of hate speech codes during Campus Hate Speech in the 1980s founded their arguments on the notion and defence of freedom of speech and have been described by Delgado as Neoconservatives and Neoliberals (1993, 2018).

This section starts by looking at the First Amendment of the American Constitution, tracing its historical evolution and the defence it enjoys in the United States. The section also looks at the relationship between freedom of information and freedom of expression and at how interpretations of freedom of expression differ in other parts of the world, such as in Spain, where it is understood as a collective right (Garcia, 2004). Next, the section looks at the different theories that emerged in this period around freedom of expression absolutism, focusing on one in particular, the 'marketplace of ideas', before finally looking at Delgado's critique of freedom of expression absolutism and the marketplace of ideas (1993, 2018).

Freedom of Expression absolutism, as its name indicates, is rooted in the First Amendment of American constitution and the idea that the right of freedom and freedom of expression is an absolute right:

> Congress will not be able to make any law regarding the establishment of religion, nor prohibiting the free practice of it; neither by limiting freedom of expression, nor by press; nor the right to peaceful assembly of persons, nor to ask the government for compensation for grievances (First Amendment United States Constitution).

According to White, the First Amendment is for the individual and only the individual, as it holds the idea that singular beings are 'capable of giving individual meaning to their life experiences. It meant that humans had the potential - the freedom to alter those experiences' (White, 1996, p. 301). The influence of the First Amendment on the modern and contemporary world is undeniable, and yet, contrary to popular sentiment, freedom of expression and the cultural value of the First Amendment has not always enjoyed the popularity it enjoys today. Experts like Rabban (1994) and White (1996) point out that freedom of expression was rarely a public concern before World War I (White, 1996, p.300). In the interwar period, and because of the defence of workers who expressed their views on workers' rights, freedom of expression began to be used commonly in the courts as a legal resource. However, the philosophical rationale for protecting self-destiny gradually surfaced in free speech jurisprudence under the guise of individual autonomy (White, 1996). Gradually, First Amendment jurisprudence became progressively protective of speech (White, 1996, p.308). Encouraged by the triumph against the Fascism and democratic advancement, the United States made the value of freedom its national standard; the value that would overthrow authoritarian systems like the communist one. As White argues, between the Second World War and the 1980s, the democratic model of politics expanded politically, accompanied by a modern consciousness. This modern consciousness understands that, through the power of reason, human beings can give individuals meaning to their life experiences by having the freedom to alter those experiences (White, 1996). It is, therefore, the concept of freedom, and freedom of expression, that is intrinsically united with the concept of individualism and democracy (O'Flaherty, 2012). Therefore, there is no surprise that, under the triumphalist flag of freedom, codes limiting racist or misogynistic expressions are thus considered a limitation to the freedom of the individual. In addition to the right of Freedom of expression is the right Freedom of information. Traditionally, those two freedoms, with the addition of freedom of assembly, are grouped together as the freedoms that enable people to share ideas, form new thinking, and join together with others to claim their rights (McGonagle and Donders, 2015). However, there is a distinction between freedom of expression and freedom of information that it is important to note. Whereas freedom of expression relates to the capacity of the individuals to express their ideas feelings and emotions freely, freedom of information is when people 'seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media' (O'Flaherty, 2015, p.63). Therefore, the specificity of freedom of information is that it focuses in the media or channel that sustains the expression as also mentioned in Resolution 59 UN. 1946, freedom of information (General Assembly 1946).

The argument known as marketplace of ideas is particularly recurrent within the frame of literature that defends Freedom of Expression absolutism. This dominant metaphor is used to explain and frame freedom of expression absolutism views, where external limitations to public discourse should not be imposed because they compromise the free movement and debate of ideas that are necessary for a democratic public life. Indeed, the best, truthful, or persuasive ideas are those that survive the democratic process of debate (Heinz, 2016). Proponents of the marketplace of ideas argument suggest that knowledge can be produced by two opposing ideas confronting one another, that is counter acting or creating counter narratives; '[through the use of freedom of expression] public outrage over racist discourse will provoke a counter-reaction in which the mistakes and ugliness of racism, or some other evil, will be exposed and subjected to criticism and condemnation, leading to a healthier [democratic] situation' (Downing, 1999, p.176). As such, racism, 'or some other evil' awaits to be counteracted, since ‘more speech is the constitutionally-mandated remedy’ (Tribe,1988 p 834).

The action of counteracting hate speech with more speech has been criticised by Delgado (1994, 2018), who questions why the individual should take, in addition to the stress and harm that hateful expressions create (Clark, 1965), the responsibility to individually confront it when the argument was not attacking the individual but historically created. Indeed, arguments that favour the use of freedom of expression as a form of defence ignore the stress that it causes in the individual. The attack is physical and psychological, it creates fear and stress, and the target does not have necessarily the resources to use fight back with words to defend themselves (Delgado, 2018). Therefore, it is not realistic to expect that someone under stress, who may be having doubts of their own worth, should overcome the situation and elaborate upon a reasoning that will convince the perpetrator that the utterance is wrong (Bemberg, 2004).

In addition to Delgado’s critique, it is possible to add, what we can refer to as, the cultural critique of Freedom of Expression. It is undeniable that freedom of expression is a fundamental right that enriches society. It is linked to words, to the intellect, to political and social life, and to culture. However, it is important to note that, when we talk about freedom of expression, we are, not only talking about freedom to express our spiritual, artistic, or individual tendencies, but also talking about a legal tool, a social technology that not all cultures and societies interpret in the same way (Viejo Otero, 2020). If we look, for example, to the Spanish of the Democratic Transition (1970s and 1980s), freedom of expression was a form of civic protest that adopted variants such as demonstrations, rallies, strikes, marches, lockdowns, conferences or concerts as a 'significant' activity (Klandermans, 1989; in Garcia, 2004, p. 6) . Instead of an individualist sentiment, freedom of expression was understood as a collective action exercised by groups and organizations that aimed at achieving social awareness. And while it is important to point out that freedom of expression as an individual right took place during these acts, this was, according to author Gloria Garcia, a civic conquest of a collective nature (Garcia, 2004), instead of an individualist right as proposed by the United States.

In sum, Freedom of expression absolutist views do not explore questions such as equality but, instead, arguments around the defence of Freedom and Freedom of expression and its defence at all cost. Increasingly, this argument has gained adherents in Europe, though it has not fully penetrated in the area of European legislation. As illustrated in the following section, Europe has its own approach to hate speech, which can be considered a hybrid between historical materialist and Neutral Viewpoint es. The next section argues that Europe’s rationale and imaginaries is Eurocentric, as it protects harm that has occurred within European borders but does not take into account the harms that Europe has created, i.e. colonial expropriation.

3.5 European Approach

The European approach is one of the latest approaches to emerge and accompany the creation and needs of the European project. The main characteristic of this approach aligns with the values of the Council of Europe, which is the institution in charge of advocating for freedom of expression and of the media, freedom of assembly, equality, and the protection of minorities (Council of Europe, 2021). In addition, the European approach considers the element of justice, by which hate speech, racism, and xenophobia are included within the body of criminal law (Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law).

Comparative Studies between America and Europe differentiate the American approach to hate speech from the European approach (see Boyle, 2001). Europe proposes hate speech regulation that limits freedom of expression on issues concerning European history, such as Nazism and the protection of minorities, whereas the United States base their approach on the fundamental defense of freedom of expression and tolerance for the intolerant (Delgado, 2018).

In order to explore all the above, this section first explains the European position in relation to Freedom, Equality, and Justice. Next, it explains how the European approach to hate speech regulation is only concerned with groups within its borders, before concluding that this approach is a hybrid between the historical materialist approach and the Neutral Viewpoint , despite their inherent contradictions.

The Council of Europe explicitly defends freedom of expression and freedom of the media but with limitations. The political actor in charge of defining limitations to freedom of expression is the European Commission against Racism and Intolerance. It was established by the Council of Europe and is pan-European, i.e. it extends across the whole of Europe not only the European Union member states. The ECRI General Policy Recommendation N15 on combating Hate speech states that 'freedom of expression and opinion is not an unqualified right and that it must not be exercised in a manner inconsistent with the rights of others' (15th Recommendation 8 December 2015, 3) such as dignity, liberty, and security of the individual (Official Journal of European Communities, 1999, p.2). The 15th Recommendation recognises that, while all individuals are subject to discrimination, some are targeted more than others:

> 'that the duty under international law to criminalise certain forms of hate speech, although applicable to everyone, was established to protect members of vulnerable groups and noting with concern that they may have been disproportionately the subject of prosecutions or that the offences created have been used against them for the wrong reasons' (15th Recommendation 8 December, 2015, p.5).

In addition, the ECRI considered that groups and forms of language and expressions are alive and evolving. As a result, anything on the 15th Recommendation 'should be applied mutatis mutandis' to those changes, meaning that the matter of hate speech is not a matter of language, but, rather, meaning (European Commission, 15th Recommendation 8 December 2015, p.5).

The recognition on behalf of the European Union, that particular groups are more targeted than others, echoes the historical materialist approach. The difference, however, is that, whereas the historical materialist approach arguments are rooted in Marxist philosophical notions of universal social justice and equality, Europe is more concerned with social cohesion among European citizens, their liberty, dignity, and security as individuals and the integration of all European citizens under the same legal framework (Official Journal of European communities, 1999, p. 4) referring particular to the 2008 Recommendation, that criminalised hate speech.

In considering all the above, it can be argued the European approach aims to favour inclusion and equality of European members and protect them from the repetition of forms of discrimination that Europe has experienced. By comparison with the other approaches, it is rather tempting to argue that this is a nuanced and advanced understanding of hate speech. However, careful observation reveals that we are now experiencing the limits of this approach, particularly when referring to the case of refugees and asylum seekers who are confined in detention systems.

3.6 Conclusion

The present chapter aimed to complement Chapter Two and sought to identify the logics, imaginaries, and key values that underpin regulator approaches to Hate speech.

To arrive to this conclusion the chapter explored the different actors involved in the previous chapter and inductively identified four approaches to regulating hate speech: Social Justice, Neutral Viewpoint , Freedom of Expression Absolutism, and European Approach. The categorisation offered in this chapter will contribute to interpreted the analysis of Facebook as it is intended to understand if Facebook follows any of these categories or if as Klonic argues in the following chapter, Social media platforms are the new governor of speech (2018).

Social Justice socio-political imagination puts equality at the centre of their concerns and sees social hate as an obstacle to achieving it. Under this perspective, hate is structural and part of the social fabric; not merely an emotion but an integral part of the arsenal by which certain social groups are exploited and oppressed. As such, this perspective to regulate hate speech seeks to acquire equality by repairing and preventing the subordination of historically vulnerable groups. The second approach, Neutral Viewpoint , places the freedom of the individual at the centre. It tends to consider hate as an emotion expressed interpersonally (McKinnon, 1987) and punishes the act of discriminating (Altman, 1993) while advocating for equal protection against discrimination to all individuals. That is, under this perspective, all individuals are equal, despite the historical processes that have determined their current social position and trajectory as members of specific social groups. The third approach this chapter has explored has been Freedom of Expression Absolutism. As its name indicates, this approach considers freedom as an unqualified right. For the proponents of this view, freedom of expression is a good in itself, which all individuals enjoy equally, and which is seen as the perfect instrument to fight discrimination. Finally, the European Approach is a hybrid between the Neutral Viewpoint and the Historical materialist approach, though it only seeks to protect domestic minorities. The four approaches are summarised below.

Approach to achieve equality	**Equality and how to achieve Equality**	**Actors that adhere this approach**	**Management Operationalization and Enforcement**
Social Justice	Oppression is an obstacle to achieve Equality. It seeks to regulate discrimination and expressions of discrimination such as hate speech by protecting historically oppressed minorities.	Soviet Union Critical Race Theorist Groups Blacklivesmatter movement	The State has to Regulate and define minorities. Universities have to regulate. Legal systems might start including determined minorities.
Neutral Viewpoint	All individuals are born equal and they are equal before the law. It seeks to protect all individuals against discrimination or 'Act wrong' on the grounds of generic categories such as gender or race. In only regulating the act of discriminating, it widens the protection of freedom of expression.	Liberal Left academics. United Nations (Chart of Human Rights) High number of States around the world, including for instance South Africa.	Relies on the Juridical system.
Freedom Absolutism	All individuals are born equal and free. Freedom of expression is the coherent form to defend against oppression and discrimination.	Neoliberal Academics and Institutions. Far Right White Supremacist Groups.	No regulation
European approach	All European citizens are equal and no one should be discriminated against. There is special mention to European minorities.	European Courts and Must member states	Relies on the Judicial system and is included as part of Criminal law.

Table 1. Approaches of Hate Speech regulation

Out of the four approaches, the Neutral Viewpoint has influenced common and popular definitions of hate speech the most, such as the aforementioned definition proposed by Almagor (1993). As such, it is possible to wonder whether the liberal, universalist, individual, and neutral model that seeks to protect everyone equally has succeeded in fighting racial and other forms of social hate or, if instead, it has only worked towards protecting the key values of liberalism. Paraphrasing the quote by Simone Weil that has opened this chapter, I argue that the Neutral Viewpoint is intrinsically incapable of fighting discrimination as it does not protect minorities. Weil aimed to illustrate that humanity repeatedly attributes a role and a practical political function to ideas that are not always the perfect practical measure. This is perhaps what can be argued in relation to the Neutral Viewpoint whose intention might have been, since 1945, to protect all individuals against discrimination, but that, due to its liberal universalist and individualistic rationale, it is intrinsically unable to fulfil the role assigned to it, which is to fight against hate, structural racism and misogyny and subordination. Evidence of its incapacity to fight hate is the constant re-emergence of hate speech.

Racist and misogynistic speech has persisted and spread through different channels such as on social media. The emergence of social media platforms is perhaps far from the narratives that evolved during the cold war. However, they have to get involved in discussions of hate speech when hate on their platforms and their effects become a growing problem. Social networking platforms such as Facebook, YouTube, and Twitter have got involved in the governance of hate. However, when social media got involved in the governance of hate, the discussion was no longer an ethical and sociological discussion about freedom and equality, but a managerial discussion that revolved around how to deal with hate speech from our own laptops, desktops, and devices. According to this new conversation, hate speech regulation is no longer only a set of codes or rules that respond to philosophical and sociological considerations about equality and how to achieve equality, but a series of intervention mechanisms such as pop-up windows, community standards, settings and written policies which serve to limit or to expand how much hateful content can travel around the platforms. It also understands how long it is going to stay online, or how rapidly or slowly hateful content is going to escalate. Therefore, social media has added to previous ethical, philosophical and legal conversations on freedom and equality, a conversation about technology, management and operability, that adds a new layer of complexity to the concept of hate speech., but that to the best of my knowledge, it remains to be analysed if the technological layer transforms how hate speech is governed, or if rather Facebook reproduces , re-creates or re-adapts old models of governance. (see Sessen 2002).

Due to its novelty, importance, and complexity , academics have turned their sights towards the platforms, their policies, and the practices that platforms were putting in place to fight hate. However, it remains to be seen if social media in general and Facebook in specific is producing new ideas around equality and freedom or, if instead, they are reproducing some of the approaches that this chapter has exposed. With the aim of clarifying what is that we know, so far, in the relationship between hate speech and social media the next chapter will explore existing literature.

Chapter 4. Social Media. The New Governors of Hate Speech?

4.1 Introduction

Social media has added to previous ethical, philosophical and legal conversations on freedom and equality, a conversation about technology, management and operability, that adds a new layer of complexity to the concept of hate speech, therefore, the purpose of this chapter is to review literature that explores social media and how social media regulates hate speech . It seeks to understand if any present literature undertakes how Facebook governs hate speech by attending to Facebook values and principles.

A review of existing literature shows that academics have placed great importance on the analysis of user behaviour (Jubany and Roiha, 2018), content moderation (Roberts, 2016, 2019; Gillespie, 2018), the role of AI (Ruha, 2019), and the relationship between social media platforms and the law (Gorwa, 2019; Suzor, 2019). However, the analysis and exploration of current literature indicates there is a gap, as none of the present studies have examined in any detail the values that underline hate speech policies and enforcement mechanisms that govern hate speech on social media. Specifically, there is a lack of studies that explain how social media values freedom or equality and the influences of these values on the governance of hate speech and hateful content. This current study is therefore important because these definitions and their underlying values guide operational decisions that, in turn, structure the platforms in ways that determine user behaviour towards hateful content. In order to expand and explore the above, the chapter is organised as follows below.

The chapter opens by exploring the social media background and ideological origin of social media platforms, which the chapter locates in the American West Coast, as part of the 'Californian ideology' (Barbrook and Cameron, 1996). It then explores the rise of social media, the techno-optimistic period, where the abundance of hateful content greatly contributed to the end of social media glorification. The third section aims to explore literature that looks at how hateful content appears and circulates on the platform, and what social media has done to tackle it, namely Socio Technical Dynamics Research. Section four of the chapter aims to provide a legal framework and focus on the effects of the US 230 Decency Act and its implications and the European Code of Conduct . The fifth section looks at the body of literature on platform governance. Finally the chapter exposes a gap in the literature, and develops the theoretical research framework to the current study.

4.2 Californian Ideology

This section discusses the Californian ideology (Barbrook and Cameron, 1996), which is argued to underpin the formation, approach, and values of the tech industry, as exemplified by the Silicon Valley companies. It incorporates ideas from the counter-cultural movements of the 1960s, combined with tech utopianism and libertarian politics.

According to Barbrook and Cameron (1996), the participants of the movement known as the 'Californian Ideology' were artists, hackers, engineers, and entrepreneurs of a liberal hippy profile, for whom Cold War rhetoric and bipolarity in the world was rigid and pernicious. While some hippies resorted to living in nature, the most active participants became tech developers who sought to drive universalist rationality, hoping that technological innovations would materialize their beliefs. West Coast ideologists aimed to implement a new form of democracy, ‘where all individuals could express themselves freely within cyberspace’ (Barbrook & Cameron, 1996, p.45). The term "electronic agora" became popular in this context. Electronic agora is a term coined by McLuhan (1964), and it was considered a virtual place where everyone could express their opinions without fear of censorship; a place that would abolish spatial boundaries and everyone would relate to one another in a community. The hope was that new technologies would facilitate free and open access to information and lead democracy in all social institutions, thus ‘breaking with the narrow politics of the post-war era' (Barbrook & Cameron, 1996, p.47). As if it were the Promised Land, the electronic agora was going to end the bipolarity of the world established by the Soviet bloc and the Western Bloc. Those who built Silicon Valley adhered to this line of thought and connected computers and the internet to the free exchange of information. With the emergence of platforms, it can be argued that there was a cultural shift. The person behind the device no longer needed computer literacy and became a device holder and a service user in a digital media environment. Indeed, the result is not necessarily what McLuhan imagined, but was, at least, a new and mediatized world (Siapera, 2018), where the idea of freedom of information, universal and free access to information, was accompanied by the arrangements and dispositions of private technical structures called "platforms" (Bogost and Monfort, 2007, 2009; Gillespie, 2010).

Platforms are 'extensible codebase of a software-based system' (Tiwana et al., 2010, p. 676) wrapped by big tech companies in democratic rhetoric (Gillespie, 2018). Democratic rhetorical narrative led platforms to initially enjoy a period of techno-optimism, where the public perceived them as a social asset; as the creators of a new social order that will facilitate and advance democracy (Cohen and Raymond, 2011). While during the 15M movement in Spain, or the Arab Spring, platforms were praised as enablers of social change, a series of scandals contributed to the disappearance of this techno-optimism. The most prominent scandal can be considered to be the 2016 US election, where digital infrastructure facilitated attempts to manipulate democratic elections. In this context, the scandal of Cambridge Analytica revealed that the personal data of millions of Facebook users were extracted and used for political purposes (Guardian, 2018; Gorwa, 2019; Owen, 2019). During this period of intense political polarisation, the leaking of Facebook moderation training materials to the Guardian and ProPublica showed that the social media approach to hate speech regulation was random, private, and not satisfactory for the general public. Indeed, these series of events greatly contributed to diminishing Facebook, Twitter, and YouTube. Critics pointed out how social media was based on a ‘cyber-utopia’ (Morozov, 2014), which did not take into account that social media platforms are in fact private for profit corporations and which take users’ personal data while widely benefiting from their voluntary workforce: the user (Fuchs, 2011, 2014; De Nardis, 2016; Srnicek, 2017). It is not surprising then, that the general public commenced to look at platforms critically, particularly in relation to the abundance, visibility, and possibility of hateful content. At the centre of concern is that platforms are private corporations with a capacity to define and self-regulate hate speech. The following section explores two relevant legal frameworks that directly influence how Facebook governs and tackles hateful content, namely the 230 Decency Act and the European Code of Conduct.

4.3. 230 Decency Act and European Code of Conduct

Social media responded to the presence of hateful content online by developing their own community guidelines, in which hate speech regulation has three aspects: Hate speech definition (where social media explains what its definition of hate speech is), Community content guidelines (example of specific forms of content that is not allowed), and enforcement mechanisms such as the flagging technique. Platforms are not subjected to any regulation but are instead allowed to self-regulate, which has led to authors like Kate Klonick (2018) to argue that online platforms such as Facebook, Twitter, and YouTube are the ‘new governors of speech’. Content published in traditional media is regulated by national and international laws. However, online platforms based -or with license to operate- in the United States are protected under the "safe harbour" regulation (Gillespie, 2018, p. 30), granted by article 230 of the Communications Decency Act (CDA), by which: 'No provider or user of an interactive computer service provider will be treated as the publisher or speaker of any information provided by another provider of information content’ (Office of Law Revision Counsel, 2020). Therefore, Article 230 of the Communications Decency Law separates platforms from content, and gives online platforms the right - not the obligation - to self-regulate (DeNardis, 2014; DeNardis & Hackl, 2015; Marsden, 2011; Milosevic, 2018; Klonick, 2018; Gillespie, 2018).

A clear case like tobacco advertisements illustrates the extent and means by which social media self-regulates. Tobacco advertising on television and sporting events is regulated in Europe by Directive 89/552 / EC, known as the Television without Frontiers Directive (European Union, 2002). The objective of this directive 'is to guarantee the free circulation of broadcasting services in the internal market of the European Union, and, *at the same time, to preserve certain objectives of public interest*' (emphasis added) (Council Directive 89/552/EEC). Under this law, all tobacco advertisements in Europe are forbidden in order to protect public health. However, and due to the protection granted by Article 230 of the Communications Decency Act, online platforms do not have the obligation to regulate in the same manner but the right to self-regulate on tobacco advertisements. As a result, there is no uniform policy on tobacco across all platforms, and while some platforms simply do not mention tobacco in their regulations (i.e.: Reddit) others have strict and detailed regulation around tobacco advertising (i.e. Facebook).

The fact that online platforms designed their own regulations is a matter of particular interest. The significance of the 230 Decency Act is that social media platforms who get involved in the governance of hate speech become, not only their own regulators but also executors of their own regulation. The problem, however, is that, ultimately, platforms are private institutions whose decisions 'will be financially motivated on forms and shapes that we might not recognise' (Gillespie, 2015, p. 1). As such, it is of interest to know how online platforms regulate or intervene in matters that were previously in the public domain (Hestres, 2013; DeNardis & Hackl, 2015; Gillespie, 2018; Milosevic, 2018).

An important observation is that policy officers on social media have the opportunity to draw from previous legal documents, whereas as executors of their own regulation, social media has particularities that emerge from their own technical nature and that may directly affect how they execute their own policies. To unpack this further, consider the following. The Soviet Union, the American Campuses, and the European Union can rely on their juridical systems which have well established procedures. However, on social media, hateful content, like any other type of content, takes the form of a set of recordable, reproducible, and achievable digital physical data that travels fast, far, and lasts for a long time after it is first published. As such, executing hate speech regulation has to address a problem of volume and scale arising from the wide circulation of content beyond local jurisdictions, and must develop appropriate intervention mechanisms that serve to limit or expand how much hateful content can travel, how many times it can be shared, how long it will remain online, how many times a video will be played, and how fast or slow hateful content will increase. Given these real and pressing issues, who is to say then that platforms will not regulate, firstly, in order to ensure their continuous operation and only secondarily (if at all) in order to address racism and discrimination?

Platforms cooperate with public legislative bodies, as illustrated in the agreement signed the EU Code of Conduct 'on countering illegal hate speech online' (European Commission, 2016). The European Commission is the body in charge of ensuring that Code of Conducts are followed. To monitor this, the European Commission relies on the work developed by NGOs known as Trusted Flaggers (European Commission, 5th Fact sheet 2020). The Code was signed in May 2016 by Facebook, Microsoft, Twitter, and YouTube. Instagram, Snapchat, and Dailymotion joined the Code in 2018 and, in September 2020, Tik Tok announced their participation. The Code specifically refers to illegal hate speech, as defined by the 2008 framework decision, and functions as a complement to European legislation fighting racism and xenophobia (Council Framework Decision 2008/913/JHA). By signing this code, companies committed to continuing efforts to reduce hateful speech content from their platforms using measures that online platforms design themselves. These measures include the internal procedures already in place, and extra training to staff to ensure the examination of all notifications of hatred within less than 24 hours. If necessary, online platforms remove such content and disable access to users, in case they breach social media community standards. In addition, for all the content that is controversial, but does not violate social media community standards, companies agreed to work and promote counter-narrative strategies and closely collaborate with NGO and civil society groups (European Commission, 2016). According to the European report form 2016-2019, the Code of Conduct has contributed to reviewing and removing large volumes of hate speech content. In 2016, social media removed a total of 28% of hate speech within 24 hours; in 2019, the percentage increased to 72%, and in 2020, on average, 90% of the notifications were reviewed within 24 hours, with 71% of the content removed (European Commission, 5th Fact sheet 2020). These numbers were looked at by the European Commission as positive, adding that the agreement 'has increased trust and cooperation between IT Companies, civil society organisations and Member States authorities in the form of a structured process of mutual learning and exchange of knowledge' (European Commission, 2019). The question, however, is why these results are framed as 'positive results'? Are public institutions legitimising social media practises? Indeed, in light of this framework, platforms are required to take content down, with platforms following accordingly. However, as it has been argued earlier, despite some important concessions to history, for example holocaust denial, the European Union's compromise approach doesn't address hateful content in their entirety. In addition, while they may have been made to take down illegal content in Europe, they do not necessarily follow similar procedures in other regions of the world. Therefore, the Code of Conduct is a very limited tool in addressing hate speech, with a very narrow focus. However, more importantly is the complicity between European Union and social media platforms. Whereas the code positively contributes to take down and reduce the presence of hateful content online, this strategy encloses a unidirectional idea proposed by some academics and platforms and availed of by the European Union, which is that hate speech is a problem of scale. This notion I argue feeds a persistent industry by which content is systematically created by users and then, subsequently, deleted by platforms. As such, whereas the Code of Conduct contributes to reducing hateful content online, it does not show signs of solving the root of the problem; that is, they do not point out, nor transform, the possibility for hateful content to be online.

In sum, this section highlights that platforms regulate themselves and enforce these regulations without external and publicly regulated oversight that have the capacity and resources to keep up with the dynamics and measures imposed on social media. It raises the question of how we can trust them to regulate on the basis of the actual problem of hate speech, as opposed to regulating on the basis of what is expedient for platforms.

Whereas the previous sections have explored social media origins and frameworks that shaped how social media governs, the following two sections explore two academic trends that look at hate speech on social media. The first section will explore Socio Technical Dynamics Research, which focuses on the abundance of hateful content on social media platforms, by looking at users' behaviours in relation to technical affordances. The second section will explore research on platform governance, which looks at social media from an architectural point of view, to understand platforms as social actors with responsibility for how speech is governed.

4.4. Socio Technical Dynamics Research

Hateful content contributed to the end of techno-optimism. In order to understand and denounce the presence of hateful content on social media platforms, research has focused on the particular 'socio technical dynamics online' (Gillespie, 2015), which refers to the study of users' behaviour on the platform, taking into account both social dimensions and technological features. A particular area of this field of research has focused on individuals who create, spread, and receive hateful content through social media platforms.

An important observation when it comes to hateful content is that online hate precedes the creation and formation of Facebook, Twitter, and YouTube. Early research conducted by Les Back points out that, despite the hopes put in cyberspace, new technologies reinvigorated racism and gave racist groups leeway to organize and operate beyond national borders (2002). Indeed, far-right groups were gathering in 'alternative media' (i.e., websites) (Atton, 2001) and gendered cyber-hate was already present in the early 1990s on gaming platforms (Jane, 2018). As such, when online media platforms like Facebook, YouTube, and Twitter appeared, hateful groups only changed the channel to gradually exploit the architectures and uses of social media to continue disseminating hateful content.

Ever since their appearance, social media platforms have empowered organised racist and supremacist groups by enabling them to exchange know-how, to develop a common vocabulary, to mainstream their ideas, and to both recruit members and spread hate 'further, faster, and wider' (Jubany and Roiha, 2018, p. 181; Froio and Ganesh, 2019). However, hateful speech is also generated by people who have no connection with hate groups (Tynes, 2006), particularly after 'trigger events'. Trigger events refer to all series of events offline that involve a hated group and the coverage of those events (Siapera, Moreo, and Zhu, 2018). Evidence points out that coverage of hateful events in the mass media and small news sites contributes to increasing hateful comments and their circulation across platforms (Jubany and Roiha, 2018).

Researchers point out and label hateful content on online platforms as a considerable problem. A study conducted in Finland shows that, out of a total of 723 Facebook users, 67% were exposed to hate speech and 21% were targeted individuals. Of the respondents in that study, 70% had accidentally reached hateful content, indicating how easily this type of material can be accessed. In addition, the study also indicates that Facebook and YouTube are the most common sites where hateful content is found (Oksanen et al., 2014). Research led by Plan International (2020) also points to Facebook as the platform with the greatest number of hate incidents. Plan International (2020) has conducted an online survey accompanied by in-depth interviews. The NGO produced an online questionnaire that reached a total of 14,071 girls living in 31 different countries across all continents. Some surveys were followed by 18 in-depth interviews. The study shows that 58% of these girls reported that they had personally experienced some form of online harassment based on their gender, nationality, race or sexual orientation. In terms of platform-to-platform distribution, 23% of girls and young women reported harassment happening on Instagram, 14% on WhatsApp, and 39% on Facebook.

A particularly illuminating case is the research carried out by Mondal Silva and Benavenuto (2017) who analysed a total of 512 million Tweets from a random sample. Researchers compiled public data available on Twitter between 2014 and 2015, with the aim of looking at the percentage of hateful content within the sample. The authors, for operational reasons, defined hateful content as ‘all offensive publications motivated by a writer against an aspect of a group of people' (2017, p.87). This definition is relevant, as it follows that the authors did not adhere to any established definition of hate speech. Instead, they sought to find what Siapera, Moreo and Zhou have referred to as 'racially loaded toxic speech’; that is, ‘contents that entrench polarization; reinforce stereotypes; spread myths and disinformation; justify the exclusion, stigmatization, and inferiorisation of particular groups', the main characteristic of which is that it is not categorized as hateful by social media platforms (2018, p.3). Therefore, and with the aim of comparing what Twitter consider hateful with the amount of ‘loaded toxic speech’, Mondal, Araujo-Silva,and Benavenuto analysed the same number of Tweets by using filters offered by Hatebase.org (a database that collects hate speech examples circulating on the network) and then analysing the same data based on their own definition of hateful content. When researchers used Hatebase.org's words, they identified 20,305 Tweets with hateful content. However, when researchers used their own definition, they found that 65% of the 512 million of tweets fell within it. Therefore, this research shows that the problem of hateful content on social media, not only appears as direct and explicit expressions of hate, but also takes the form of denigrations and threats against persons or groups without the use of any racial slur or epithet. This can happen even when adopted in a loving way: ‘I love my country clear and pure’ (Les back, 2002), or under commonly accepted formulas that are not considered hateful or which breach social media community content guidelines (but which still segregate or are considered discriminatory); for example, patriotism on Facebook (Siapera, Viejo-Otero, 2017). As such, it is possible to argue that, whereas academic research and social media platforms agree that hate speech is a serious problem, it seems that they differ on what constitutes hateful content. In addition, studies in the field indicate that hateful comments generally target individuals on the basis of race, gender, sexual orientation, and gender focusing on blacks, gays, and women. Thus, it can be argued that although no new kind of racism, misogyny, or homophobia appears on social media, new forms of targeting the same historically oppressed groups do appear (Siapera and Viejo, 2020).

The following section explores a different trend of research, platform governance that takes a different approach, as it studies social media platforms from what can be considered an ‘architectural’ viewpoint.

4.5. Platform governance

In 2015, Gillespie wrote a brief statement or manifesto called "Platforms intervene", which opens with the following paragraph:

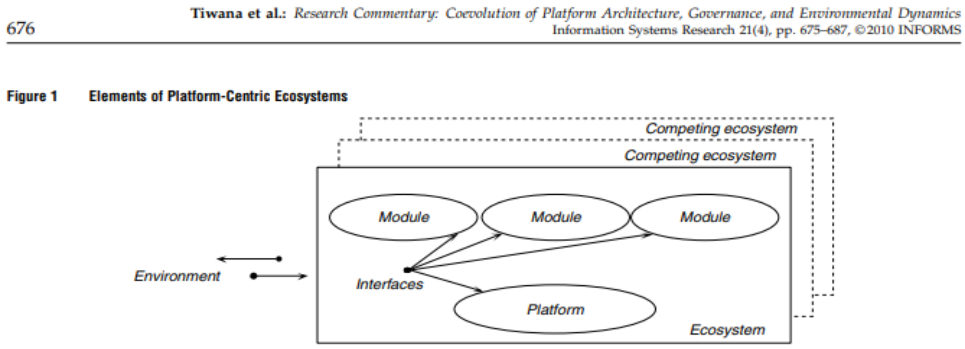
> It is still tempting to study social dynamics on platforms while ignoring the platforms themselves, treating them as simply there, irrelevant or designed in the only way imaginable. But recent work on the technical partner dynamics, context-specific realities and political economic dynamics of social media have made it clear that platforms, in their technical design, economic imperatives, regulatory frameworks and their public nature, have different consequences for what users can and, in fact, do (Gillespie, 2015, p.1)

Platform governance is a line of academic research that seeks to understand the effects that platforms are having on society, and the responsibility that platforms might have on their own regulation, while also seeking to overcome the limitations of studying the behaviour of users on the network (Gorwa, 2019; Gillespie, 2015). Gillespie notes that, given the protection that 230 CDA offers, among all possible decisions, social media could have chosen not to govern (2018). However, in the balance of running the platform for the user, sustaining the business model, and engaging with external public bodies, large technology companies started to govern hateful content, entering what Foucault refers to as a governance problem, understanding governance to be the set of techniques and power relations that aim to intervene in the life of individuals, or users (in the present study) (2007). Platform governance is a complex concept (Gorwa, 2019), which in this particular field of research refers to:

> the layers of governance relationships that structure interactions between key parties in today's platform society, capturing the growing body of work addressing the political effects of digital platforms, as well as the complex challenges that the governance of platform company’s presents (Gorwa, 2019, p. 855).

Platform governance looks at the architectural side of online platforms, their means of governance, and the impact of those structures in the social environment. Platform governance encompasses several topics of research such as: the architectural side of platforms, the data ecosystem, algorithm governance, and the influence of external regulation.

The article published by Tiwana et al. (2010) is a relevant technical example of research that looks at social media from its architectural side or technical construction. Tiwana et al. define platforms as 'the extensible codebase of a software-based system' formed by modules. Modules are all the 'add-on software subsystems' (Katz and Shapiro, 1994; Tiwana et al., 2010, p. 676) connected through interfaces, which are 'designed in a particular way that describes how the platform and modules interact and exchange information' (Katz and Shapiro, 1994; Tiwana et al., 2010, p. 676).



676

Tiwana et al.: *Research Commentary: Coevolution of Platform Architecture, Governance, and Environmental Dynamics*
Information Systems Research 21(4), pp. 675–687, ©2010 INFORMS

Figure 1 Elements of Platform-Centric Ecosystems

Figure 3 Coevolution of Platform Architecture, Governance and environmental dynamics (Tiwana et al., 2010, p.676)

Overall, platforms, modules, and interfaces are a technical ecosystem embedded in society, which Tiwana et al. refer to as 'Environment'. Their main argument that the authors pursue is that technical ecosystems-platforms ultimately depend on who owns them and who designs them, who, they suggest, might ultimately 'retain sufficient control to ensure the integrity of the platform and relinquishing enough control to encourage innovation by the platform's module developer' (Tiwana et al., 2010, p.675). Tiwana et al.'s perspective, therefore, offers us a techno-organic vision of platforms that emphasize how social media is technically constructed while highlighting the importance of the role of their owners in the construction and functions of the platform. Tiwana et al. (2010) point out that the environment (i.e. the social fabric) and the ecosystem (e.g. the platform, Facebook, Twitter etc.) mutually inform each other. The relationship between the 'environment' and the 'ecosystem' is of great interest to scholars, who have focused on how social media platforms construct or mediate this relationship. The work developed by Jose Van Dijck (2013) has proved especially fruitful. She critically analysed the Terms of Services and Policies of Platforms. Social networks establish a contractual relationship with their users through the Terms of Service (Van Dijck, 2013; Suzor, 2018). Policies derive from Terms of Service. On them, platforms cover a wide range of topics such as users' rights to connect, inform, and be informed, as well as social rules on property rights, identification, privacy, safety, and penalties for misconduct (Van Dijck, 2013). Policies do not follow any particular national law (Skinner, 2011; Paslawsky, 2012) and they constantly change and update (van Dijck, 2013). However, in Europe, following the General Data Protection Regulation EU 2018, every policy, security measure, or change in the privacy settings has to be notified to the user, who can accept, manipulate, and reject the particularities of how data is shared with third parties. Despite all these recent innovations, what has not changed is the fundamental transaction between the user and the platform established in the Terms of Service, by which the user accesses a service in exchange for her/his data. This is because the fundamental economic source of social media platforms is data (van Dijck, 2013; Srnicek, 2017).

Online platforms create and perpetuate a data industry whose dynamics influence how online platforms govern. In conjunction, the data industry and the platforms conform to a 'data ecosystem'. Data ecosystem refer to 'the architectural components and operational models that together extract, analyse and traffic with data' (Demchenko and Membrey, 2014). In the online environment, some platforms are dedicated to more general themes (i.e. Facebook), while others are topic specific (i.e. LinkedIn). Waller and Anderson (2019) point out the differences between generalist users and specialist users and indicate that, whereas generalist users on generalist platforms tend to produce more diverse data, platform that focus on specific kind of content (i.e.: Linkedin) tend to produce less data.

For the data industry, a single unit of data has multiple uses. For example, the data that a sports app generates can be used by outdoor clothing companies or biomedical corporations. Like oil, data is a material that must be extracted, refined, and used in various ways. Their uses are endless and social media does not rule data out (Srnicek, 2017). Cheney-Lippold refers to this as a 'datification-dominated ecosystem' (2017, p. 192), by which online platforms transform part, if not most, of our lives and our narratives into computable data that social network platforms, such as data processors, recombine and reuse endlessly for their own benefit. Therefore, social media platforms are mostly data extractors (Srnicek, 2017), with a tight, if not dominant, relationship between the data ecosystem and the ways platforms govern (van Dijck, 2013; Srnicek, 2017).

Gillespie argues that, at the centre of social media governance, is the activity of moderation (2018). Broadly speaking, moderation is the 'governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse' (Grimmelmann, 2015, p. 6). The centre of moderation is, for Gillespie, the community standards guidelines (2018), which regulate how users should behave on platforms in a friendly and detailed manner (Milosevic, 2018; Gillespie, 2018). As community standards call for user collaboration, in appearance they do not establish a vertical order-to-command relationship but that of 'governmentality', by which the user is willing to participate in the regulation of the platform by regulating themselves (Foucault, 2007). Per Community Standards, almost all popular social media 'prohibit hate speech' (Gillespie, 2018); although, as with the tobacco example above, not all platforms defined hate speech in the same way (Alkiviadou, 2016).

Users are expected to, both respect the Community Standards and report the presence of hateful content. In order to report hateful content, users have a series of interphases available known as 'flagging' (Gillespie, 2018, p.87). Flagging refers to the series of steps that the user has to go through in order to report hateful content, and it is composed of buttons, symbols, and popup windows that allow the user to select the type of content they are reporting. In addition, users have different interfaces available, referred to as settings, that appear beside each single comment and that serve to block, hide, or ignore comments (Gillespie, 2018). Langlois (2013) has looked at the different sets of symbols, protocols, software, and cultural practices implemented by social media platforms interfaces. For the author, the governance process focuses on how platforms create meaning, which works through different techno-cultural articulations like establishing equivalences between technical processes (i.e., reporting a comment) and cultural practices (i.e., combating hate speech). Therefore, flagging forms part of the governing system of these media platforms and, as a result, when studying governance, Langlois puts in evidence that buttons and interfaces are relevant subject for the study of governance as their design can culturally inform us of the ideology they held.

All flagged reports go to human review teams who systematically review flagged hateful content. Content reviewing systems have been of major interest to scholars of platform governance. Authors have researched the role of hashtags in moderation (Gerrard, 2018), the effects of social media moderation in social media engagement (Cao, Meadows, Wong, Xia, 2021), users' reaction to moderation decisions (Myers West, 2018), the inconsistency between moderation decisions and norms (Jubany and Roiha, 2018), and the figure of the moderator (Roberts, 2019). Sarah Roberts (2016, 2019) research is particularly interesting, as it describes the work of human review moderators as if it were a factory discarding damage production, their work consists in the systematic and mechanical review of the most violent, explicit, and radical expression of hatred on the platform. Commercial Content

Moderation (Roberts, 2016) resembles the work carried out by call centres and customer care centres, which are industries known for lacking job stability and where workers suffer stress and ethical and emotional damage as a result of being exposed to a systematic display of hateful and violent content (The Verge, 2019). According to Roberts, some workers carried out their duty, 'putting values aside' (2016) and continued comparing the material they reviewed against given community content guidelines. Reviewers are separated by teams regarding their language and culture, and they receive an urgent list of material that directly affects the group to which they belong. Numerous reports corroborate the moral and mental harm that human groups of moderators suffer and who do not articulate their complaints as discrimination, but as labour instability. It is, therefore, possible to point out that, within the social media industry, hatred has been industrialized, and gathered in offices, where workers are responsible for eliminating the most derogatory expressions from Facebook, and which content, whereas does not attack directly as individuals, affects them as group. In addition, human moderation activity has led to the creation of a vast industry. Multiple companies have specialized in content and content moderation for large and small platforms. The problem of hate (and other kinds of problematic) content has been industrialized, giving birth to multiple companies that offer remote work of moderation and classification of content, as well as companies that develop detection systems of 'objective content', such as applift.com.

Research points out that the combination of flagging report systems and human moderation has not been entirely efficient. In the first place, research points out that individuals are not inclined to user flagging report systems as much as social media platforms would have desired. According to Siapera, Moreo and Zhu (2018), there is what can be referred to as 'diffusion of responsibility' (Darley and Latane 1968), since users and bystanders consider that it is not their responsibility to report hateful comments nor consider it useful. In the second place, there is a problem of volume of hateful content (Gillespie, 2020), in which users do not have the capacity or the will to report and review the volume of hateful content online. Thirdly, there is the added problem of security, which has become a major issue in moderation activity. This was particularly evident after the 2019 White Supremacist Attack on Al Noor Mosque in Christchurch, New Zeeland. During the following 21 hours, the video and versions of that video were uploaded at least 1.5 million times (Independent, 2019), which would make human flagging and moderation insufficient measures for a problem of such magnitude. As a result, platforms like Facebook, Twitter, and YouTube have implemented what Gorwa, Binns and Katzenbach define as 'algorithmic commercial content moderation' (2020, p.3), which mostly consists on algorithm solutions to remove, geo-blocking, reduce visibility, distribute content or take down content (Gorwa, Binns, Katzenbach, 2020; Elkin-Koren and Maayan, 2020).

Algorithm content moderation has raised numerous critiques. O'Neill famously argues that 'algorithms are opinions embedded in code' (2016, p.21). Our actions online are shaped by algorithms decisions that do not necessarily respond to our desires but are shaped by the information that the algorithm has been feed with. As such, a petition or a post made on Facebook will make a judgment of how to use it better, targeting selected friends, and, therefore, governing and intervening in our personal communication intentions. Benjamin (2019) explores the bias with which developers build algorithms. Developers do not consider the privilege structures in which they are embedded, such as white privilege, and re-produce them in their algorithms, as illustrated by the experiment commonly known as Mitch McConnell's vs. Barack Obama's face. In September 2020, a non-academic experiment showed evidence that Twitter algorithm favoured the circulation of Mitch McConnell's face (white) and hid Barack Obama's face (black). Benjamin's research, therefore, aligns with Tiwana et al. (2010), where the authors argue that ownership matters. That is, even when algorithms might have evolved by their own means, there is an initial code or owner signature that settles its function or direction.

Of a different order, within platform governance, scholars have also looked at how external regulation affects how platforms govern speech. Klonick (2018) has looked into the influences of the First Amendment and how social media governs speech. The author, whose central argument is that social media are the new governors of speech, concludes that the analogy established between online platforms and the First Amendment should be 'largely abandoned' (2018, p. 1603). Even when influenced by the principle of freedom of expression, platforms ultimately self-regulate hate speech. Also, Suzor (2018) explores the Terms of Service as a contractual public mechanism that can be subject to the general rule of law. Following this, Suzor's body of work explores how to regulate platform governance, through both 'legal rules and social obligation' (2018, p.9).

In sum, Platforms Governance is a field of study shared by scholars from different fields such as media, computational science, law, culture studies, and sociology. Platform governance does not focus on socio technical dynamics that are the result of user behaviour and platform technological affordances. Instead, this trend of research targets platforms from an architectural, legal, cultural, technical and computational angle and aims to understand how online platforms govern content and how it affects the social fabric. As it has been seen, platform governance literature focuses on how platforms are technically constructed and the role of ownership, which is highlighted by Tiwana et al. (2010). This section further explained the relationship that platforms develop with their users through the Terms of Service (van Dijck, 2013; Suzor, 2018), the role of Community standards, the use of flagging reporting systems, the activity of human activity, and the algorithms employed for moderation (Roberts, 2016; Gillespie, 2018, 2020; Gorwa, Binns, Katzenbach, 2020; Elkin-Koren and Maayan, 2020).

Platform Governance researchers commonly criticise the lack of transparency and the lack of explanation for the continued inconsistencies and contradictions between the set of rules and how those rules are operationalised and implemented (Jubany and Roiha, 2018, Gillespie, 2018; Suzor, 2018). This lack of transparency has academics moving in circles, conducting micro-analysis and searching for clues that explain what social media policy officers have in mind, or what kind of information they possess that we, as researchers, do not. As a consequence, studies tend to be antagonistic, descriptive, or even complicit with social media platforms. It seems evident to me that there is a lack of broader understanding that explains how all those tools are designed, of what the overarching purpose of these techniques is, and what the actual governing tendencies that social media platforms perform are, not at technical but at an ideological level. In considering this, the following section proposes to look at social media as actors who have the same agency as the European Union, American University, or the Soviet bloc, and whose principles and values shape their technical tools and explain their governing intentions.

4.6 Gap in Literature

Foucault, understands governance as all the infinitesimal relations and power relations that control and shape order behaviours and actions (2007). However, it is also Foucault who pointed out that power relation reproduces previous systems of knowledge (2002). In these terms, no matter for instance how much agency that Facebook and other companies grant to their workers, its workers and designers are already embedded and reproduce prior systems of knowledge, values and beliefs (see Guha 2019).

Literature does not so far explored what are the systems of knowledge and beliefs that Facebook creates or perpetuates and how they inform Hate speech policies. This constitutes a gap in platform literature. Indeed, for Miller and Rose, who base their studies on Foucault, the study of principles and values is crucial. In particular, the authors indicate that the study of governance involves the understanding of all possible techniques that intervene in the life of individuals, the knowledge that shape the design of those techniques, and how in turn the techniques secure this values. (Miller and Rose, 2008). In considering the analysis conducted on chapter two and three, the fundamental concepts that historically shaped how hate speech regulation is approached are equality and freedom. As such, and considering that Facbeooks does refer to the regulation of hateful content as 'hate speech', the dissertation seeks to principally understand how Facebook understand and define Freedom and Equality, and what other set of concepts Facebook is creating and that influence how the platform governs hate speech.

In addition to Miller and Rose, among the few who have looked in the ideological dimensions of governance of platforms is Terranova (2015), whose work is explored below. Terranova argues that Foucault's analysis of liberalism and neoliberalism produces a peculiar reading of social networks. In particular, she uses the notion 'securing the social' (2015, p.112). By securing the social, Terranova understands the need that social media platforms have to motivate the creation of content and ensure that users stay on the platform by 'maximizing circulation, minimizing error or loss and ensuring an overall expansive stability (Terranova, 2015, p.124), tasks that are achieved through the implementation of 'dispositives of security'. While security is understood as the need that emerges in liberalism to secure freedom Since securing freedom for liberalism does not accept coercive measures, the result is a display of mechanisms and techniques that control and secure the freedom of circulation of people and goods, i.e. the police on a public park, the way pathways and benches are arranged in the park, and the signs or lack of signs that says 'grass is growing, please keep off the grass'. These are all governing techniques that regulate and govern the park by looking at how people circulate. This idea of how spaces are securitized and how circulation is controlled works as an analogy of how social media regulates content. The theoretical supposition, therefore, is that social media platforms have built a governing apparatus that has led to the creation of a 'dispositive of security', whose ultimate aim is to protect freedom of expression while maximising the possibilities of safety of both users and content. There is a need therefore to study the dispositive that Facebook has created in order to ensure and control the flow of content and the role of this dispositive within Facebook’s apparatus of governance. This analysis is important, because on the one hand it allows us to test the extent to which Foucault's ideas on security, territory, and freedom are valid in the new digital territories that have emerged in the 40 years since he gave these lectures; and on the other, because this analysis may reveal the rationale behind Facebook's definition and approach to hate speech, which, if Foucault is right, is far removed from the achievement of social justice or the philosophical motivations and drivers of the positions we identified so far in the previous chapters.

4.7 Conclusion

This chapter aimed to explore a new actor involved in the governance of hate speech, social media platforms. It has looked at its origins and reviewed relevant literature that explores the presence of hateful content and how platforms govern hate speech. In addition, the chapter has identified a gap in the literature, and has described the theoretical framework that empirical research embeds.

The idea of a global digital agora emerged in California as a form to overcome the rigidity imposed by the tensions between the soviet and the Western Bloc. The Californian Ideology (Barbrook and Cameron, 1996) was forged in the 1970s and it associated freedom, individualism, and equality with technology. One of the main influences of this movement was the idea of electronic agora devised by McLuhan (1964), where all individuals will join together with no geographical or physical borders through their individual devices. As if it were the Promised Land, the electronic agora was bringing to an end the bipolarity of the world imposed by the Cold War. As such, those adhering to this line of thought connected computers to the free exchange of information on the so called platforms. The result is not necessarily what McLuhan imagined, but at least a new and mediatized world (Siapera, 2018), where the idea of freedom of information, freedom of movement, universal and free access to information, is mediated by private platforms.

Platforms initially enjoyed a period of optimism, until 2016 when a series of scandals such as the United States 2016 elections and Cambridge Analytic showed that platforms were not the social change generator that was expected, but instead were private companies influencing the social fabric for their own benefit (Fuchs, 2014). As such, the general public, press, and academia began to take a closer look at how platforms were acting and the problems that they were perpetuating.

Among academics concerned with hateful content on the platforms, a body of work emerged that looked at the relationship between users, technical affordances, and how they contribute to the creation and propagation of hateful content. This trend of research has been coined by Gillespie as 'socio-technical dynamics research' (2015, p.1) and it commonly agrees upon the fact that social media has a problem with the volume of hateful content it fosters.

Scholars concerned with how platforms tackle hate speech initially found out that social media in the US enjoys legal protection through the 230 Decency Act. By virtue of this law, platforms have the right that not the obligation to self-regulate. In the case of Europe, scholars have been looking at the cooperation and agreement between the European Commission and media platforms known as Code of Conduct. The European Code of Conduct aims to complement Council Framework Decision 2008 to combat Racism and Xenophobia and stipulates that all platforms, that voluntarily adhere to the Code, have to take harmful content down in a maximum of 24 hours.

In considering that platforms have the privilege to self-regulate, a growing group of researchers have focused on how platforms govern. Platform governance research understands that social media includes actors with agency, values, and motives that influence the social fabric by their means of governing platforms. Gradually, researchers of this field have uncovered some of the mechanisms and techniques that social media platforms use to govern. In this context, Tiwana et al. offered a technical explanation of how platforms are organised and the role of ownership in the design of platforms. Jose van Dijck offered in 2013 one of the first systematic studies of Platforms Governance, where she focused on the study of Terms of Service, an issue later developed further by Suzor (2018), who focused on regulating terms of service from a general law perspective. Both van Dijck and Suzor observe that social media established, with the terms of service, a contractual relationship that determines how users should behave and what kind of content they can upload. The specific set of rules and regulations, along with the mechanisms to enforce the rules, are further studied by authors such as Roberts (2019) and Gillespie (2010, 2015, 2018, 2020), whose persistent work in unveiling social media governance of content have provided valuable insights around the role and value of Community Standards, the activity of users and human moderation and the developments of automatic detection to moderate content. Together, the range of scholars who have put their efforts into the study of social media platforms from an architectural viewpoint have provided a detailed analysis of how social media operates and manages content and by what means.

One question, however, arises and it is the lack of studies that aim to understand the ideological settings that underpin Facebook and that would explain how platforms shape their hate speech regulations. According to Miller and Rose, the study of governance, not only relates to the study of techniques or mechanisms of intervention, but also to the study of the knowledge that underpins these interventions (2008). Indeed, whereas platform governance has provided a detailed map of how platforms manage and intervene with content, the field lacks a study of how those techniques are shaped by a given knowledge and by concepts that Facebook might be creating that influence their governing practices.

More in particular, and in line with chapter two and three, the field lacks systematic study that looks at how social media platforms understand, create and interpret value equality or freedom and how it affects the governance structure. As such, the present dissertation proposes an empirical research that aims to uncover social media principles and values around hate speech governance and, in particular, how platforms understand equality and freedom. With this aim, the present research theoretically frames this study by looking at Foucauldian studies by Miller and Rose (2008), Terranova Studies on Social Networks, and the lecture at the college of France by Foucault on Security, Territory and Population (1978)

Chapter 5. Researching Facebook. Methodology and Methods

> 'Discourses can be treated as practices that systematically form the objects of which they speak'
>
> (Foucault, 2010, p. 35).

5.1 Introduction

This chapter discusses the research methodology used to explore how Facebook governs hate speech.

According to Mills (2010), methodology, both reveals the researcher's sociological imagination and makes the field of research readable for others. The elements that compose the field are not always clear and objective, but it is the work of the researcher to extract them, organize them, and give them a complete meaning. In considering this, the present chapter seeks to explain both my 'sociological imagination' and the tools I have used to decipher the present field of research.. To this end, the chapter is organized as follows:

The first section of this chapter justifies why I have chosen Facebook as a topic of research and why this choice is relevant. The second section explores the ontological and epistemological position for this research, which I locate between constructivism and the critical paradigm. The third section outlines the research design .The fourth and fifth sections explain in detail how I collected and analysed data, respectively.

5.2 Field of Research

As Gillespie points out, it matters how Facebook sets and enforces the rules (2018). Whereas, in 2013, Facebook had 835 million users, as of the 31st of March 2021, Facebook has 2.85 billion monthly active users and experienced an increase of 10% year-over-year

(Facebook Investor, 2021). The amount of users makes Facebook the first social network to have billions of users. Having such a large number of users implies that the Facebook platform can influence and affect an immense number of individuals worldwide with its hate speech regulation. To reinforce this idea, I would also like to point out how Facebook informs its users about its policies. If we compare Facebook Hate speech community standards with, for instance, the 15th Recommendation by the European Union to combat racism and other forms of discrimination, it is evident that Facebook's capacity to explain what (for them) is hate speech is more accessible for users that the European legal document. In addition, Facebook's transparency policy has to be taken into account, which provides for those users who want to further understand the rationale behind all decision making around hate speech policies. In sum, I have chosen Facebook due to its number of users and its capacity to communicate and influence users with their definition of hate speech.

5.3 Epistemological aspects and research paradigm

The aim of defining my research paradigm is to allow the reader to understand how I perceive and understand knowledge.

According to Guba and Lincoln (1994), a paradigm is a set of fundamental beliefs that represent the inquirer's worldview, and which consist of a series of the interconnected components: ontology, epistemology, and methodology. Ontology aims to answer the question: 'what is the form and nature of reality and what is there that can be known?' Epistemology aims to understand how to approach the question of knowledge and how knowledge can be understood. Finally, methodology refers to how the researcher finds what there is to be known. Every paradigm is based upon its own ontological and epistemological assumptions and, as a researcher, I situate myself at the critical paradigms, whereas it is the latter what leads this research. A constructivist approach implies that we need to study the conditions under which reality is produced, while the critical paradigm reminds us that what makes ideas 'real' is the system of knowledge, the formation of culture, and the relations of power in which these concepts are located (Tuhiwai, 2012 p50)In considering this, the importance of constructivism is that it is a way of thinking, a style of thought, that redirects our attentions towards the set of relations between people and objects that are not necessarily visible to the eye, whereas critical paradigms, in particular critical analysis, delves into the question of power attached to the historical context. Under the critical paradigm, the duty of the researcher is to identify with precision the small and disperse events that brought a concept into existence. As a result, the subject of the study is naked to the eye and the researcher is capable of altering the subject of study and provide a different reading or interpretation.

The critical discourse analyst needs to be clear about the fact that a critique is always situated within a discourse (Jagger and Maier, 2009, 36). As such, it is important that all research is premised upon a variety of assumptions and positions that researchers should make transparent in their work (Mir and Watson, 2000). This is referred to as positionality of the researcher, and is concerned, more specifically, with the ability to understand where the researcher is located in relation to the research, and how it might condition the research approach. Literature on this topic is limited and mostly framed in ethnographic research. However, I will try to show how my personal views may have impacted the present research. My academic and professional trajectory is the area of social justice. As such, in the early stages of this research it came to my attention that the dominant definitions of hate speech do not contemplate categories related to class or minorities. This made me turn my attention to the notion of equality, specifically how it was defined by the dominant body of literature on hate speech that emerged in the 1980s (see Chapter Two). My conclusion was that the notion of equality held in this literature does not consider how the left has traditionally understood equality and, in fact, how it neglects the contribution of the left. Specifically, Hobsbawm argues that the political project for equality of the left has always been universalist (1996, 7). However, the starting point of the left is that it acknowledges that, historically speaking, not all individuals have access to the same assets and equality needs to be facilitated. This has provided a point of reference for this research, which looks to make the case that there are more perspectives on hate speech regulation than what is proposed by the dominant literature. My own position, therefore, is a critical one, coming from a social justice perspective, and seeking to interrogate dominant/hegemonic patterns of knowledge surrounding hate speech.

5.5 Research questions

According to Bryman (2012), a research design is a framework for the generation of evidence that is suited both to a certain set of criteria and to the research questions in which the researcher is interested. Chapter One outlined the formulation of my research questions, which were informed by the genealogical research undertaken in Chapters Two, Three and Four. They are the following:

How does Facebook regulate and govern hate speech? What are Facebook's principles and values around hate speech definition? What techniques has Facebook developed to enforce its policies? What values underpin Facebook’s operational system?

5.6 Research Design.

This dissertation covered a period of time in which Facebook created, developed and established its apparatus of governance for hateful content.

Facebook governing systems have greatly evolved during the time of writing this dissertation, to an extent that, during the first three years of this dissertation, the platform appeared to be improvising its governing mechanisms, creating them on a regular basis. As a result, the Facebook governance apparatus has been gradually created by an infinitesimal series of techniques. A setting button, or an informal post by Facebook’s CEO Mark

Zuckerberg, may eventually play out a role in the governance of Facebook. Even the smallest or peripheral Facebook documents do not stand alone, but work in relation to another, or enclose information that filters and influences different techniques.

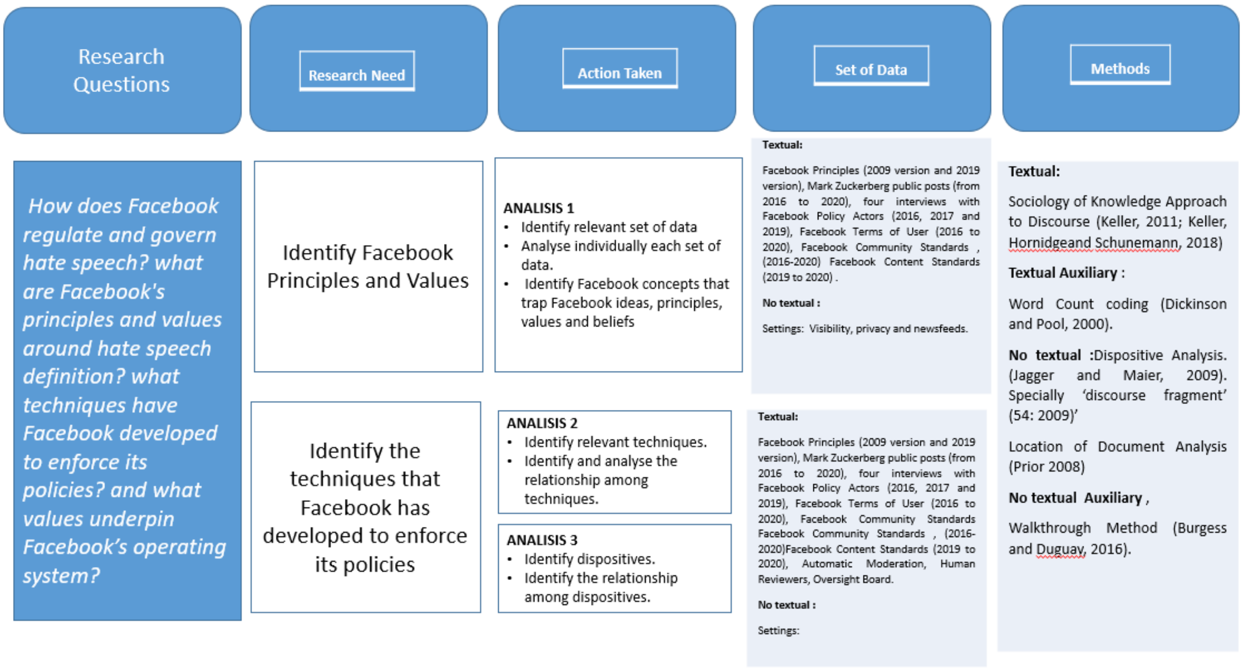
In addition, Facebook strategies are not permanent, but constantly evolving. This has demanded a time-linear observation, documentation and, finally, a chronological analysis of all Facebook policies and practices that relate to the governance of hateful content. Tracking the various documents across time reveals that some of the practices and dynamics have been solidifying in the period that the present study took place, whereas others have been discarded by the company. This was particularly clear when analysing the 'counter narrative programmes' and the Rights and Responsibilities documents. If, in 2016, those two practices were central to the analysis and revealed important information, they soon became redundant as techniques of governance.4 Others, like Community Standards, content standards, human moderation or automatic detection, evolved and became central to Facebook governance (Gillespie, 2018). As such, while I study all techniques that emerged throughout the period of this study, in the findings chapters I focus on the techniques that were sustained and solidified by Facebook.5 Over time, I also observed that certain dynamics among techniques became established, for exampe the direct relationship between policy makers, sector managers and human reviewer's activity. This observation reveals that Facebook does not alter its practices; rather, it solidifies and selects those that work 'better'. This aspect became particularly clear after I conducted the fourth interview with a Facebook Public Policy informant, as explained below.

[4] Their analysis, however, held important relevant information that this thesis took into consideration.

[5] Also took into accountability the information provided by techniques that are not in use, but that at a point revealed important, background information useful for the analysis, i.e. the relationship between Rights of Responsibilities and Facebook Principles, which this dissertation explains in detail in the analysis section of the present chapter.

I initially undertook interviews with informants, asking for information that explained what Facebook was doing, what solutions they were putting in place, how their policies were operationalised, how they understood hate speech, what they were learning, what innovations they had created to tackle hate speech, or what was the latest product solution. When, in early 2019, the conversations shifted to what was the specific role of the Public Policies, I began to understand that roles and dynamics involved in the governance of content were more established than the previous interview (2016, 2017), and that there were also better established channels of influence between different sections or departments within Facebook. For example, Zuckerberg's Letters (as they were informally referred to by the Facebook policy makers interviewed) were reference points for policy activity. Policies refer to principles and values, and human reviewer's dependence on public policies, etc.

In sum, this dissertation, the data selected and the methods chosen to analyse data represent a period of time in which Facebook began to get involved in the governance of hate, the creation of diverse techniques, the establishment of some of those techniques, and the growing relationship between techniques. The following sections are a schematic explanation (as opposed to a chronological one) of how I selected, collected and analysed data in a constantly evolving environment.



Research Questions

How does Facebook regulate and govern hate speech? what are Facebook's principles and values around hate speech definition? what techniques have Facebook developed to enforce its policies? and what values underpin Facebook's operating system?

Research Need

Identify Facebook Principles and Values

Identify the techniques that Facebook has developed to enforce its policies

Action Taken

ANALISIS 1

- Identify relevant set of data
- Analyse individually each set of data.
- Identify Facebook concepts that trap Facebook ideas, principles, values and beliefs

ANALISIS 2

- Identify relevant techniques.
- Identify and analyse the relationship among techniques.

ANALISIS 3

- Identify dispositives.
- Identify the relationship among dispositives.

Set of Data

Textual:

Facebook Principles (2009 version and 2019 version), Mark Zuckerberg public posts (from 2016 to 2020), four interviews with Facebook Policy Actors (2016, 2017 and 2019), Facebook Terms of User (2016 to 2020), Facebook Community Standards , (2016-2020) Facebook Content Standards (2019 to 2020) .

No textual :

Settings: Visibility, privacy and newsfeeds.

Textual:

Facebook Principles (2009 version and 2019 version), Mark Zuckerberg public posts (from 2016 to 2020), four interviews with Facebook Policy Actors (2016, 2017 and 2019), Facebook Terms of User (2016 to 2020), Facebook Community Standards Facebook Community Standards , (2016-2020)Facebook Content Standards (2019 to 2020), Automatic Moderation, Human Reviewers, Oversight Board.

No textual :

Settings:

Methods

Textual:

Sociology of Knowledge Approach to Discourse (Keller, 2011; Keller, Hornidgeand Schunemann, 2018)

Textual Auxiliary :

Word Count coding (Dickinson and Pool, 2000).

No textual :Dispositive Analysis. (Jagger and Maier, 2009). Specially 'discourse fragment' (54: 2009)'

Location of Document Analysis (Prior 2008)

No textual Auxiliary ,

Walkthrough Method (Burgess and Duguay, 2016).

Figure 4 illustrates research design

5.7 **Data selection**

The empirical research consists of a qualitative analysis of a diverse and large set of data, namely: Facebook Principles (2009 version and 2019 version), Mark Zuckerberg's public posts (from 2016 to 2020), four interviews with Facebook Policy Actors (2015, 2016, 2017 and 2019), Facebook Terms of Serive (2016 to 2020), Facebook Community Standards and Facebook Content Standards (2019 to 2020), a range of non-textual settings, i.e. visibility settings and newsfeed settings, reporting mechanisms that include buttons, drop-down menus or clickable links. Additionally, the specific analysis of Public Policy interviews, Mark Zuckerberg posts, Community Standards, About FB documents and literature by Roberts (2016, 2019) and Gillespie (2019) reveal the existence of the Oversight Board, automatic detection and human reviewer teams, which I included in the list of techniques and, therefore, to the analysis of the series of Facebook techniques.In addition, I have used Facebook Transparency Reports such as About FB documents, which inform on Facebook new initiatives, and Facebook investor, a Facebook-run site dedicated to informing on Facebook financial benefits.

5.7.1 Data Collection

5.7.1.1 Monitoring Facebook site

I have monitored the Facebook platform on a daily basis from October 2016 to October 2020. Monitoring involved using the website on a daily basis, exploring all the tools and settings that Facebook was making available, and manually annotating in a diary all the changes that were introduced on the site. In addition, I often took note of the links between sections, understanding them as bridges that link sections or information and their possible relation. By monitoring the Facebook site, I became very familiar with my field of study, specifically the continuing changes of its policies and enforcement mechanisms. During the writing period, and because it was easier to access the main documents, I used the 'Platform Governance Archive' developed by Magalhanes and Katzenbach (2020) which has an archive of all Facebook documents from previous years.

5.7.1.2 Principles of Facebook

I have monitored, extracted, compiled and analysed the Principles of Facebook. Facebook has published two versions, one in 2009 and a second in 2019. The Principles of Facebook are colloquially known inside the platform as the DNA of the company. The different versions of Facebook Principles have been extracted manually.

5.7.1.3 Monitoring and extracting Mark Zuckerberg's publications

All Mark Zuckerberg's written posts and publications, from May 2016 to November 2020, were manually extracted. The total number of posts extracted was 55, although the total of posts analysed was 223 as they were more directly concerned with hate speech. I copied, pasted and printed all the posts and transcripts in chronological order, from the 11 May 2016 to December 2020. I used the Zuckerberg Files https://zuckerbergfiles.org/, a database which allows researchers to redirect to the original post. The Zuckerberg Files is housed in Digital Commons at the University of Wisconsin-Milwaukee and consists of two digital collections. The "Transcripts" collection includes full-text transcripts of all content in the digital archive of Zuckerberg's public statements, while the "videos" collection represents a subset of the group with archived and available video files that document Zuckerberg's appearances. The virtue of this database is that it reconnects the researcher to the source – that is, to the primary data source – eliminating the burden of scrolling back. I have also used secondary data from different sources such as the transcript from a leaked discussion between Mark Zuckerberg and Facebook employees, dated 1 October 2019, as reported by Casey Newton.

5.7.1.4 Facebook policy officers

I conducted a total of four in-depth interviews with Facebook. I conducted the first interview with Patricia Cartes in November 2015, thanks to personal connections. When I conducted the interview, Cartes was working for Twitter, but Cartes had previous and precious professional experience on Facebook when Facebook had only two people in Europe. Interestingly, Cartes was part of the team that had initially decided how governing hate speech should function and could look like; so, in that regard, the interview conducted with Patricia Cartes provided important insights into Facebook. She offered information around the ideas behind the governance design of Facebook at the time, and around the decisions that led to selecting one particular design to govern hate speech over another. On 1 February 2016, I did a formal in-depth interview with Facebook Policy Risk Manager, Content Policy, Siobhan Cummiskey. The interview took place at Facebook headquarters in Dublin. During these interviews, Cummiskey informed me of the different techniques, mechanisms and systems that Facebook was putting in place at the time. The interview helped me understand and make sense of the data I was extracting.

This interview with Cummiskey was followed a year later by an in-depth interview with Aibhinn Kelleher on 8 June 2017. I gained my interview with Kelleher after I was invited to Facebook Headquarters in Dublin to participate in a training day about counter-narratives. By then, Aibhinn Kelleher was responsible for policymaking at Facebook and was in charge of counter-narrative initiatives, a topic I was researching in-depth during 2016 and which is included in the analysis of this dissertation. The interview was held at Dublin Headquarters, and it revolved around Facebook values and principles, including the topic of equality and the value of Mark Zuckerberg's posts. Finally, a fourth interview was conducted on Facebook, again with Siobhan Cummiskey. It took place on 21 January 2019, at Facebook Headquarters. By 2019, Siobhan Cummiskey was Director of Public Policy, Campaigns and Programs, in Europe, Middle East and Africa. The conversation revolved around the mechanisms, values and developments of Facebook governance of hate speech in the period between 2016 and 2019. The interviews were recorded on an audio recording device and transcribed verbatim, as it helped me become familiar with the data (Braun and Clark, 2017). All the interviewees signed a consent form and allowed me to use their names.

5.7.1.5 Terms of Service and policies:

I have monitored, extracted and compiled the Terms of Service and its changes. Between 2016 to 2019, the Terms of Service established the list of user rights and obligations. The Terms of Service has registered five changes in this period: 18/05/2017, 26/05/18, 3/08/2019, 17/10/2020 and 24/10/2020. The changes within documents were monitored manually, but for the final analysis I used Magalhanes and Katzenbach's project (2020) https://pga.hiig.de/.

5.7.1.6 Community Standards and Content:

I have monitored, extracted, compiled, and analysed the Community Standards. Community Standards is the document that regulates users' behaviour on the platform. I have monitored and recorded its changes from 2016 until 2020. The total number of versions during this period is 26, particularly when Facebook added the Content Standards. Specifically, the analysis initially focused on four documents: version 1 of 30/04/2015, version 2 of 28/05/2018, version 3 of 7/07/2019, and version 4 of 11/08/2020, as they are representative of all the changes introduced. The 2015 version defines what is hate speech, whereas the 2018, 2019 and 2020 versions define hate speech and list specific examples of what is allowed and what is not allowed. The more detailed analysis relied on version 4 of 11/08/2020.

Community standards have, since 2018, a section dedicated to content standards (under a hyperlink that reads READ MORE). In 2018, it included a section called 'Recent Updates', in which Facebook provides the most recent changes to content standards. These changes are also regularly announced by Mark Zuckerberg. Content Standards is an always-changing document, content-wise. No extraction was needed, as the section 'Recent updates' specified the changes. I noted the changes in a document, and the analysis that is reflected in Chapters Six and Seven relies on the 12/09/2020 version of 'recent updates', when holocaust denial was introduced (see Chapter Seven).

5.7.1.8 Oversight Board:

I have monitored, extracted, compiled and analysed all the references to the Oversight Board published on the Facebook website. I have followed and monitored the Oversight Board since its inception in November 2018 until its implementation in September 2019, and its first step from September 2019 to October 2020. The interview conducted with Facebook key informant Cummiskey in 2019, the page About FB, and then a dedicated page to the Oversight Board were the main sources of information for this analysis.

5.7.1.9 Settings:

I collected all objects and settings involved in the governance of contents, focusing on those relevant to hate speech, namely general account settings, privacy settings, specific settings to regulate visibility regarding other publications, newsfeed settings, and settings to activate the flagging report system. I collected these on a regular basis, capturing their location on the platform using the Snipping Tool, which allows users to easily screenshot images.

5.7.1.10Automatic Detection

For the analysis of Automatic Detection, I rely on interviews I conducted, Facebook Transparency Reports (https://transparency.facebook.com/), and relevant academic articles.

5.7.1.11 Human Reviewers

In order to gather knowledge about moderation, I have relied on interviews I conducted, the work developed by Roberts (2016, 2018) and Gillespie (2018), and secondary data, specifically leaked information about moderation published by the Guardian newspaper: https://www.theguardian.com/news/series/facebook-files.

5.8 Data analysis.

The following section and subsections explain in detail how I have analysed the selected data.

The analysis performed has been threefold. I first looked at the set of data separately as standalone units of information (Analysis 1 Analysing the data as carriers of information). Next, I looked at the set of data in their role of governance techniques (Analysis 2, Analysing the data as techniques and in terms of their location). Third, I analysed the relationship between techniques (Analysis 3, Analysing the relationship between techniques). The following paragraphs are a re-creation of the analysis I conducted.

5.8.1 Analysis 1. Analysing the data as carriers of information.

For this analysis, the set of data is analysed as 'standalone' units of information. I use this analysis to identify what principles and values Facebook has produced to govern hate speech.

All data is discursive in nature, but some of the data analysed is textual, while some is non-textual. As such, I applied two different methods to identify the set of ideas and how they manifest in the different forms of data, namely the Sociology of Knowledge Approach to Discourse, or SKAD (Keller, 2011; Keller Hornidgeand Schunemann, 2018) for textual analysis, and dispositive analysis (Jagger and Maier, 2009) for non-textual analysis.

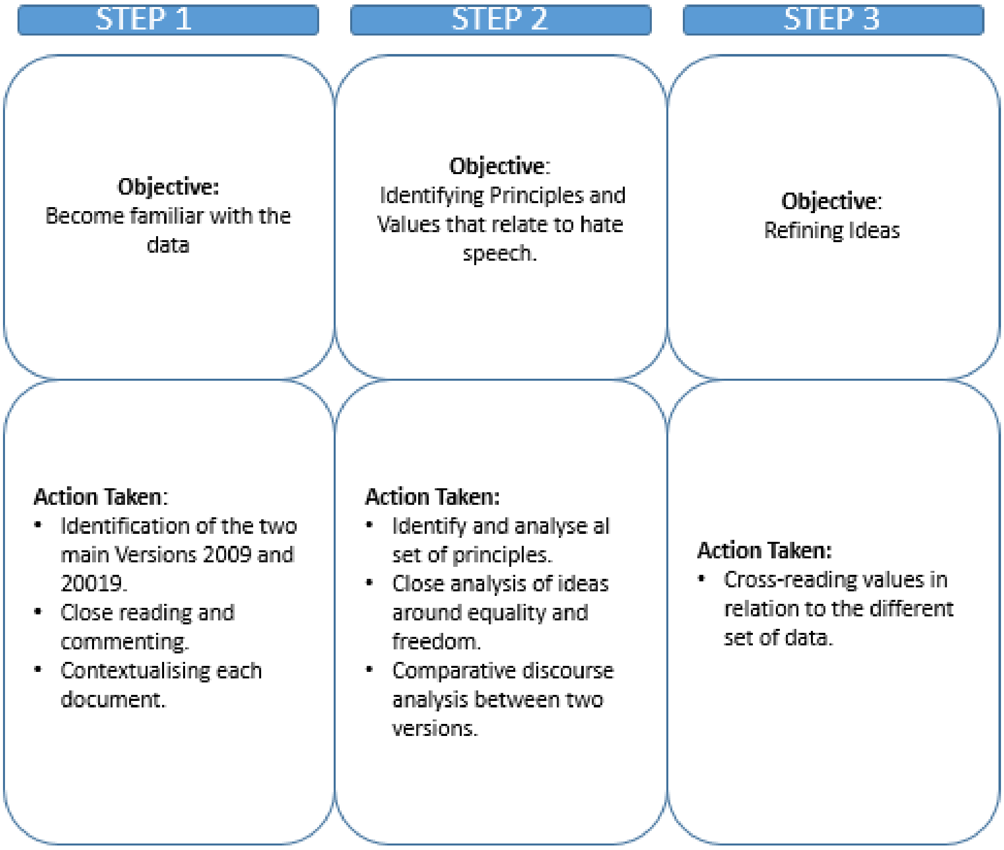
SKAD is concerned with the social production, circulation and transformation of knowledge. SKAD draws from Berger and Luckmann, and Foucault's, writings in Archaeology of Knowledge (2010). Berger and Luckmann coined the term 'conversation machinery' in relation to the production of knowledge that emerges between the interaction of subjects and, in particular, its institutionalisation (Keller, 2011, p.44). Foucault reflects on the meaning of knowledge, and the particularity of its creation, which he views as a historically situated practice, very much embedded in everyday life and its interactions (2010). Similarly, Berger and Luckmann aim to locate and identify the creation of knowledge typified through interactions and institutionalised as valid knowledge. It is important, therefore, to locate notions, words and concepts that Facebook is creating, as it is important to understand how Facebook uses them, how they permeate into daily Facebook narratives and how they conform discourses strands. This means looking at concrete information such as the idea of 'Platform for all' that Facebook has created, and the meaning and uses that Facebook grants to this. In addition, it is also part of the analysis to understand what system of knowledge this information contributes to and from which it is shaped. In sum, with SKDA analysis I aim to understand how Facebook has constructed its own body of knowledge, how it has institutionalised its own ideas about hate speech policies and content management, and how it is making them, ultimately, transmittable to the general public.

I complement SKDA analysis with word count coding (Dickinson and Pool, 2000) in order to look at frequency of use and the amount of times keywords appeared in Mark Zuckerberg's public post through time.

For non-textual analysis, I relied on a specific aspect of the Dispositive Analysis (Jagger and Maier, 2009), which draws on a Foucauldian understanding of discourse and explores the discursive relation of objects. The discursive relation of objects is treated by Foucault in Archaeology of Knowledge (2010) and is also tackled by Foucault in Security Territory and Population (1978, 2007) – although, in this case, Foucault refers to it in a rather scattered manner. Jagger and Maier gathered both texts and reconstructed a body of knowledge around the notion of objects, detecting and solving a blind spot in Foucault's theory, which is the mediation between the textual and non-textual practices (Wodak and Meyer, 2009). Specifically, I made use of what the authors refer to as 'discourse fragment' (2009, p.54), which I used for the analysis of buttons, displayable windows, and all those elements that are part of Facebook's governance of hateful content. This specific interpretation of discourse fragment has to be understood and read alongside Langlois' body of work (2013). The authors propose that buttons and interfaces are relevant subjects for the study of governance as their design reveals something about the ideas and values of those who designed these tools. As such, the analysis has consisted of understanding the significance of the location of buttons, their disposition on the site, and an analysis of what knowledge they correspond to. The discourses analysis of the non-textual symbols or discursive fragments has limitations for the present analysis, however, since Jagger and Maier have developed this based on material objects rather than digital objects. As such, I have complemented their work by using what is known as the ‘walkthrough method’ (Burgess and Duguay, 2018). Burgess and Duguay argue that the walkthrough method contributes to systematically and forensically stepping through the various stages of an application, its buttons, windows and stages that the user has to go through to perform an action on a website or an app. For the authors, ‘the walkthrough method establishes a foundational corpus of data upon which can be built a more detailed analysis of an app’s intended purpose, embedded cultural meanings and implied ideal users and uses’ (2018, p.887).

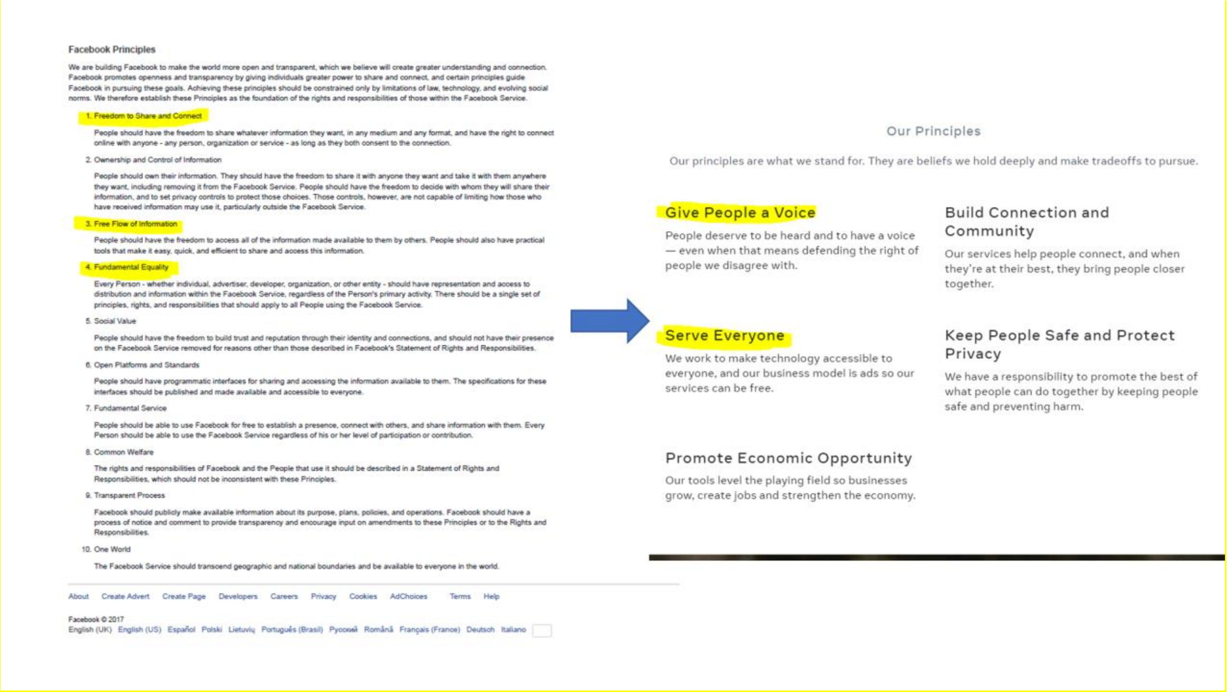
The following paragraphs and figures explain in detail how I analysed each set of data. The order of figures has been selected to follow the internal logic of the analysis – although, as has been pointed out above, the dynamic nature of Facebook meant that, at some point in time, I would need to look back and alter the order of analysis to better understand the relationships. This, however, does not alter the findings and the following figures are intended to guide the reader through the interactive steps of the process of the analysis.

5.8.1.1 Facebook Principles



STEP 1	STEP 2	STEP 3
Objective: Become familiar with the data	**Objective**: Identifying Principles and Values that relate to hate speech.	**Objective**: Refining Ideas
Action Taken: • Identification of the two main Versions 2009 and 20019. • Close reading and commenting. • Contextualising each document.	**Action Taken:** • Identify and analyse al set of principles. • Close analysis of ideas around equality and freedom. • Comparative discourse analysis between two versions.	**Action Taken:** • Cross-reading values in relation to the different set of data.

Figure 5 Illustrates phases of Facebook Principles research and methods applied



Facebook Principles

We are building Facebook to make the world more open and transparent, which we believe will create greater understanding and connection. Facebook promotes openness and transparency by giving individuals greater power to share and connect, and certain principles guide Facebook in pursuing these goals. Achieving these principles should be constrained only by limitations of law, technology, and evolving social norms. We therefore establish these Principles as the foundation of the rights and responsibilities of those within the Facebook Service.

1. Freedom to Share and Connect

   People should have the freedom to share whatever information they want, in any medium and any format, and have the right to connect online with anyone - any person, organization or service - as long as they both consent to the connection.

2. Ownership and Control of Information

   People should own their information. They should have the freedom to share it with anyone they want and take it with them anywhere they want, including removing it from the Facebook Service. People should have the freedom to decide with whom they will share their information, and to set privacy controls to protect those choices. Those controls, however, are not capable of limiting how those who have received information may use it, particularly outside the Facebook Service.

3. Free Flow of Information

   People should have the freedom to access all of the information made available to them by others. People should also have practical tools that make it easy, quick, and efficient to share and access this information.

4. Fundamental Equality

   Every Person - whether individual, advertiser, developer, organization, or other entity - should have representation and access to distribution and information within the Facebook Service, regardless of the Person's primary activity. There should be a single set of principles, rights, and responsibilities that should apply to all People using the Facebook Service.

5. Social Value

   People should have the freedom to build trust and reputation through their identity and connections, and should not have their presence on the Facebook Service removed for reasons other than those described in Facebook's Statement of Rights and Responsibilities.

6. Open Platforms and Standards

   People should have programmatic interfaces for sharing and accessing the information available to them. The specifications for these interfaces should be published and made available and accessible to everyone.

7. Fundamental Service

   People should be able to use Facebook for free to establish a presence, connect with others, and share information with them. Every Person should be able to use the Facebook Service regardless of his or her level of participation or contribution.

8. Common Welfare

   The rights and responsibilities of Facebook and the People that use it should be described in a Statement of Rights and Responsibilities, which should not be inconsistent with these Principles.

9. Transparent Process

   Facebook should publicly make available information about its purpose, plans, policies, and operations. Facebook should have a process of notice and comment to provide transparency and encourage input on amendments to these Principles or to the Rights and Responsibilities.

10. One World

   The Facebook Service should transcend geographic and national boundaries and be available to everyone in the world.

About Create Advert Create Page Developers Careers Privacy Cookies AdChoices Terms Help

Facebook © 2017
English (UK) English (US) Español Polski Lietuvių Português (Brasil) Русский Română Français (France) Deutsch Italiano

Our Principles

Our principles are what we stand for. They are beliefs we hold deeply and make tradeoffs to pursue.

Give People a Voice

People deserve to be heard and to have a voice — even when that means defending the right of people we disagree with.

Build Connection and Community

Our services help people connect, and when they're at their best, they bring people closer together.

Serve Everyone

We work to make technology accessible to everyone, and our business model is ads so our services can be free.

Keep People Safe and Protect Privacy

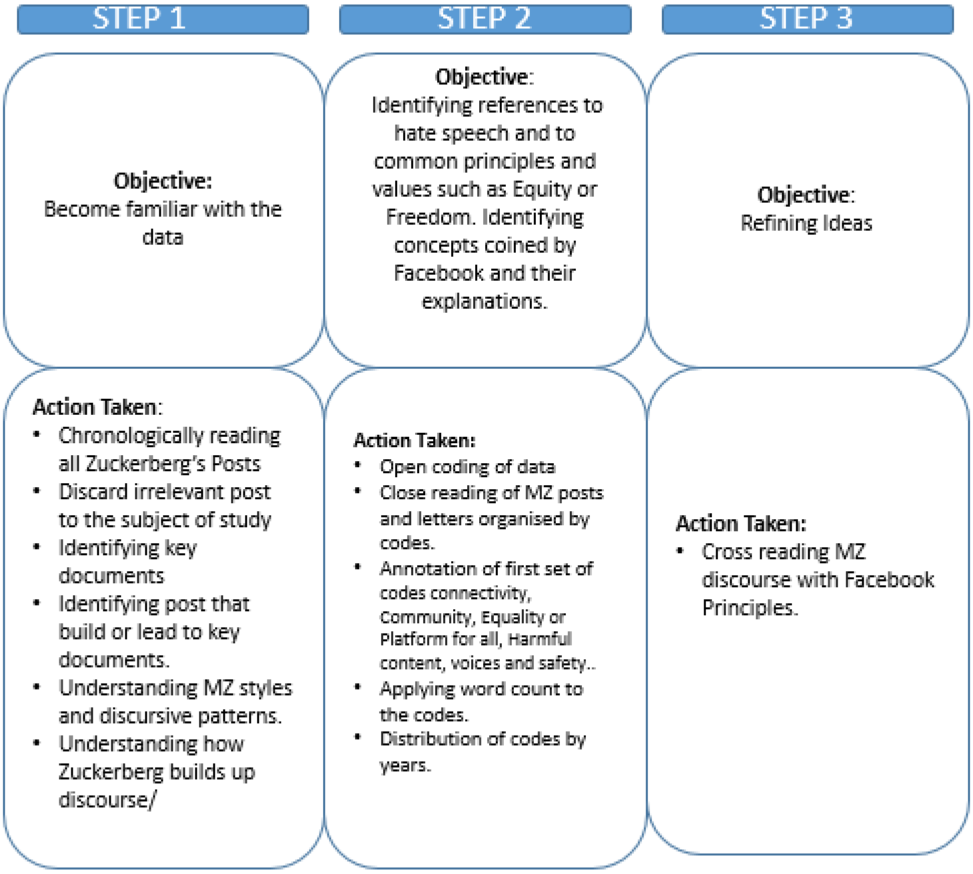
We have a responsibility to promote the best of what people can do together by keeping people safe and preventing harm.

Promote Economic Opportunity

Our tools level the playing field so businesses grow, create jobs and strengthen the economy.

Figure 6 2009-2019 Facebook Principles to 2019 -2021 Facebook Principles

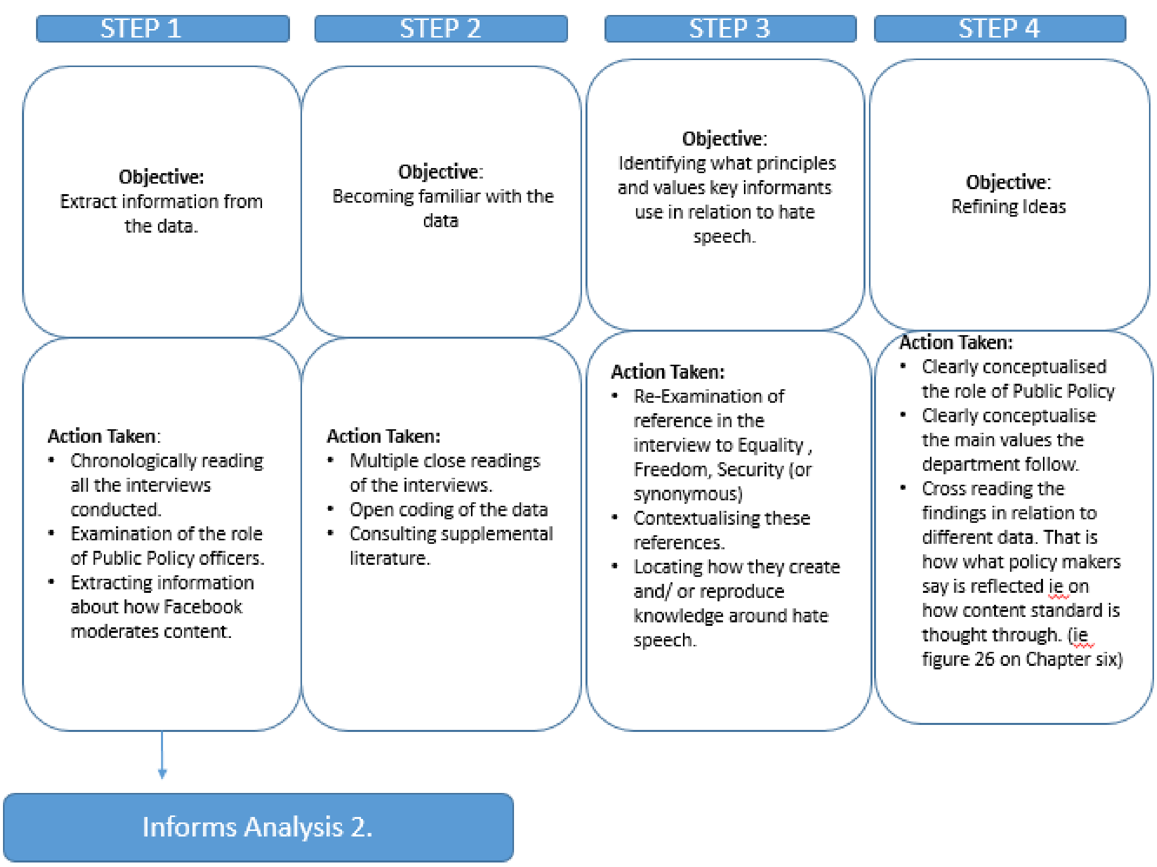
5.8.1.2. Mark Zuckerberg's publications



STEP 1	STEP 2	STEP 3
Objective: Become familiar with the data	**Objective**: Identifying references to hate speech and to common principles and values such as Equity or Freedom. Identifying concepts coined by Facebook and their explanations.	**Objective**: Refining Ideas
Action Taken: • Chronologically reading all Zuckerberg's Posts • Discard irrelevant post to the subject of study • Identifying key documents • Identifying post that build or lead to key documents. • Understanding MZ styles and discursive patterns. • Understanding how Zuckerberg builds up discourse/	**Action Taken:** • Open coding of data • Close reading of MZ posts and letters organised by codes. • Annotation of first set of codes connectivity, Community, Equality or Platform for all, Harmful content, voices and safety.. • Applying word count to the codes. • Distribution of codes by years.	**Action Taken:** • Cross reading MZ discourse with Facebook Principles.

Figure 7 illustrates how I analysed MZ public post

5.8.1.3. Facebook Policy Officer's Interviews.



STEP 1

Objective:
Extract information from the data.

Action Taken:

- Chronologically reading all the interviews conducted.
- Examination of the role of Public Policy officers.
- Extracting information about how Facebook moderates content.

STEP 2

Objective:
Becoming familiar with the data

Action Taken:

- Multiple close readings of the interviews.
- Open coding of the data
- Consulting supplemental literature.

STEP 3

Objective:
Identifying what principles and values key informants use in relation to hate speech.

Action Taken:

- Re-Examination of reference in the interview to Equality , Freedom, Security (or synonymous)
- Contextualising these references.
- Locating how they create and/ or reproduce knowledge around hate speech.

STEP 4

Objective:
Refining Ideas

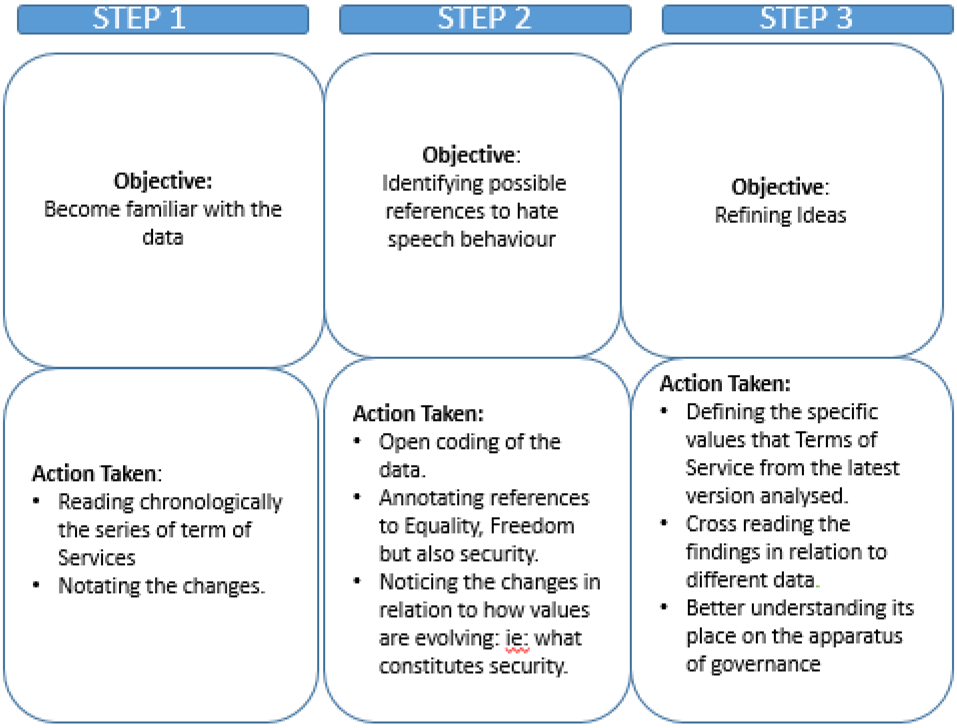
Action Taken:

- Clearly conceptualised the role of Public Policy
- Clearly conceptualise the main values the department follow.
- Cross reading the findings in relation to different data. That is how what policy makers say is reflected ie on how content standard is thought through. (ie figure 26 on Chapter six)

Informs Analysis 2.

Figure 8 illustrates how I analysed Public Policy Interviews

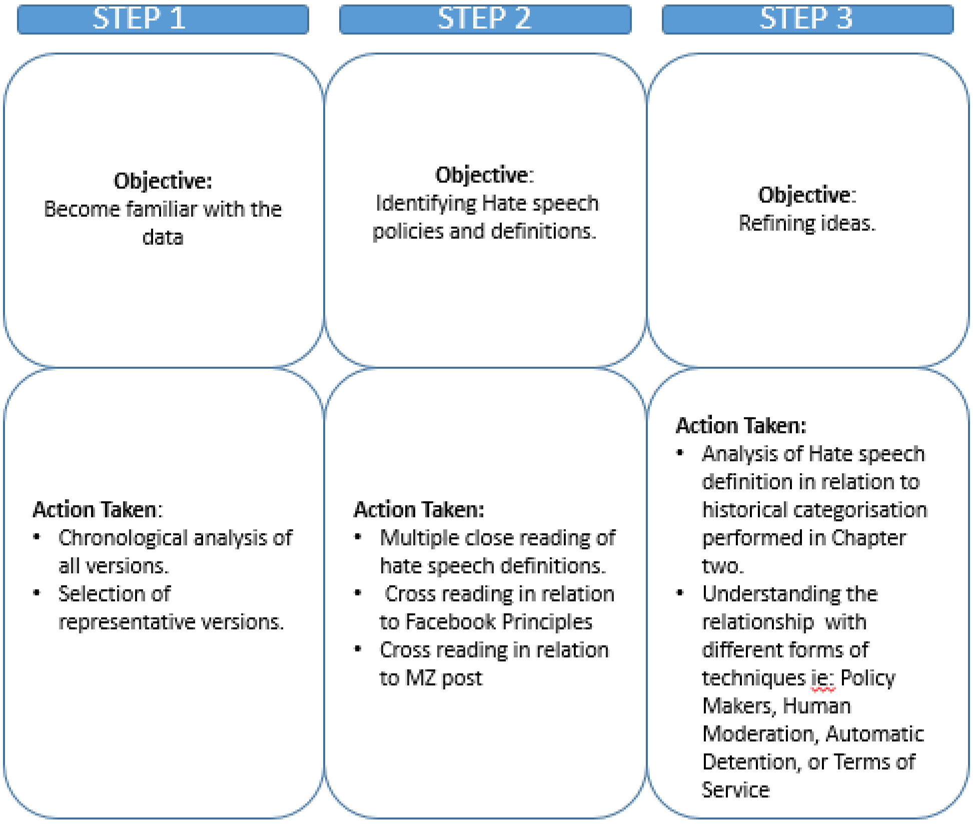
5.8.1.4. Terms of Service



STEP 1	STEP 2	STEP 3
Objective: Become familiar with the data	**Objective**: Identifying possible references to hate speech behaviour	**Objective**: Refining Ideas
Action Taken: • Reading chronologically the series of term of Services • Notating the changes.	**Action Taken:** • Open coding of the data. • Annotating references to Equality, Freedom but also security. • Noticing the changes in relation to how values are evolving: ie: what constitutes security.	**Action Taken:** • Defining the specific values that Terms of Service from the latest version analysed. • Cross reading the findings in relation to different data. • Better understanding its place on the apparatus of governance

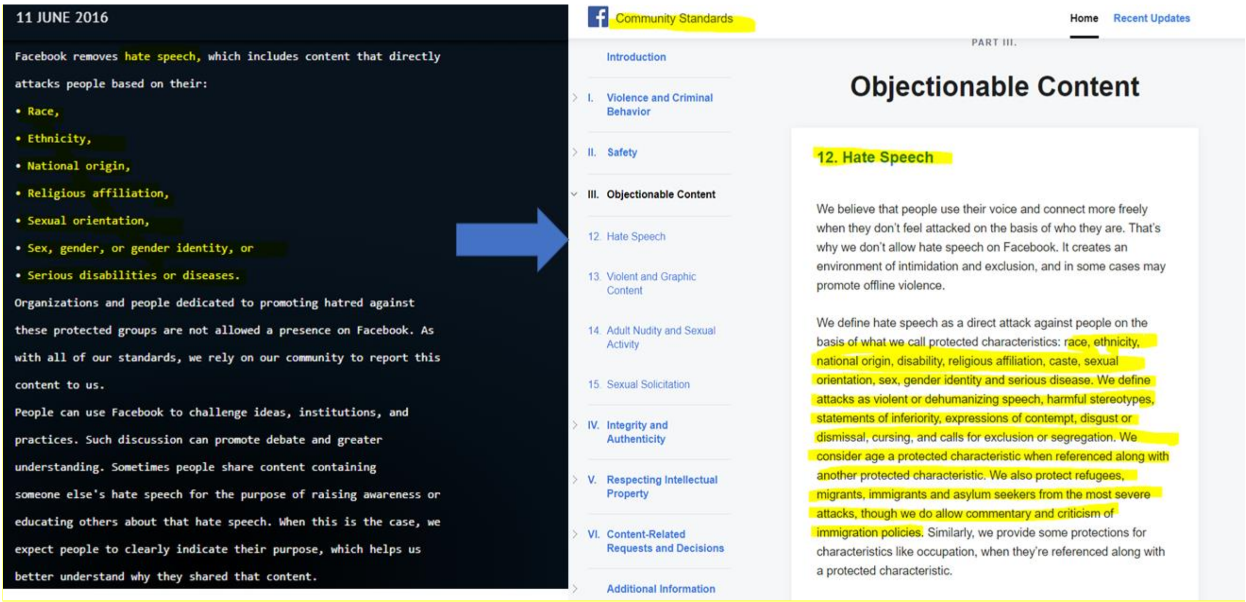
Figure 9 illustrates how I analysed Terms of service

5.8.1.5. Community and Contend Standards



STEP 1	STEP 2	STEP 3
Objective: Become familiar with the data	**Objective**: Identifying Hate speech policies and definitions.	**Objective**: Refining ideas.
Action Taken: • Chronological analysis of all versions. • Selection of representative versions.	**Action Taken:** • Multiple close reading of hate speech definitions. • Cross reading in relation to Facebook Principles • Cross reading in relation to MZ post	**Action Taken:** • Analysis of Hate speech definition in relation to historical categorisation performed in Chapter two. • Understanding the relationship with different forms of techniques ie: Policy Makers, Human Moderation, Automatic Detention, or Terms of Service

Figure 10 Illustrates how I analysed community and content standards



11 JUNE 2016

Facebook removes hate speech, which includes content that directly attacks people based on their:

- Race,
- Ethnicity,
- National origin,
- Religious affiliation,
- Sexual orientation,
- Sex, gender, or gender identity, or
- Serious disabilities or diseases.

Organizations and people dedicated to promoting hatred against these protected groups are not allowed a presence on Facebook. As with all of our standards, we rely on our community to report this content to us.

People can use Facebook to challenge ideas, institutions, and practices. Such discussion can promote debate and greater understanding. Sometimes people share content containing someone else's hate speech for the purpose of raising awareness or educating others about that hate speech. When this is the case, we expect people to clearly indicate their purpose, which helps us better understand why they shared that content.

Community Standards
Home
Recent Updates
Introduction
I. Violence and Criminal Behavior
II. Safety
III. Objectionable Content
12. Hate Speech
13. Violent and Graphic Content
14. Adult Nudity and Sexual Activity
15. Sexual Solicitation
IV. Integrity and Authenticity
V. Respecting Intellectual Property
VI. Content-Related Requests and Decisions
Additional Information

PART III.

Objectionable Content

12. Hate Speech

We believe that people use their voice and connect more freely when they don't feel attacked on the basis of who they are. That's why we don't allow hate speech on Facebook. It creates an environment of intimidation and exclusion, and in some cases may promote offline violence.

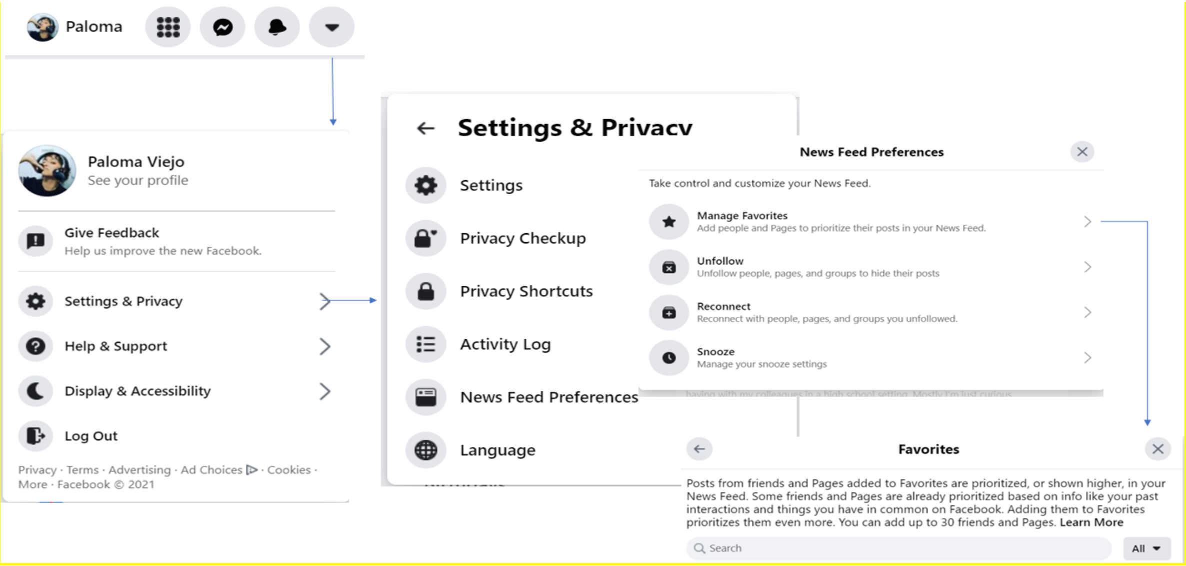
We define hate speech as a direct attack against people on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing, and calls for exclusion or segregation. We consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants and asylum seekers from the most severe attacks, though we do allow commentary and criticism of immigration policies. Similarly, we provide some protections for characteristics like occupation, when they're referenced along with a protected characteristic.

Figure 11 shows an example of how community standards change over time

5.8.1.6. Settings

STEP 1	STEP 2	STEP 3
Objective: Become familiar with the data	**Objective:** identifying how settings reflects principles and values around equality or freedom.	**Objective:** Refining ideas
Action Taken: Using all buttons. Exploring the web	**Action Taken:** Annotating all the steps , arrows, location of the arrows. Making comments and notes in relation to how these bottoms might relate to above named principals and values. (Walkthrough Methods)	**Action Taken:** • Answering the following questions: • What is the layout? • How to these relate to other techniques. How does it reproduce values?

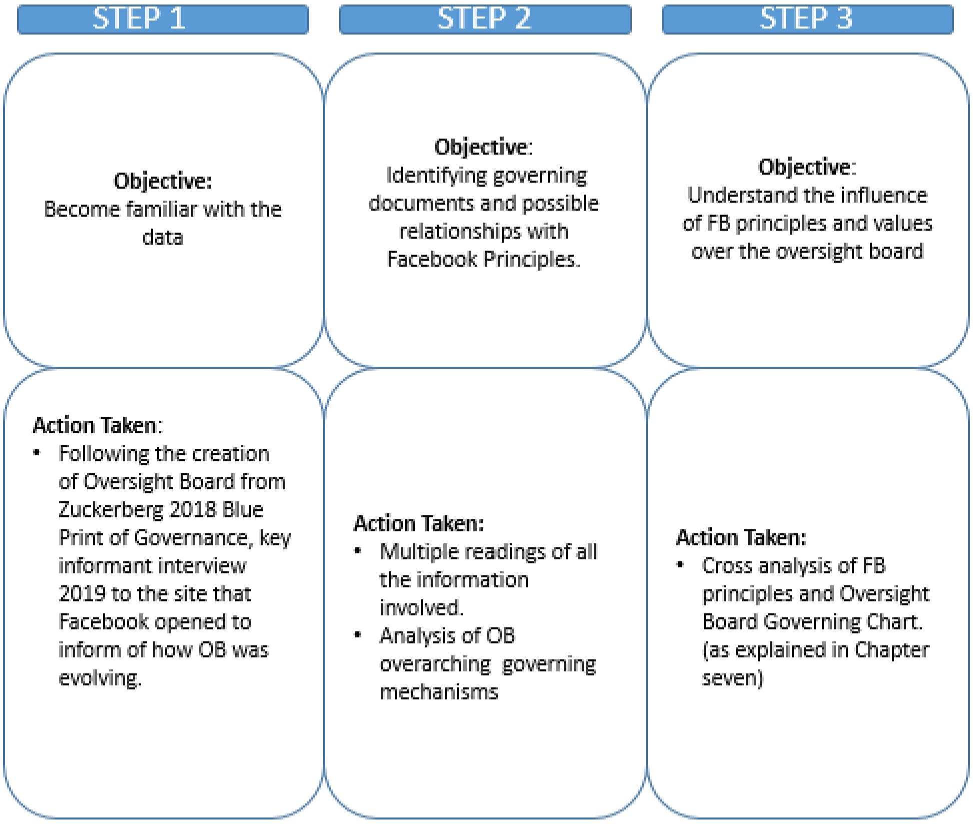
Figure 12 shows how I analysed settings


Paloma
Paloma Viejo
See your profile
Give Feedback
Help us improve the new Facebook.
Settings & Privacy
Help & Support
Display & Accessibility
Log Out
Privacy · Terms · Advertising · Ad Choices ▷ · Cookies · More · Facebook © 2021
Settings & Privacv
Settings
Privacy Checkup
Privacy Shortcuts
Activity Log
News Feed Preferences
Language
News Feed Preferences
Take control and customize your News Feed.
Manage Favorites
Add people and Pages to prioritize their posts in your News Feed.
Unfollow
Unfollow people, pages, and groups to hide their posts
Reconnect
Reconnect with people, pages, and groups you unfollowed.
Snooze
Manage your snooze settings
Favorites
Search
All

Posts from friends and Pages added to Favorites are prioritized, or shown higher, in your News Feed. Some friends and Pages are already prioritized based on info like your past interactions and things you have in common on Facebook. Adding them to Favorites prioritizes them even more. You can add up to 30 friends and Pages. **Learn More**

Figure 13 Shows News feeds and how I used walkthrough method steps

5.8.1.7 Oversight Board



STEP 1	STEP 2	STEP 3
Objective: Become familiar with the data	**Objective**: Identifying governing documents and possible relationships with Facebook Principles.	**Objective**: Understand the influence of FB principles and values over the oversight board
Action Taken: • Following the creation of Oversight Board from Zuckerberg 2018 Blue Print of Governance, key informant interview 2019 to the site that Facebook opened to inform of how OB was evolving.	**Action Taken:** • Multiple readings of all the information involved. • Analysis of OB overarching governing mechanisms	**Action Taken:** • Cross analysis of FB principles and Oversight Board Governing Chart. (as explained in Chapter seven)

Figure 14 shows how I analysed The Oversight Board

5.8.1.8 AI and Human Reviewers

For AI and Human Reviewers analysis I relayed in the work developed by Sarah Roberts on human reviewers (2016, 2019) and Gorwa et al (2020) for Automatic Detention.

In sum, above I have reflected on a series of figures that demonstrate how I analysed the different set of data, having in all cases as reference how Facebook was producing, circulating, and transforming knowledge related to hate speech and the values around hate speech as SKDA or Dispositive analysis informs. The main contribution of this analysis was that it a) helped the analysis to determined how Facebook produces ideas around freedom or equality and the values it gives to these principles, and b) contributed to uncover 'discourse strands' (Jagger and Maier, 2009, p46).

Jagger and Maier refer to 'discourse strands' (2009, 46), that is, the way in which specific and recurring text, symbols and strategies permeate all techniques. Jagger and Maier understand that discourse strands lead the researcher to understand the emergence and solidification of 'knowledge', and therefore how specific realities are created (2009). According to Jagger and Maier, to uncover discourse strands 'it is necessary to analyse longer periods of time' (2009, p51), which this dissertation understood as larger forms of data – that is, to identify the discourse that flows across different forms of data.

The importance of this analysis was to identify the relevance of Facebook Principles and its importance, as informed by the following statement by van Dijck:

> 'If we look more closely at Facebook's governance, the first thing to notice is that there are at least five different ToS levels: Facebook Principles, Statement of Rights and responsibilities, Privacy Policy, Data use Policy and platform Policy. All policies are derived from Facebook Principles (Van Djick, 2013, p 60)

The analysis, therefore, continued van Dijck's research 'clue' and to understand the values trapped on Facebook Principles, and how these values are reflected in the rest of governing techniques. By the end of Analysis 1, the main document that served as reference for the analysis of all techniques was Facebook Principles. As such, I revisited all data, having Facebook Principles as a reference ('Cross-reading findings in relation to other sets of data'). Specifically, to have Facebook Principles as reference served to identify a discursive flow that shows how Facebook's discourse on hate speech is formed by three values, i.e. Equality, Voice and Security. Public policies are ruled by Equity, Voices and Security. Flagging report layout responds to individual freedom and equality of users. Content standards are created respecting the principle of Voice, etc.

In considering this, I conducted a second analysis that specifically looks at the relationship between techniques.

5.8.2 Analysis 2. Analysing the data as techniques and in terms of their location.

If Analysis 1 looked at the forms of data as units of standalone information, Analysis 2 commenced to look at the forms of data as techniques of governance. It, therefore, ‘objectivises’ the forms of data and focuses on their location on the Facebook site, and the symbolic meanings of their location (Prior, 2008). The method serves to identify relationships or levels established among governing techniques as van Djick argues (2013, p60). The techniques subject of this analysis were Facebook Principles, Terms of Service, Community and Content Standards, Settings and Location, AI, Human Moderators and The Oversight Board and Mark Zuckerberg Post.

Prior observes that, in social research, sociologists and other social scientists tend to see documents primarily as sources of evidence and as receptacles of inert content. Consequently, the key strategies for data exploration have been associated with several content styles or thematic analyses, forgetting that a document is an object and that it is also susceptible to analysis (2008). Therefore, inspired by Prior (2008), I have sought to understand the importance of the location of certain techniques on Facebook's site and their symbolic meaning, by looking at how many links a document has and the actual location of the document on Facebook site and on Facebook as a corporation.

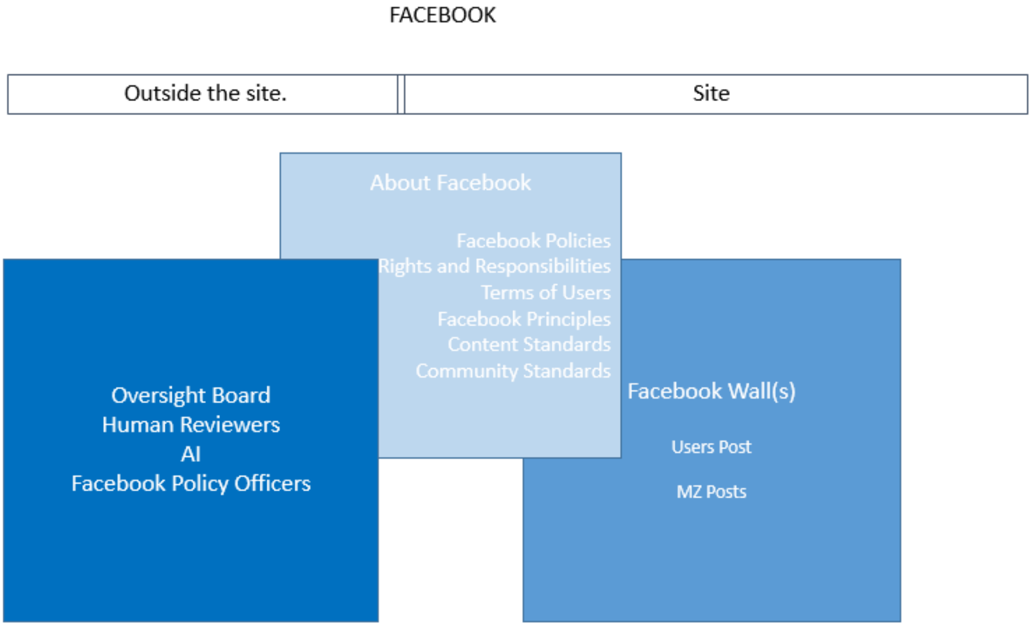


Figure 15 represents Facebook where techniques can be found having Facebook site as reference.

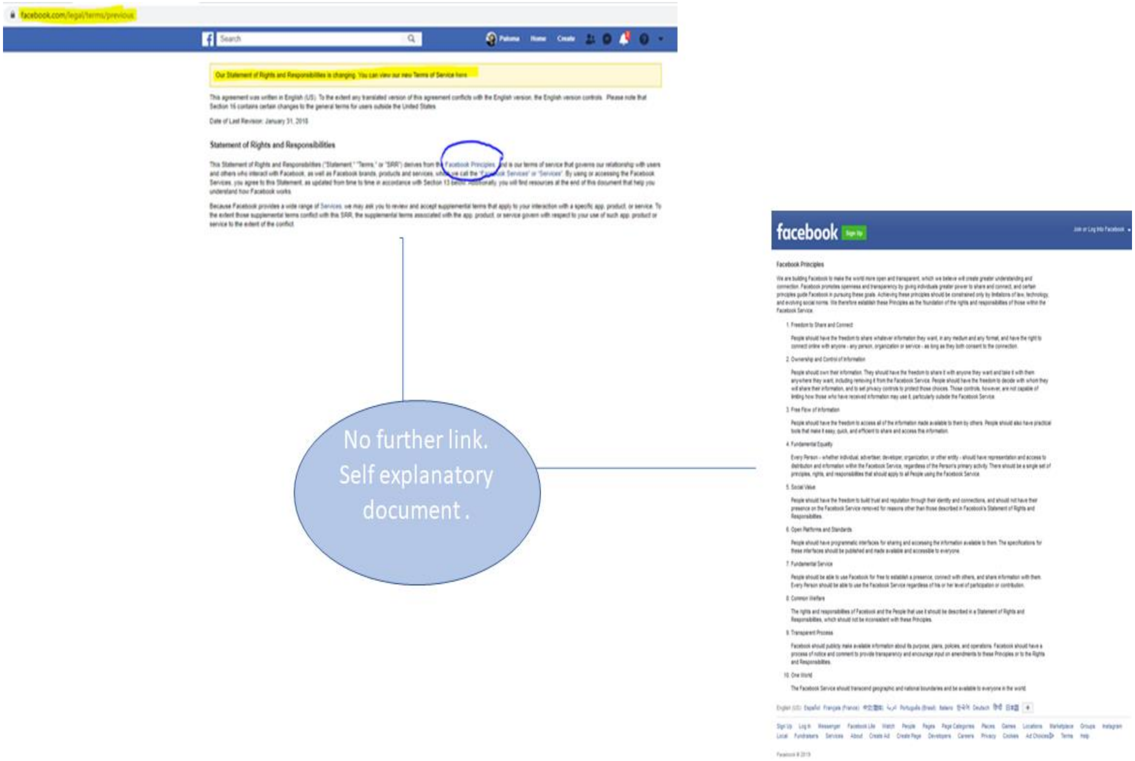
The analysis aims to understand how techniques relate to each other, and questions whether techniques can act independently (as, for example, is the claim regarding the Oversight Board), or if techniques are somehow related to each other. Indeed, links, and their disposition, do contain information about the importance of the documents and their relationship. For instance, Facebook Principles lack hyperlinks; this shows that Facebook Principles do not accommodate further explanations. In contrast, Community Standards is a document populated with many links, which indicates that the document is in need of further explanation. It also indicates that it is a document that changes content often. Below, I provide examples of how I conducted the location analysis.

5.8.2.1. Facebook Principles and leadership Location.

The first version of Facebook Principles used to be accessible through the Statement of Rights of Responsibilities.

> Date of last revision: the 31st of January 2018
>
> Statement of Rights and Responsibilities

> This Statement of Rights and Responsibilities (“Statement,” “Terms,” or “SRR”) derives from the Facebook Principles, and is our terms of service that governs our relationship with users and others who interact with Facebook, as well as Facebook brands, products and services, which we call the “Facebook Services” or ““Services”” (…)



Our Statement of Rights and Responsibilities is changing. You can view our new Terms of Service here

This agreement was written in English (US). To the extent any translated version of this agreement conflicts with the English version, the English version controls. Please note that Section 16 contains certain changes to the general terms for users outside the United States.

Date of Last Revision: January 31, 2018

Statement of Rights and Responsibilities

This Statement of Rights and Responsibilities ("Statement," "Terms," or "SRR") derives from the Facebook Principles, and is our terms of service that governs our relationship with users and others who interact with Facebook, as well as Facebook brands, products and services, which we call the "Facebook Services" or "Services". By using or accessing the Facebook Services, you agree to this Statement, as updated from time to time in accordance with Section 13 below. Additionally, you will find resources at the end of this document that help you understand how Facebook works.

Because Facebook provides a wide range of Services, we may ask you to review and accept supplemental terms that apply to your interaction with a specific app, product, or service. To the extent those supplemental terms conflict with this SRR, the supplemental terms associated with the app, product, or service govern with respect to your use of such app, product or service to the extent of the conflict.

Facebook Principles

We are building Facebook to make the world more open and transparent, which we believe will create greater understanding and connection. Facebook promotes openness and transparency by giving individuals greater power to share and connect, and certain principles guide Facebook in pursuing these goals. Achieving these principles should be constrained only by limitations of law, technology, and evolving social norms. We therefore establish these Principles as the foundation of the rights and responsibilities of those within the Facebook Service.

1. Freedom to Share and Connect

   People should have the freedom to share whatever information they want, in any medium and any format, and have the right to connect online with anyone - any person, organization or service - as long as they both consent to the connection.

2. Ownership and Control of Information

   People should own their information. They should have the freedom to share it with anyone they want and take it with them anywhere they want, including removing it from the Facebook Service. People should have the freedom to decide with whom they will share their information, and to set privacy controls to protect those choices. Those controls, however, are not capable of limiting how those who have received information may use it, particularly outside the Facebook Service.

3. Free Flow of Information

   People should have the freedom to access all of the information made available to them by others. People should also have practical tools that make it easy, quick, and efficient to share and access this information.

4. Fundamental Equality

   Every Person - whether individual, advertiser, developer, organization, or other entity - should have representation and access to distribution and information within the Facebook Service, regardless of the Person's primary activity. There should be a single set of principles, rights, and responsibilities that should apply to all People using the Facebook Service.

5. Social Value

   People should have the freedom to build trust and reputation through their identity and connections, and should not have their presence on the Facebook Service removed for reasons other than those described in Facebook's Statement of Rights and Responsibilities.

6. Open Platforms and Standards

   People should have programmatic interfaces for sharing and accessing the information available to them. The specifications for these interfaces should be published and made available and accessible to everyone.

7. Fundamental Service

   People should be able to use Facebook for free to establish a presence, connect with others, and share information with them. Every Person should be able to use the Facebook Service regardless of his or her level of participation or contribution.

8. Common Welfare

   The rights and responsibilities of Facebook and the People that use it should be described in a Statement of Rights and Responsibilities, which should not be inconsistent with these Principles.

9. Transparent Process

   Facebook should publicly make available information about its purpose, plans, policies, and operations. Facebook should have a process of notice and comment to provide transparency and encourage input on amendments to these Principles or to the Rights and Responsibilities.

10. One World

    The Facebook Service should transcend geographic and national boundaries and be available to everyone in the world.

English (US) Español Français (France) 中文(简体) العربية Português (Brasil) Italiano 한국어 Deutsch हिन्दी 日本語
Sign Up Log In Messenger Facebook Lite Watch People Pages Page Categories Places Games Locations Marketplace Groups Instagram Local Fundraisers Services About Create Ad Create Page Developers Careers Privacy Cookies Ad Choices Terms Help
Facebook © 2019

Figure 16 Illustrates the relationship between documents. Facebook Principles presents no further links and has just one link to access it

The location was through a single specific link: https://www.facebook.com/principles.php. This changed in November 2019, when it was possible to access the Principles through the ‘About Facebook’ page https://about.facebook.com/company-info. In addition to the link, I also annotated the location within the site, as Figure 25 in Chapter Six indicates[6].

[6] **Figure illustrates how location of document provides information in relation to the document. This specific example is further explained in Chapter Six, and is used to explained the symbolic location of Facebook Principles in relation to Facebook Leadership or Mark Zuckerberg.**

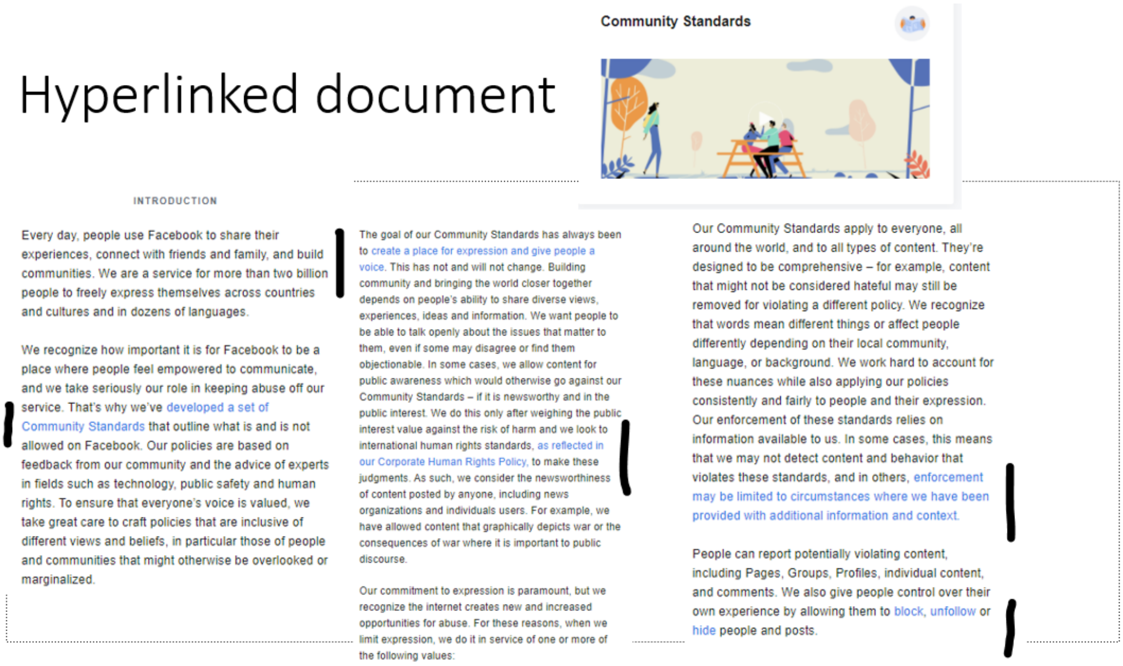
5.8.2.2. Terms of Service Location.

Facebook Terms of Services is accessible through https://www.facebook.com/terms.php and policies are accessible through https://www.facebook.com/policies. These links lead to both the Terms of Service and Policies. Before this most recent update, the User Terms, policies, and community standards were posted in separate links. However, by the time this dissertation was finalised, the three parts of this policy were amalgamated into a single link. It is important to note that Terms of Service and access to Policies vary depending on the use of Facebook or the Facebook app on mobile devices.

The Terms of Service is an overarching set of policies that have changed name and location several times in the last few years. It was first known as Right and Responsibilities and connected Facebook Principles with all Facebook Policies. Later, Rights and Responsibilities was no longer in use and Facebook published Facebook Policies and Facebook Terms of Service. In addition, the Terms of Service's contents have changed on several occasions and, since 2018, some of the content was partially replaced or relocated on Facebook's Community Standards.

5.8.2.3. Community Standard and content standard location.

The Community Standards were originally part of Facebook Policies and Terms of Service. Perhaps it is the document that has undergone the most changes concerning its location and content. Community Standards can be accessed through the link https://www.facebook.com/communitystandards/ and it is a hyperlinked document with several external references.



Hyperlinked document

Community Standards

INTRODUCTION

Every day, people use Facebook to share their experiences, connect with friends and family, and build communities. We are a service for more than two billion people to freely express themselves across countries and cultures and in dozens of languages.

We recognize how important it is for Facebook to be a place where people feel empowered to communicate, and we take seriously our role in keeping abuse off our service. That's why we've developed a set of Community Standards that outline what is and is not allowed on Facebook. Our policies are based on feedback from our community and the advice of experts in fields such as technology, public safety and human rights. To ensure that everyone's voice is valued, we take great care to craft policies that are inclusive of different views and beliefs, in particular those of people and communities that might otherwise be overlooked or marginalized.

The goal of our Community Standards has always been to create a place for expression and give people a voice. This has not and will not change. Building community and bringing the world closer together depends on people's ability to share diverse views, experiences, ideas and information. We want people to be able to talk openly about the issues that matter to them, even if some may disagree or find them objectionable. In some cases, we allow content for public awareness which would otherwise go against our Community Standards – if it is newsworthy and in the public interest. We do this only after weighing the public interest value against the risk of harm and we look to international human rights standards, as reflected in our Corporate Human Rights Policy, to make these judgments. As such, we consider the newsworthiness of content posted by anyone, including news organizations and individuals users. For example, we have allowed content that graphically depicts war or the consequences of war where it is important to public discourse.

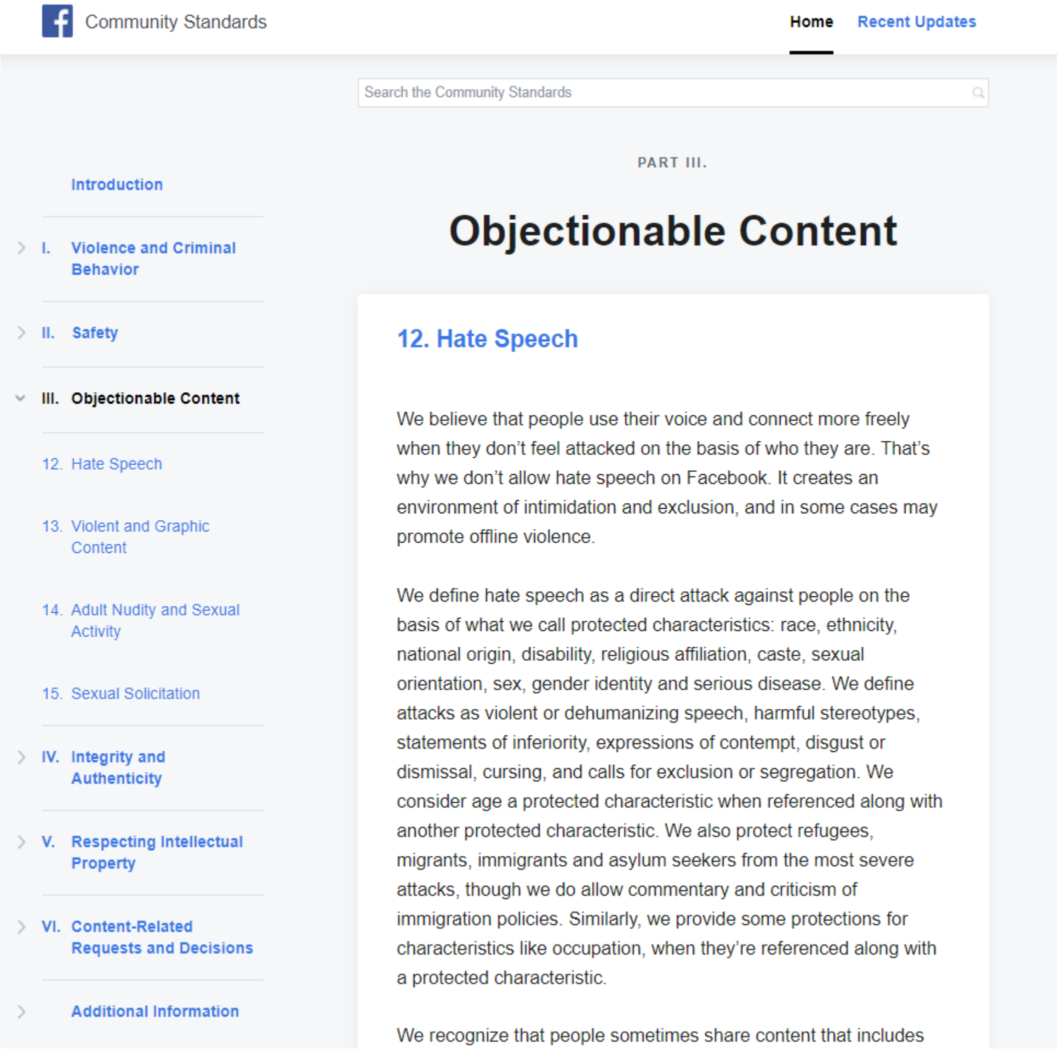
Our commitment to expression is paramount, but we recognize the internet creates new and increased opportunities for abuse. For these reasons, when we limit expression, we do it in service of one or more of the following values:

Our Community Standards apply to everyone, all around the world, and to all types of content. They're designed to be comprehensive – for example, content that might not be considered hateful may still be removed for violating a different policy. We recognize that words mean different things or affect people differently depending on their local community, language, or background. We work hard to account for these nuances while also applying our policies consistently and fairly to people and their expression. Our enforcement of these standards relies on information available to us. In some cases, this means that we may not detect content and behavior that violates these standards, and in others, enforcement may be limited to circumstances where we have been provided with additional information and context.

People can report potentially violating content, including Pages, Groups, Profiles, individual content, and comments. We also give people control over their own experience by allowing them to block, unfollow or hide people and posts.

Figure 17 Illustrates Facebook Community Standards, a document that gives access and it is accessed from different areas on the site.

In addition, it is particularly notable that hate speech policies are included in Community Standards, Section III, Objectionable Content, which has important implications for understanding hate speech governance, as noted in Chapter Six. In addition, moderation systems, i.e. Oversight Board and Human Reviewers, are located as part of the Community Standards under section *VI Content related request and decisions* (Figure 18 below). It is also noticeable that as Figure 29 will illustrate (Chapter 7), Content Standards are located on Hate Speech Policies under the section read more. As such it is understood that Content Standards are dependant of the content of Community standards on hate speech.



Community Standards

Home Recent Updates

Search the Community Standards

Introduction

I. Violence and Criminal Behavior

II. Safety

III. Objectionable Content

12. Hate Speech

13. Violent and Graphic Content

14. Adult Nudity and Sexual Activity

15. Sexual Solicitation

IV. Integrity and Authenticity

V. Respecting Intellectual Property

VI. Content-Related Requests and Decisions

Additional Information

PART III.

Objectionable Content

12. Hate Speech

We believe that people use their voice and connect more freely when they don't feel attacked on the basis of who they are. That's why we don't allow hate speech on Facebook. It creates an environment of intimidation and exclusion, and in some cases may promote offline violence.

We define hate speech as a direct attack against people on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing, and calls for exclusion or segregation. We consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants and asylum seekers from the most severe attacks, though we do allow commentary and criticism of immigration policies. Similarly, we provide some protections for characteristics like occupation, when they're referenced along with a protected characteristic.

We recognize that people sometimes share content that includes

Figure 18 shows where hate speech policies is located within the Community Standards.

5.8.2.4. Settings Location.

The location of the settings is discursively important and it was noted on Analysis 1.

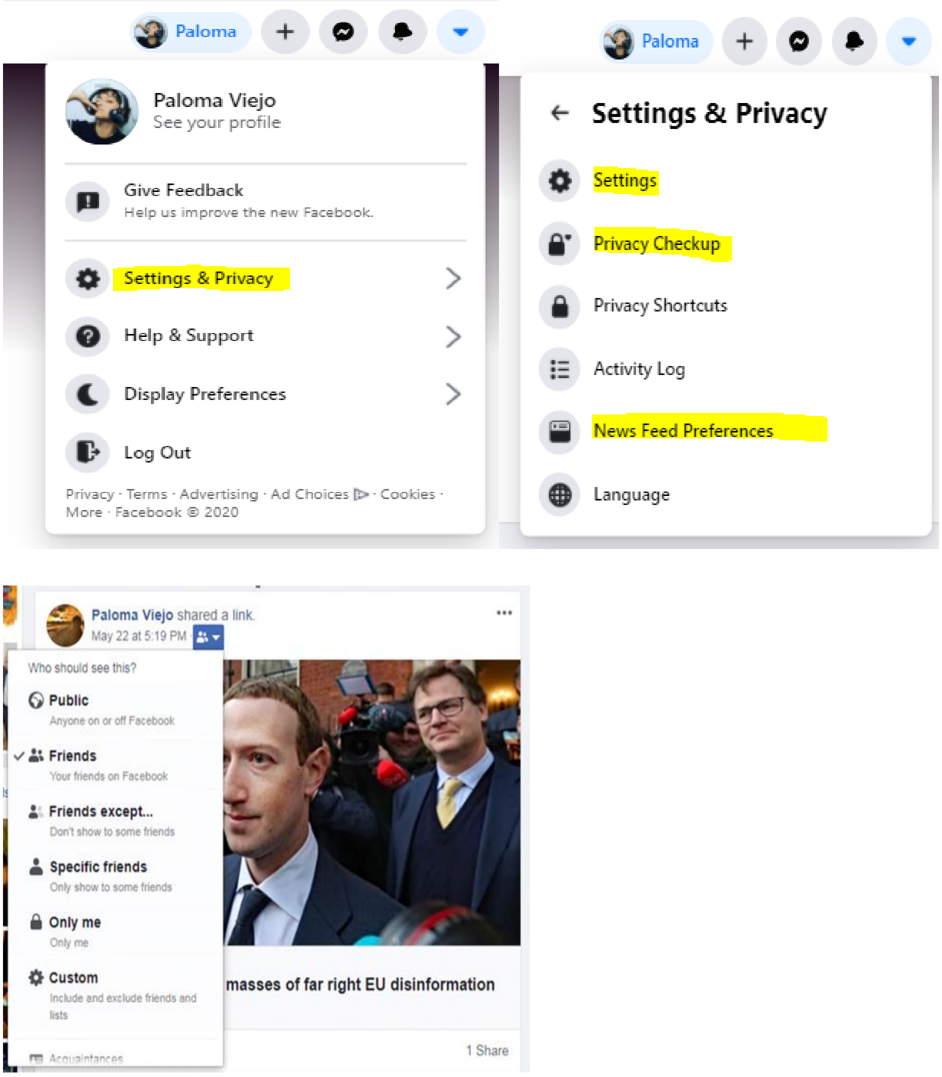


Figure 19 illustrates settings location on the site.

5.8.2.5 Human Reviewers, AI, and The Oversight Board.

As Sara Roberts Points out Human Reviewers Labour is a task specifically designed to do not be visible (2019), therefore Human Reviewers location is outside the Facebook site. Automatic Detection is neither visible, perhaps one of the more secrets activities of Facebook enterprise also hidden. Finally, the Oversight Board is located outside the site, if it is bundle to Facebook by a series of Principles as Chapter Seven will illustrate.

5.8.2.6 Mark Zuckerberg Post

Mark Zuckerberg Post are located within the site, on Mark Zuckerberg Wall, it is open to the public and its content flows across the platform freely. Therefore, these post are in all user's walls if they follow Mark Zuckerberg and public opening Mark Zuckerberg wall.

As a result of and informed by Analysis 1 and Analysis 2, I identified a particular flow between techniques. Therefore, the immediate relationship among techniques is illustrated in Figure 20. In blue, I point out how Facebook Principles and Mark Zuckerberg techniques are the source of all principles and values (see evidence in Chapter Six). Evidence of the analysis also points out that the Policy department and the Community Standards are driven by Facebook Principles (see also Chapter Six). At the same time, the Policy department oversees the creation of Content Standards (Chapter Seven), which regulates users' behaviour, and serves as a guide to automatic moderation and human moderators (see Chapter Seven).

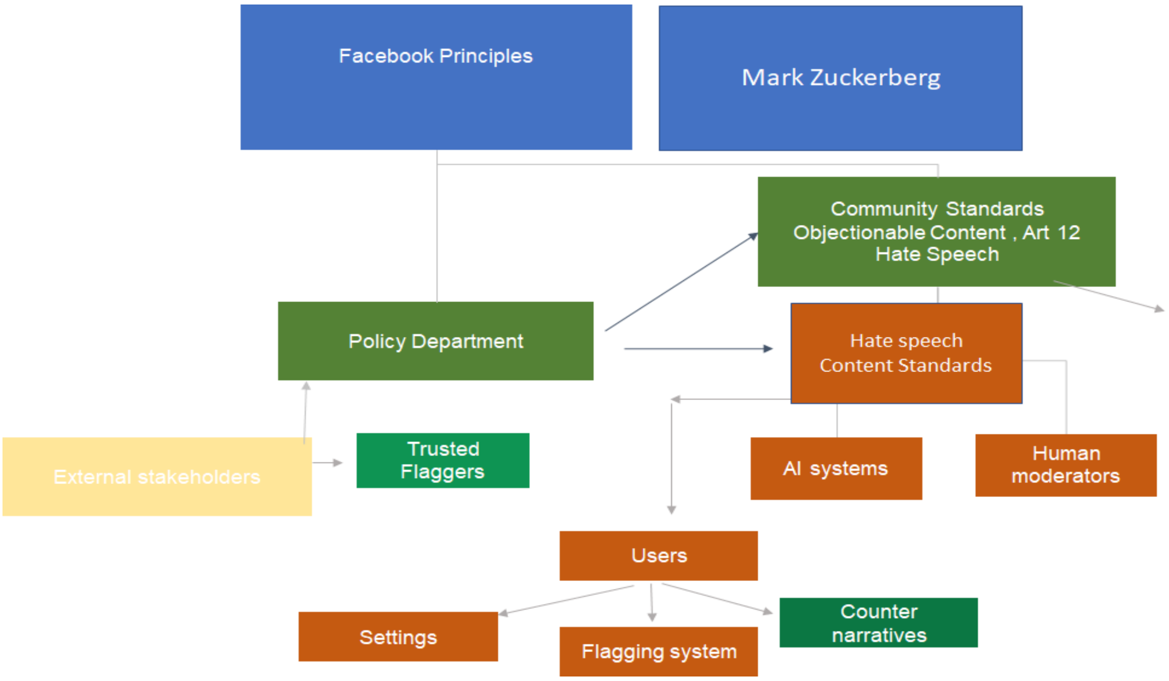


Figure 20 illustrate Facebook association of techniques as dispositives. The ideological dispositive is in blue, the disciplinary dispositive is in green, and the security dispositive is in red.

5.8.3 Analysis 3. Analysing the relationship between techniques

Techniques do not work in isolation, but in relation to others, even when they all have their independent task to fulfil. The aim of this analysis is to reorganise techniques into dispositives and, together draw Facebook's Apparatus of Governance. The analysis has been twofold. Informed by Analysis 1 and 2, I grouped the techniques according to their role in relation to their Principles, and in relation to how each technique functions to enforce Facebook Policies on hate speech. Each grouping of techniques constitutes a dispositive (Foucault, 2007), as explained in Chapter One.

Each dispositive was informed by data, and by Miller and Rose's and Foucault's studies of governance. Specifically, ideology, discipline, and enforcement were drawn from Miller and Rose's governing studies (2008) and Foucault's studies on governmentality. Ideology refers to all the techniques that contained the series of beliefs and values that Facebook applies to inform its hate speech policies and operational systems The next category is disciplinarian, which I drew from Foucault's Security Territory and Population (1978), where I look at all techniques with the authority to arrange other techniques around the platform (e.g. policy officers). Finally, Enforcement Mechanisms derives from Milosevic (2018), who studied the techniques that social media employ to report and punish bullying content.

5.8.3.1 Dispositive of Ideology

This refers to the group of techniques that influence or determine the values embedded in Facebook's apparatus of governance. Under this category, the present dissertation identifies two techniques: Facebook Principles and Mark Zuckerberg's posts.

5.8.3.2. Dispositive of discipline.

This includes all the techniques with the capacity to create policies and the authority to implement them. I have identified policy officers and everything that is associated with their work and which also functions as techniques, namely: community standards, content standards, selection and training of human reviewers, disciplining trusted flaggers, creating and implementing counter narrative programmes, and engaging with external stakeholders.

5.8.3.3. Dispositive of enforcement:[7]

This includes all documents and tools by which the governance of hate speech is executed on the platform. The present dissertation includes under this category: User Settings, Automatic Detection, Human Reviewers, and Oversight Board.

The results are illustrated in Chapter Eight, wich delves on the table below.

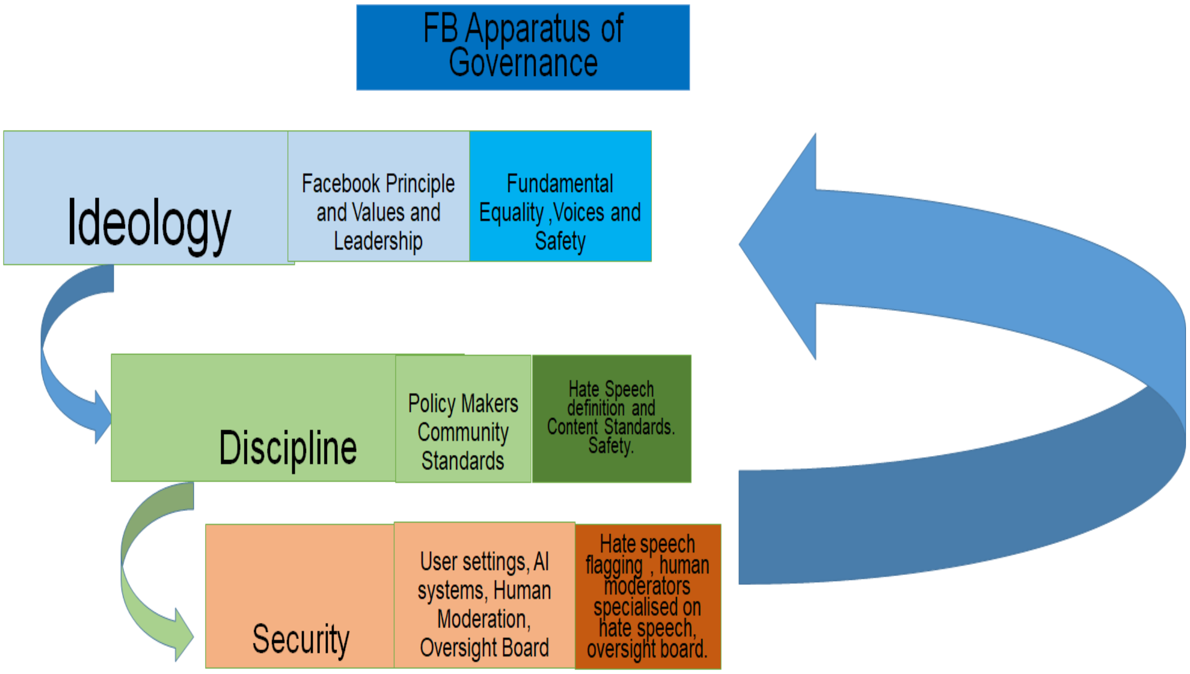

FB Apparatus of Governance
Ideology
Facebook Principle and Values and Leadership
Fundamental Equality ,Voices and Safety
Discipline
Policy Makers Community Standards
Hate Speech definition and Content Standards. Safety.
Security
User settings, AI systems, Human Moderation, Oversight Board
Hate speech flagging , human moderators specialised on hate speech, oversight board.

Figure 21 Illustrates Facebook Governance apparatus according to the relationship between dispositives. (also in Chapter 8)

[7] Later referred to as dispositive of security.

5.9 Conclusion

The methodology reflects the ambition to find out how Facebook governs hate speech and its values and principles that feed into how it conceptualises and operationalises hate speech. The chapter has explained that the fundamental reason for focusing on Facebook is its number of users and its capacity to influence users with its policies. Next, the chapter defines the research paradigm that underpins this dissertation, which I situated between constructivism and critical paradigm, with the latter driving the research. The chapter described the research design, the data selection, data collection and data analysis, while functioning as an example of the complexity of researching a dynamic environment as Facebook.

Chapter 6. Facebook Principles and Values

6.1 Introduction

Chapter Six is the first empirical chapter and it analyses the Facebook Principles and their influence on Facebook's definition of hate speech. The chapter aims to answer the following research question: What are Facebook's principles and values around their definition of, and approach to, hate speech?

A principle is widely defined as a basic idea or rule that explains and controls how something works (Cambridge Dictionary, 2020). Facebook Principles is a company-drafted document, first posted on the platform on 3rd June 2009 (Platform governance Archive, 2021). This document captures Facebook's socio-political vision, determines the design of its policies and operations, and is colloquially known within the company as Facebook DNA (Siapera and Viejo Otero, 2021). The document has had two versions; the first version was on site from 2009 to 2019 and comprised 10 Articles, with a second version introduced in November 2019 that comprised five Articles.

Jose van Djick (2013) says of Facebook Principles that it is a hierarchically superior document that influences all terms of Service and Policies on Facebook (2013). However, what makes Facebook Principles one of the most important documents on Facebook? And what are the actual means by which it influences the rest of Facebook policies and strategies? To answer this question, the chapter is organised as follows: firstly, it analyses the location of Facebook Principles; secondly, it examines the relationship between Facebook Principles and the posts of Facebook's CEO, Mark Zuckerberg. This specific analysis exposes the close relationship between those techniques. Next, the chapter explores the specific principles that Facebook aims to push forward to justify its hate speech policies and the values that it grants to the following principles: Connectivity and Community Equality, Freedom or Voices, and

Security. The chapter then looks at how Facebook Principles permeate and influence different techniques such as Facebook public policies and user settings. Finally, based on these analyses, the chapter unveils how Facebook Principles on equality and freedom shape Facebook's hate speech policy.

The chapter concludes with two findings. First, taking into account Facebook Principles and the values Facebook gives to them, Facebook approaches hate speech through the Neutral Viewpoint as seen in Chapter Three. Second, taken together, Mark Zuckerberg and Facebook Principles constitute a dispositive ideology.

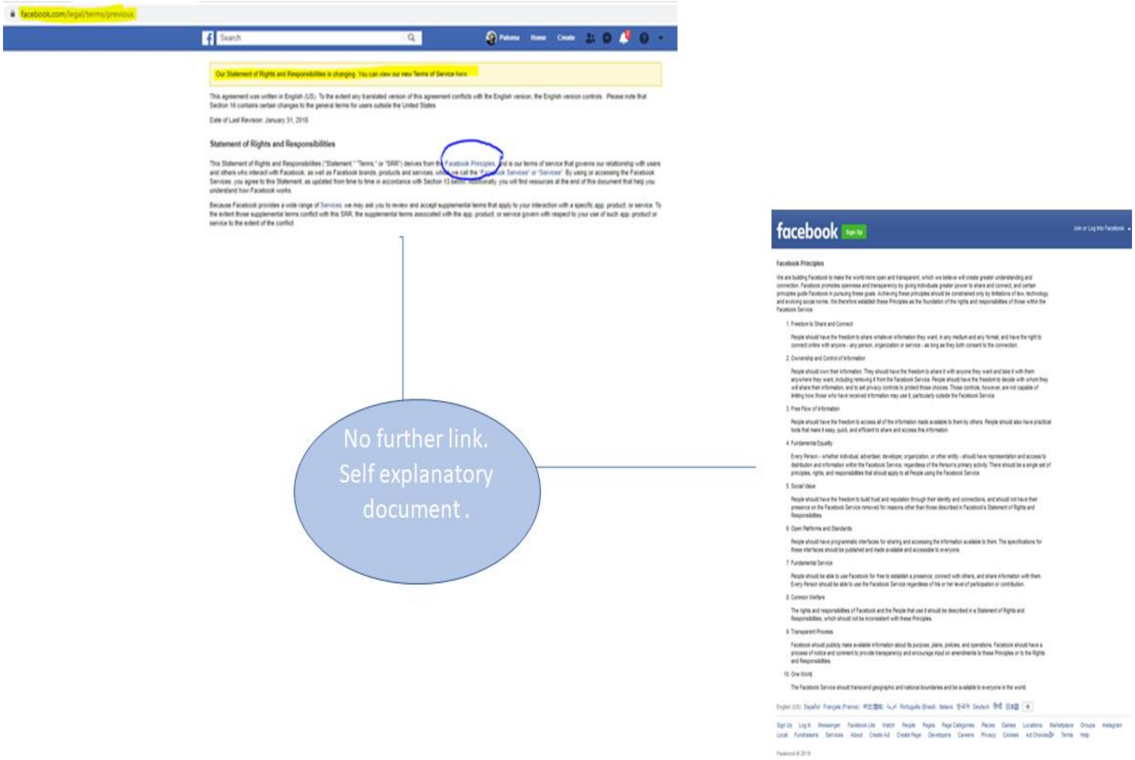
6.2 Facebook Principles Location

During the period 2009 to 2019, Facebook Principles was a document that had never changed its content. It did, however, change its location – or, more specifically, the route by which the document could be accessed. During this decade, users accessed the document through the 'Rights and Responsibilities' page, as shown in the following paragraph extracted directly from the Facebook page and which is currently no longer available.

The statement of Rights and Responsibilities reads as follows:

> This Statement of Rights and Responsibilities ("Statement," "Terms," or "SRR") derives from the Facebook Principles, and is our terms of service that governs our relationship with users and others who interact with Facebook, as well as Facebook brands, products and services, which we call the "Facebook Services" or "Services". By using or accessing the Facebook Services, you agree to this Statement. (Facebook, 2020. Retrieved from https://www.facebook.com/legal/terms/previous. Last retrieved on 05/11/2020) (emphasis added).

Once the link was clicked, Facebook Principles presented as a simple, not dynamic, document. It did not offer any other link. Accessing Facebook Principles through the Rights and Responsibilities gave the user the feeling of arriving at a foundational document. In July 2018, the document lost this feature and was no longer accessible through the Statement of Rights and Responsibilities. Instead, its location could only be found via an active search in the search bar https://www.facebook.com/principles.php. Despite the change on how to access it, the document continued to offer no other link, and maintains its aesthetics of a Decalogue table of the law. That is a primordial, simple document that contains 10 fundamental principles that refer to Facebook Principles.



Our Statement of Rights and Responsibilities is changing. You can view our new Terms of Service here.

This agreement was written in English (US). To the extent any translated version of this agreement conflicts with the English version, the English version controls. Please note that Section 16 contains certain changes to the general terms for users outside the United States.

Date of Last Revision: January 31, 2018

Statement of Rights and Responsibilities

This Statement of Rights and Responsibilities ("Statement," "Terms," or "SRR") derives from the Facebook Principles, and is our terms of service that governs our relationship with users and others who interact with Facebook, as well as Facebook brands, products and services, which we call the "Facebook Services" or "Services". By using or accessing the Facebook Services, you agree to this Statement, as updated from time to time in accordance with Section 13 below. Additionally, you will find resources at the end of this document that help you understand how Facebook works.

Because Facebook provides a wide range of Services, we may ask you to review and accept supplemental terms that apply to your interaction with a specific app, product, or service. To the extent those supplemental terms conflict with this SRR, the supplemental terms associated with the app, product, or service govern with respect to your use of such app, product or service to the extent of the conflict.

Facebook Principles

We are building Facebook to make the world more open and transparent, which we believe will create greater understanding and connection. Facebook promotes openness and transparency by giving individuals greater power to share and connect, and certain principles guide Facebook in pursuing these goals. Achieving these principles should be constrained only by limitations of law, technology, and evolving social norms. We therefore establish these Principles as the foundation of the rights and responsibilities of those within the Facebook Service.

1. Freedom to Share and Connect

   People should have the freedom to share whatever information they want, in any medium and any format, and have the right to connect online with anyone - any person, organization or service - as long as they both consent to the connection.

2. Ownership and Control of Information

   People should own their information. They should have the freedom to share it with anyone they want and take it with them anywhere they want, including removing it from the Facebook Service. People should have the freedom to decide with whom they will share their information, and to set privacy controls to protect those choices. Those controls, however, are not capable of limiting how those who have received information may use it, particularly outside the Facebook Service.

3. Free Flow of Information

   People should have the freedom to access all of the information made available to them by others. People should also have practical tools that make it easy, quick, and efficient to share and access this information.

4. Fundamental Equality

   Every Person - whether individual, advertiser, developer, organization, or other entity - should have representation and access to distribution and information within the Facebook Service, regardless of the Person's primary activity. There should be a single set of principles, rights, and responsibilities that should apply to all People using the Facebook Service.

5. Social Value

   People should have the freedom to build trust and reputation through their identity and connections, and should not have their presence on the Facebook Service removed for reasons other than those described in Facebook's Statement of Rights and Responsibilities.

6. Open Platforms and Standards

   People should have programmatic interfaces for sharing and accessing the information available to them. The specifications for these interfaces should be published and made available and accessible to everyone.

7. Fundamental Service

   People should be able to use Facebook for free to establish a presence, connect with others, and share information with them. Every Person should be able to use the Facebook Service regardless of his or her level of participation or contribution.

8. Common Welfare

   The rights and responsibilities of Facebook and the People that use it should be described in a Statement of Rights and Responsibilities, which should not be inconsistent with these Principles.

9. Transparent Process

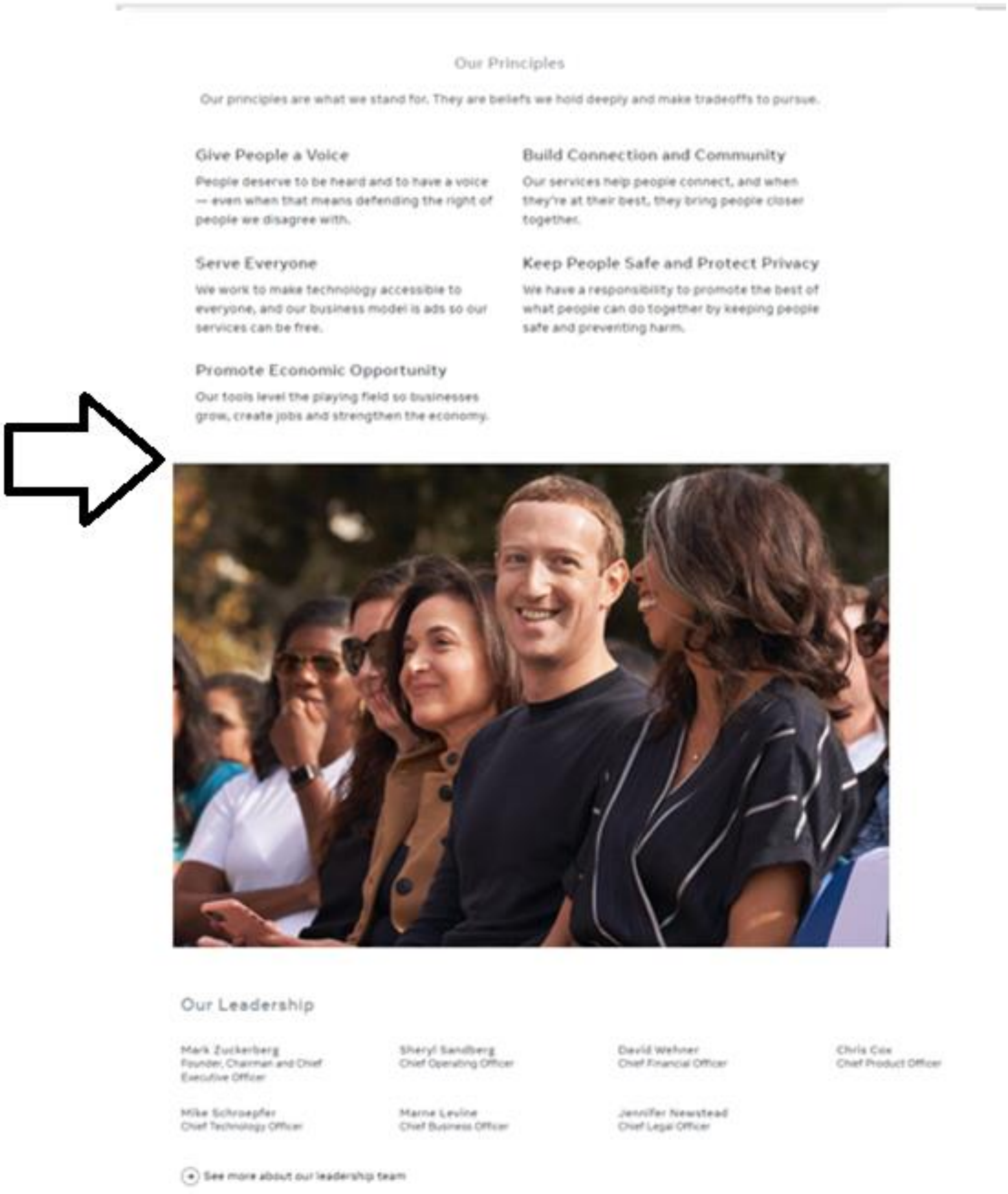
   Facebook should publicly make available information about its purpose, plans, policies, and operations. Facebook should have a process of notice and comment to provide transparency and encourage input on amendments to these Principles or to the Rights and Responsibilities.

10. One World

   The Facebook Service should transcend geographic and national boundaries and be available to everyone in the world.

Figure 22 illustrates the relationship between documents. Facebook Principles presents no further links and has just one link to access it.

The change of location in 2018, however, propounds a deeper change, which arrived in November 2019. Whereas Facebook only announced a change of the logo and general Facebook outlook, the company also changed access, localisation, the aesthetic appearance and content of Facebook Principles (see Zuckerberg, 4th November 2019). Facebook Principles could be accessed through the following link https://about.facebook.com/company-info. Examining the way in which information is displayed and positioned on the page, it is of particular significance that the location of Facebook Principles was right above leadership names, symbolically indicating the relationship among Principles and leadership, an issue that is unpacked in the following section. Specifically, the location of this document and the explicit linking between the Principles and leadership allude to the tightening of the relationship and point outs to a more hierarchical structure as below figure 23 illustrates.



Our Principles

Our principles are what we stand for. They are beliefs we hold deeply and make tradeoffs to pursue.

Give People a Voice

People deserve to be heard and to have a voice — even when that means defending the right of people we disagree with.

Build Connection and Community

Our services help people connect, and when they're at their best, they bring people closer together.

Serve Everyone

We work to make technology accessible to everyone, and our business model is ads so our services can be free.

Keep People Safe and Protect Privacy

We have a responsibility to promote the best of what people can do together by keeping people safe and preventing harm.

Promote Economic Opportunity

Our tools level the playing field so businesses grow, create jobs and strengthen the economy.

Our Leadership

Mark Zuckerberg
Founder, Chairman and Chief Executive Officer

Sheryl Sandberg
Chief Operating Officer

David Wehner
Chief Financial Officer

Chris Cox
Chief Product Officer

Mike Schroepfer
Chief Technology Officer

Marne Levine
Chief Business Officer

Jennifer Newstead
Chief Legal Officer

See more about our leadership team

Figure 23 illustrates the position of Facebook Principles and shows the symbolic relation among Principles and Facebook Leadership

6.3 Facebook Principles and Mark Zuckerberg,

This section aims to explore the relationship between Facebook Principles and the posts of Mark Zuckerberg, arguing that both techniques are the source of knowledge that underpins the values that shaped Facebook's schemes of intervention.

The analysis of the location of Facebook Principles reveals both the importance of the document and the symbolic relationship between Facebook Principles and Mark Zuckerberg. However, this is not only a symbolic relationship. Indeed, the analysis of Mark Zuckerberg's posts indicates that Facebook Principles are of his specific authorship.

In Founders Letter 2012, Five years later (2017), Zuckerberg suggested that he wrote the first version of Facebook Principles and values himself.

> RE: Facebook Mission and Facebook Business
>
> As I said above, Facebook was not originally founded to be a company. We've always cared primarily about our social mission, the services we're building and the people who use them. This is a different approach for a public company to take, so I want to explain why I think it works. I started off by writing the first version of Facebook myself because it was something I wanted to exist. (Zuckerberg, 1st February 2017)

Indeed, Mark Zuckerberg's public posts on his Facebook profile provide evidence that the Facebook CEO has always been vocal regarding Facebook's ideology as a company that would change society profoundly and make it 'more transparent, efficient and tolerant' (Zuckerberg, 20 April 2009). However, the question remains: which values and value system does Zuckerberg draw upon? In order to answer this question, and in order to understand the relationship between Zuckerberg's posts and Facebook Principles, I analysed a total of 280 of Mark Zuckerberg's written posts and publications from May 2016 to November 2020. During the analysis, I first became familiar with his written style and patterns. Next, I identify chronologically how Zuckerberg creates and builds up specific discourses to then relate them with Facebook Principles. To back up this finding, I run a series of word counts. Finally, I compared Mark Zuckerberg's posts with Facebook Principles. The results are outlined below.

6.3.1 Zuckerberg discourse patterns

Mark Zuckerberg's posts can be described as ideological-techno-corporate crafted with an optimistic and proactive tone, lacking historical or anthropological complexity. Arguments typically adopt the following pattern: first, Zuckerberg tends to present an ideological/reflective vision for society; second, he offers an operational explanation or solution of how technology can carry out and accomplish these visions; and third, he tends to present a Facebook product which will contribute to this ideological end. To illustrate this point, I have included two random examples from 28 April 2021 and 1st April 2021. The examples used are thematically different; however, the structure of both posts follows the above-mentioned pattern.

Post 1: Business needs.

> ‘For a lot of people, online commerce is less about websites and shops, and more about messaging. People want to get support and make purchases right from a chat. (problem)
>
> Businesses using the WhatsApp Business API are already sending more than 100 million messages per day. (Technological solution)
>
> Since we introduced "Click-to-WhatsApp" ads nearly 1 million advertisers have already started using them too. Now the next step here is we're going to make it possible for them to create those "Click-to-WhatsApp" ads from within the WhatsApp Business app. (Facebook contributes to solving the problem)
>
> (Zuckerberg, 28 April 2021)

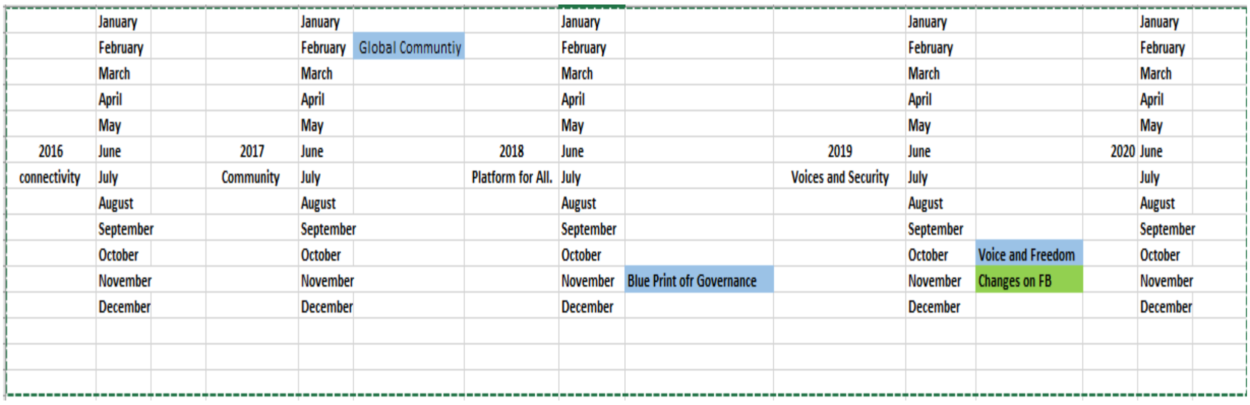
Post 2: Covid-19 Vaccination.

> If we're going to stop Covid, we need everyone who's eligible to get vaccinated. People are more likely to get vaccinated if they see friends, family and people they trust doing it too. (Problem)
>
> So we're launching a new Covid vaccine profile frame that you can add to your profile pic, partnering with the CDC and US Department of Health and Human Services to launch new Covid vaccine profile frames. (Technological solution)
>
> It lets you easily show your support and tells people that you've been vaccinated. And we'll show you in the News Feed your friends who have put up this profile frame. I'll update mine after I get vaccinated, which I plan to

> do as soon as I'm eligible. I hope everyone else does the same! (Facebook contributes to solving the problem)
>
> (Zuckerberg, 1st April 2021)

6.3.2 Chronological analysis.

Following the familiarisation with Zuckerberg's writing style, the analysis focused on how Zuckerberg builds discourse throughout the years. The analysis seeks to contribute to understanding the main ideas that Facebook wants to create and solidify, and the general direction of the platform. In addition, the chronological analysis also reveals a particular characteristic of Zuckerberg's posts, which is the gradual construction of discourse. Once it is matured, he presents it in a different format such as a letter, as Figure 24 illustrates. Overall, the present section shows that Zuckerberg's discourse progressed chronologically, illustrating the changes and direction that Facebook has taken.



	January		January			January			January			January
	February		February	Global Communtiy		February			February			February
	March		March			March			March			March
	April		April			April			April			April
	May		May			May			May			May
2016	June	2017	June		2018	June		2019	June		2020	June
connectivity	July	Community	July		Platform for All.	July		Voices and Security	July			July
	August		August			August			August			August
	September		September			September			September			September
	October		October			October			October	Voice and Freedom		October
	November		November			November	Blue Print ofr Governance		November	Changes on FB		November
	December		December			December			December			December

Figure 24 illustrates the distribution of topics per years and the relevant documents

#2016 Zuckerberg emphasises the role of 'connecting' and being connected to communities. Zuckerberg clearly discussed this topic in his keynote speech in Peru during the Innovation Summit (19 November 2016) and even through the promotion and celebration of Facebook Connectivity Lab (17 November 2016).

#2017 focuses on the idea of community (see post 6 June 2017). The notion of community was firstly tackled in Zuckerberg's letter, regarding building a Global Community, issued on 17 February 2017 (which this chapter will examine in a following section). This is followed by a series of posts and interviews in which Zuckerberg presented the 'New mission of Facebook', which 'is to bring the world closer together' (Zuckerberg, 22 June 2017). This is the year that Facebook reinforced the role of Community Standards.

#2017- 2018. In addition to Connectivity and Community, Facebook began to introduce the notion of 'platform for all'. These years represent the end of techno optimism that began to be visible on the platform. Zuckerberg defended the position of a platform that treats all users and their views equally – a 'platform for all':

> Every day I work to bring people together and build a community for everyone. We hope to give all people a voice and create a platform for all ideas. (Zuckerberg, 27 September 2017)

#2018-2019 During this period, Facebook changes drastically. The focus is on not losing the foundational values, and in particular the value of voice, while introducing safety measures.

> 'We're a very different company today than we were in 2016, or even a year ago. We've fundamentally altered our DNA to focus more on preventing harm in all our services, and we've systematically shifted a large portion of our company to work on preventing harm. (Zuckerberg, 28 December 2018)
>
> 'The difficulty lies in fighting harm while allowing freedom' (Zuckerberg, 17 October 2019).

#2019-2020. Facebook continues with the notion of Safety and Voice, but focuses on the Oversight Board (6 May 2020).

6.3.2 Zuckerberg Word count and frequency

In order to back up previous information, I extracted the most repeated themes and topics of Connectivity, Community, Equality (or Platform for all), Harmful content, Voices and

Safety, and using the Zuckerberg Transcripts database I ran an analysis that aimed to understand the time and frequency of the words. This analysis was run to back up previous observations.

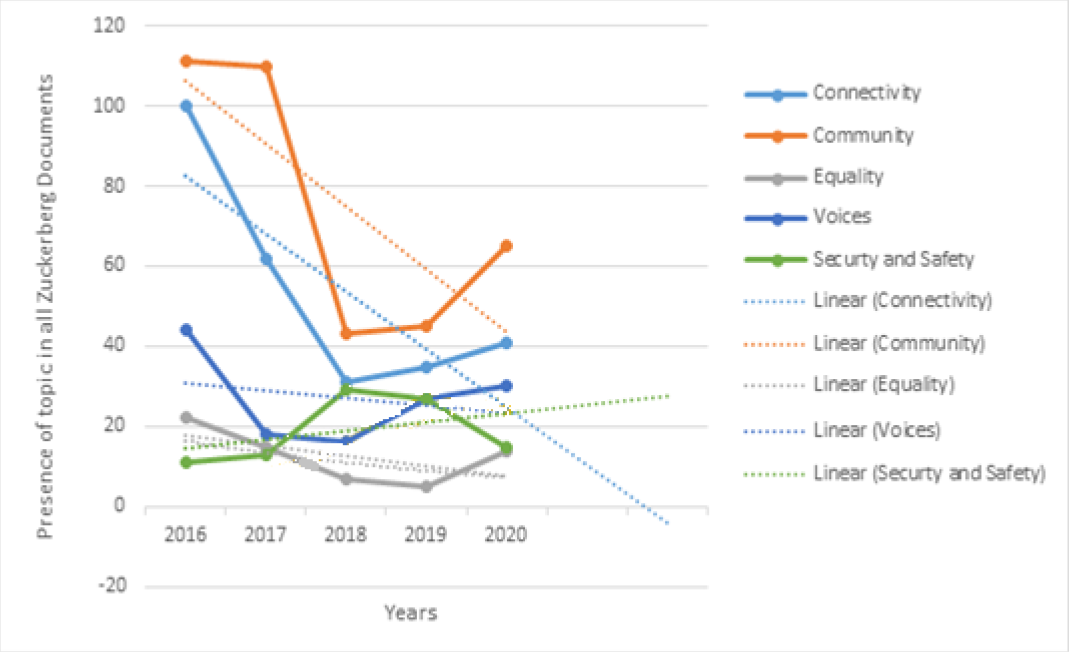


Figure 25 Illustrates How many times Zuckerberg has repeated a term and its evolution per year.In addition, and with dot-bar, the figure presents a linear forecast, predicting if a topic will be more or less repeated.

First, the analysis showed that Connectivity and Community are the most repeated topics throughout the years, while Equality (and Platform for All), Voices, Security and Safety appear in Zuckerberg's discourse less often. In terms of frequency, and since 2018, Community and Connectivity have the tendency to decrease, while Equality and Platform for all follow a regular pattern. Since 2017, Safety and Security and Voices are more often repeated.

6.3.3. Zuckerberg themes and Facebook Principles

The final step of the analysis illustrates topics that overlaps between Facebook Principles and Mark Zuckerberg discourse.

TIMELINE	Themes in Zuckerberg's posts	Facebook Principles
2016-2017	Connectivity Global Community Community (Building a platform Collaborative, safe, informed, civically engaged, inclusive).	Article 10 One world (Version 2009)
2017-2018	Platform for all or (Equality) Voice and Freedom of Expression. Connectivity	Article 4 *FundamentalEquality*. (Version 2009) Article 1 *Freedom to Share and Connect* (Version 2009) Article 3 *Free Flow of Information.* (Version 2009)
2019-2020	Voice and Safety Connectivity	Give People a Voice Keep People Safe and Protect Privacy Build Connection and Community (Version 2019)

Table 2 shows the evolution of Mark Zuckerberg discourse and its relation with Facebook Principles

In considering the similarity between Facebook Principles and Zuckerberg discourse, this chapter shows that there is a tight relationship between Zuckerberg and Facebook Principles, revealing that both Facebook Principles and Mark Zuckerberg constitute together a dispositive. This dispositive marks the general direction of Facebook and constitutes what this dissertation will refer to as Dispositive of Ideology. The section below outlines the fundamental principles that this dispositive has created or re-created.

6.4 Facebook's Fundamental Principles

The analysis of Facebook Principles and Mark Zuckerberg posts contributes to an understanding of Facebook's Fundamental Principles. The analysis of Facebook Principles reveals how Facebook defines two of the fundamental values attached to hate speech: Equality and Freedom. In considering this, the following paragraphs focus on how Facebook defines the values Equality and Freedom. In addition, the section includes the analysis of Connectivity, Community or Safety. In the discussion of previous moments in the history of hate speech, Connectivity, Community or Safety would not be considered influential principles. However, Facebook includes them in its Principles, and indeed their formation and rationale serves Facebook to justify how hate speech is understood and how hateful content is operationalised. Specifically, Connectivity justifies Facebook's technological design, Community justifies the need to intervene in the life of Facebook users, while Equality justifies that there is one single set of rules, applied to all users. Freedom, however, is a rather more complex term that Facebook defines in connection with information and safety, to such an extent that it ended up developing a unique and custom-made definition of Freedom in order to accommodate enforcement mechanisms that provide safety for users and enforce hate speech policies. The following sections unpack these notions in more detail.

6.4.1 Connectivity.

The idea of connectivity is reflected in Article 10 One World (Facebook Principles, 2009-2019): ‘The Facebook Service must transcend geographical and national borders and be available to everyone in the world’ (Facebook Principles, 2009-2019). It is also presented as a mission: Connectivity as a mission is ‘to give people the power to share and make the world more open and connected’ (Facebook Principles November, 2019).

As indicated by Barbrook and Cameron (1996), McLuhan’s (1964) ‘electronic agora' was a very influential idea for early internet developers. In fact, Zuckerberg’s comments refer to the same idea of connecting the world in one space, revealing the influence of McLuhan’s aspirations. In particular, Mark Zuckerberg recalls in one of his posts that, when he was a student, it was evident to him that 'someone was going to build a community to connect the whole world. We never thought it could be us’ (Zuckerberg, 16 August 2016; interview 22 August 2016). Digital technology was rising and connectivity in the world was inevitable, which Zuckerberg, among all possibilities that human imagination can conceive, interpreted as ‘the ability to share information with all your “friends” at the same time and in the same place’ (Zuckerberg, 5 September 2016). In order to facilitate this particular idea of Connectivity, Zuckerberg et al. (2006) created and patented a series of technological methods known as News Feed that serve as a platform for all users to connect. Zuckerberg promoted the idea of connectivity in 2016 at several conferences, meetings, and through posts. These include the Summit in Peru on 16 November 2016, where Zuckerberg presented Facebook’s ‘safety check’ (2016, ‘10.18), by which users can connect with their loved ones and check if they are safe in case of disasters. The analysis of posts for the year 2016 reveals that, by that time, Facebook’s idea of connectivity blends with supporting external communities, particularly those who use Facebook, which Zuckerberg refers to as ‘our community’ (18 November 2018). This understanding shifts in 2017, where the notion of connectivity and community turns inwards, referring instead to the Facebook Community.

> For the past decade, Facebook has been focused on making the world more open and connected -- and we're always going to keep doing that. But now it's clear we have to do more. We also need to bring people closer together and build common understanding. One of the best ways to do that is by helping people build community, both in the physical world and online. (Zuckerberg, 6 June 2017)

6.4.2 Community.

With the aim of reinforcing what is a Community for Facebook, Zuckerberg wrote on the topic in his letter 'Building Global Community' on 17 February 2017. Building a Global Community is not an ordinary post, but one of the first governance documents issued by Zuckerberg that tacitly informs users of Facebook's intention to govern (Terranova, 2017). It was a letter that, according to Terranova, reflected that Facebook was formulating what Foucault would describe as a 'political rationality' (2007, p.278).

'Building global community' is the perfect example of what has been defined as Zuckerberg's 'corporate sociotechnical imaginary' (Haupt, 2021). It follows the same pattern that the majority of Zuckerberg's posts display: first exposing a problem, then a technological solution made by Facebook, and concluding by explaining how Facebook will contribute to solving the problem except for the last part of the letter, when Zuckerberg describes the problems arising around harmful content and the role of the community standards, which Zuckerberg describes as a challenge and a problem. Building a Global Community was followed in 2018 by 'Blueprint for Facebook Governance and Enforcement' (15 November 2018), which is different from Building a Global Community, as it does not present an ideal version of governance, but presents practical solutions and enforcement to manage content. In this section, I focus on the document Building a Global Community, the content of which, as Terranova has argued, reveals Facebook's 'social aspirations' (2017) and, more specifically, what is and what constitutes Community Standards.

6.4.2.1 Building a global Community

When analysed in detail, the most striking characteristic of Building a Global Community is that it is attached to Zuckerberg's own perceptions of the world. Zuckerberg opens 'Building a Global Community' with an interesting statement: 'the power of traditional institutions is weakening' (2017). By traditional institutions, Zuckerberg refers to churches, associations, or families. Where he draws the statement that traditional institutions are weakening is unknown, as Zuckerberg does not state where his thoughts emerged from.

1) Communities: With no further explanation, Zuckerberg states that, since the well-being of small communities is the very foundation of the social fabric, Facebook embarked on the mission to strengthen communities, followed by a solution, which is a FB product – Facebook Groups. 'Facebook Groups' is a feature that enables communication among a limited number of users. Zuckerberg continues by saying that communities not only want to be stronger, but safe, referring to safety as protection from external threats such as terrorism or climate change.

2) Safety: Building a Global Community does not refer to safety as an internal matter, but to the notion of how Facebook can provide security measures and guarantee the safety of the social fabric, since, according to Zuckerberg (2017), governments might not have adequate infrastructure. Therefore, in order to contribute to the security of the community, Zuckerberg continues the promotion of the use of 'Safety Check' (Zuckerberg, 12 June 2016, 21 June 2016, 6 February 2017, September 2017), an app that facilitates cooperation and self-organisation in case of emergency.

3) Civic engaged community. A third feature of any Facebook community is a community that is civically engaged, understood as 'citizens who do not participate in institutional or self-govern organizations' (Zuckerberg, 17 February 2017). That is, Zuckerberg suggests that the lack of civic engagement with the establishment is a problem. Therefore, in order to

promote engagement in civic affairs, Facebook created systems to participate in collective decision-making, which this dissertation has not explored further but which include, for example, systems that enable voting registration in the United States.

4) Informed and Inclusive Community: The fourth idea for a Facebook community that Zuckerberg spells out in this letter is that of creating informed and tolerant individuals – in other words, creating a 'platform for all views' by which Facebook is open to all types of opinions and speech (Zuckerberg, 17 February 2017). The problem, however, that Zuckerberg encounters and describes in this letter is that of misinformation (Zuckerberg, 18 November 2016, 15 December 2016, 17 February 2017). Zuckerberg proposes that control of misinformation can be achieved by 'reducing sensationalist pieces of content in News Feed and identifying sensational publishers and facilitating the connection to the internet for those who do not have access to basic information' (Zuckerberg, 17 February 2017).

5) Community Standards: Finally, Zuckerberg talks about the problem of hateful content, for which Zuckerberg changes the recurrent pattern of: problem-technological solution-Facebook product, to present dealing with hate speech as a personal struggle, a challenge and a problem. This time, the solution is not a technological solution and a Facebook product, but a more complex system that combines the creation of Community Standards with AI.

> We need a system where we can all contribute to setting the standards. Although this system is not fully developed, I want to share an idea of how this might work. The guiding principles are that the Community Standards should reflect the cultural norms of our community, that each person should see as little objectionable content as possible, and each person should be able to share what they want while being told they cannot share something as little as possible. The approach is to combine creating a large-scale democratic process to determine standards with AI to help enforce them. (Zuckerberg, 17 February 2017)

Building a Global community is initially a compilation of problems that Zuckerberg perceives that, with the related products that Facebook creates, can be resolved. But this approach reaches a limit when it comes to harmful content. A difference between harmful content and weak communities, or safety from natural disasters, is that harmful content is an internal problem that does not only affect individuals but also damages Facebook's reputation. It is no surprise, therefore, that Zuckerberg ends the letter with a piece on Community and Content Standards, which can be interpreted as the most important message that this letter carries, since it is the first time that Zuckerberg says that Facebook is willing to intervene in the type of content that users are uploading (Terranova, 2017). It is, therefore, since the publication of Building a Global Community, that the significance of Facebook Community Standards has increased, becoming one of the central governing techniques to influence the activity of the users (Gillespie, 2018). The importance of the Community Standards for the present case is that it is where hate speech policies are nested.

6.4.3 Fundamental Equality

Article 4 of Facebook Principles is Fundamental Equality, which is defined as follows:

> Every person – whether individual, advertiser, developer, organization, or other entity – should have representation and access to distribution and information within the Facebook Service, regardless of the person's primary activity. There should be a single set of principles, rights, and responsibilities that should apply to all people using the Facebook Service. (Facebook Principles, 2019)

Due to Article 4 on "Fundamental Equality", all user profiles are strictly the same. For example, a page of a public institution and a personal profile with a handful of friends are treated in the exact same manner. Since there are no criteria for differentiating between users, Facebook applies an arithmetic understanding of equality between units or profiles. Each profile or page is, for Facebook, a unit that operates under the same rules as other units. Therefore, this way of understanding equality loses ethical coherence to acquire arithmetic coherence and is reductionist in the substance of identity, since it does not take into account any difference between profiles (Siapera and Viejo Otero, 2021). Whether it is the profile of a public institution, a celebrity with an antiracist agenda, or an individual with a small group of connections, all identities are reduced to arithmetically equivalent units. Under fundamental equality, users are units with no added cultural values; and, in practical terms, the page/profile of an account that posts materials attacking others and the page/profile of an account that supports others are treated in the exact same manner:

> when you have two communities attacking each other and of course we never say that just because one community is more under attack than the other we should treat a single report differently, if any individual reports any piece of content and there is an attack on any of those particular categories, it doesn't matter who is perpetrating it, it will be removed. (Aibhinn Kelleher, Key Informant 2 A, 2017)

As such, Facebook, under the argument that it is not 'in the business of assessing which group has been disadvantaged or oppressed' (Zuckerberg, 2019), makes it clear that it does not distinguish between users in any way, walking away from historically oppressed groups, affected groups or minorities, and taking content down for all users, independently of what group the users belong to (Siapera and Viejo Otero, 2021). This approach to equality echoes the American neutral viewpoint reflected on Neutral Viewpoint as seen in Chapter Three. In light of this, the chapter discusses how Facebook understands freedom.

6.4.4 Freedom.

Facebook has always aimed to govern. Whether talking about freedom, equality, community, or security, it has always cultivated a vision. When it comes to principles of freedom, Zuckerberg supported a vision of a free and open internet (12 July 2017). However, during the period of writing this thesis, Facebook's vision has changed. At first, in line with the idea of a free and open internet, Facebook Principles sought to guarantee freedom of expression and freedom of information, which implied that Facebook would allow expression and the circulation of that expression, or content, with no limitation of any kind, similar to the freedom of expression of an absolutism approach. Facebook's approach to freedom was reflected in Facebook Principles Article 1 Freedom to Share and Connect, and on Article 3 Free Flow of Information:

Article 1. Freedom to Share and Connect

> People should have the freedom to share whatever information they want, in any medium and any format, and have the right to connect online with anyone – any person, organization or service – as long as they both consent to the connection.
>
> Article 3. Free Flow of Information
>
> People should have the freedom to access all of the information made available to them by others. People should also have practical tools that make it easy, quick, and efficient to share and access this information.

As observed, the two articles draw upon a notion of freedom that implies that users should be free to create all kinds of content, and that content should flow freely across the platform with no interference as the Californian Ideology envisioned. Problems, however, emerged when Facebook had to introduce measures to tackle harmful content. In particular, there are two significant interventions by Mark Zuckerberg that relate to how Facebook was altering its notion of freedom. The first comes from December 2018, when Zuckerberg announced that Facebook was 'altering its DNA to combat harm'; that is, Facebook was changing its Principles. Second, and particularly relevant, is Zuckerberg's keynote at Georgetown University Standing for Voice and Freedom of Expression in October 2019. This intervention is particularly significant, as Zuckerberg introduces the term 'Voice' as a synonym of freedom of expression, expressing that Facebook does not consider freedom of expression in absolute terms, but with limitations (Zuckerberg, 17 October 2019). Based on these, Facebook justifies the introduction of a new principle: safety.

As the table below illustrates, Facebook adapted or manipulated its relationship with freedom, specifically by eliminating Article 3 Free Flow of Information to introduce the Safety Article.

2009-2019 Version	2019 -2021 Version
Freedom to Share and Connect People should have the freedom to share whatever information they want, in any medium and any format, and have the right to connect online with anyone - any person, organization or service - as long as they both consent to the connection.	**'Give People a Voice'.** People deserve to be heard and to have a voice — even when that means defending the right of people we disagree with (Facebook Principle, 2021).
Free Flow of Information People should have the freedom to access all of the information made available to them by others. People should also have practical tools that make it easy, quick, and efficient to share and access this information.	**'Safety'** We have a responsibility to promote the best of what people can do together by keeping people safe and preventing harm (Facebook Principle, 2021).

Table 3 Evolution of freedom on Facebook Principles.

6.4.5. Safety

Facebook uses techniques dedicated to safety, whose ultimate aim is to protect the Principle of Voice. Commonly, safety is understood as 'the condition of not being in danger' (Cambridge Dictionary, 2021, unpaginated). In its principles, Facebook refers to its responsibility 'to promote the best of what people can do together by keeping people safe and preventing harm' (Facebook Principles, 2021). This, in essence, translates into 'Facebook is a safe place for users to use their voice without the fear of encountering attacks' (Cummiskey, 2019). As such, Facebook is not concerned necessarily with the well-being of the user, but with creating an environment so the user can feel safe to upload content. It is, therefore, possible to argue that Facebook has, technically, allowed freedom of expression but limited freedom of information. In light of this, the following section explores and explains the significance that Facebook has on freedom of voice to introduce safety.

6.4.6. Voice and Safety

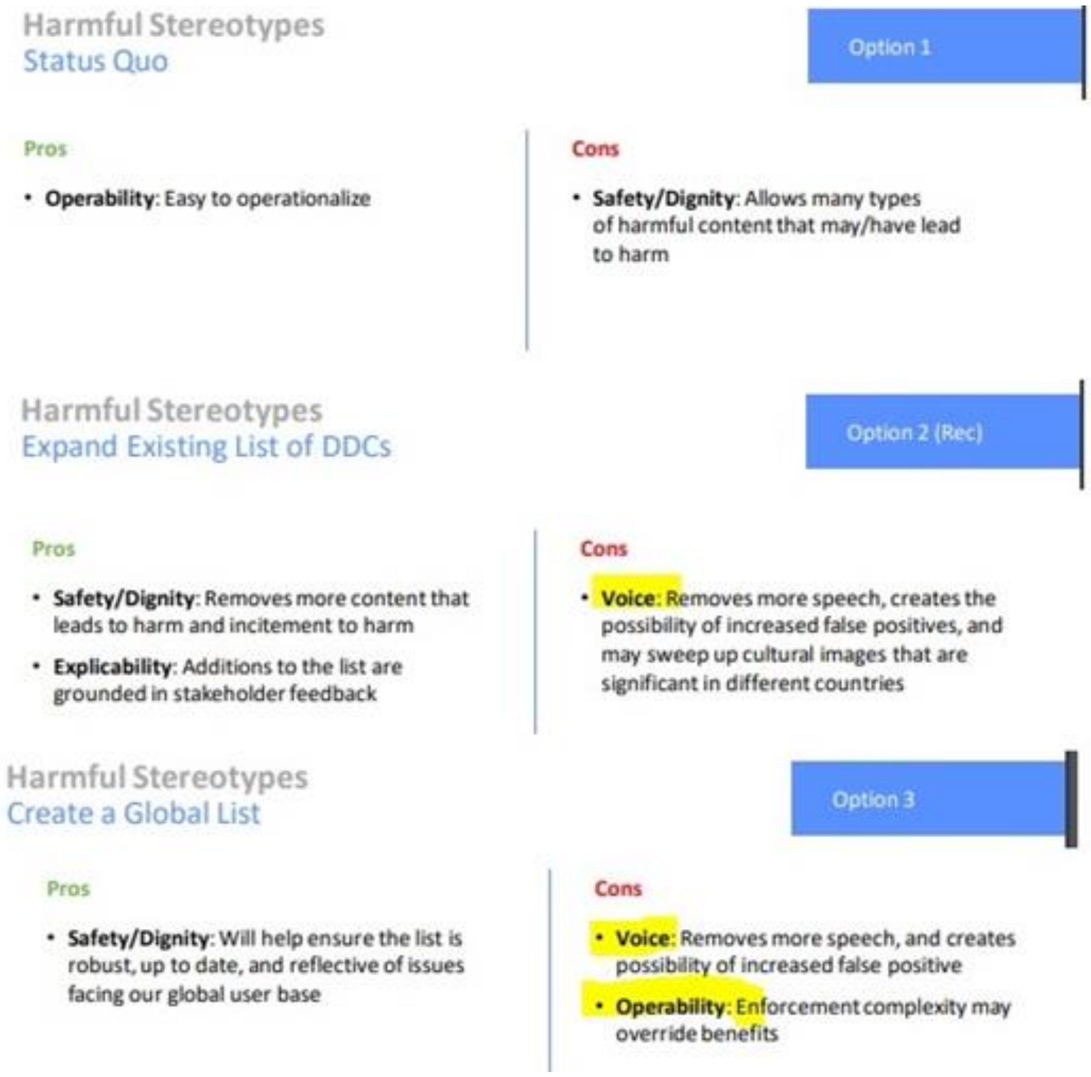
The relationship between Voice and Safety is a subject that requires close attention. Facebook's responsibility in keeping users away from harm, in combination with Facebook's need to protect and guarantee all users' voice, is understood on Facebook as a tension. This tension is such that it is not uncommon to find posts by Zuckerberg commenting on this struggle, as we have seen from the quote earlier: 'I've struggled with the tension between standing for free expression and the harm caused by minimizing or denying the horror of the Holocaust' (Zuckerberg, 2020). Underneath this tension, what lies is the struggle that Facebook encounters in its quest to provide freedom without limitations. As pointed out above, the Free Flow of Information was substituted in November 2019 by the mission of Safety, and Facebook's real struggle is to keep the principle of Voice intact. While Safety is understood as a mission to protect the user (Facebook Principles, 2021), following the logic of the key informant's comment that safety is there to safeguard voice, its mission, rather, is to protect the voice of the user. It is, therefore, possible to argue that the main activity of Safety is to protect Voice, which is both the main economic source of Facebook, and one of its principal ideological values, namely freedom of expression.

The protection of freedom of expression is 'not a foregone conclusion based on technological necessity, but a decision made for ideological and "business reasons"' (Roberts, 2019, p. 35). Voice guarantees the production of content and the extraction of data, which is the fundamental economic source for Facebook (van Dijck, 2013; Srnicek, 2017). Several industries have come to rely on users uploading content, for example advertising, intermediaries who analyse users' behaviours, objectionable content industry [8] and even moderation industry and jobs that depend on users uploading content. Importantly, the

[8] I refer to the objectionable content industry as a series of companies and organisations that are in charge of detecting Fake news, identifying hate speech content, and creating algorithms to harvest hate speech, i.e. Doubleverify, Newtral, applift.com, etc.

revenue source for a platform such as Facebook is founded on the principle of 'Voice' or the capacity for users to upload content.

The principle of Voice was first introduced by Zuckerberg in October 2019. It was defended by Zuckerberg at Georgetown University, where he made the point that this is the reason why the service exists: 'I've focused on building services to do two things: give people a voice, and bring people together. These two simple ideas -- voice and inclusion -- go hand in hand' (2017, unpaginated). An example of the role that the principle of Voice has in the Facebook decision-making process can be found below. Figure 26 presents a decision-making process around harmful racist stereotypes and the protection of particular groups. The image that the figure illustrates emerges from Facebook's Product Policy Forum minutes from 11 August 2020. The example illustrates a recommendation on harmful stereotypes and it involved 60 external engagements and 9 working groups. The recommendation was to 'Expand Designated Dehumanizing Comparisons (DDCs) list to add stereotypes based on evidence that they are linked to harm or are likely to incite harm' (Facebook, 2020). Designated dehumanising comparisons are mostly generalisations, or behavioural statements involving particular groups, generally minority groups, that perpetuate or create social subordination. At the time of writing, DDCs are included in Community Standards Tier 1 and are considered priority content to take down. The particular subject matter of this working group was to decide between a) leaving the content standards as they are, b) expanding the list of DDCs, or c) creating a global list.



Harmful Stereotypes
Status Quo

Option 1

Pros

- **Operability**: Easy to operationalize

Cons

- **Safety/Dignity**: Allows many types of harmful content that may/have lead to harm

Harmful Stereotypes
Expand Existing List of DDCs

Option 2 (Rec)

Pros

- **Safety/Dignity**: Removes more content that leads to harm and incitement to harm
- **Explicability**: Additions to the list are grounded in stakeholder feedback

Cons

- **Voice**: Removes more speech, creates the possibility of increased false positives, and may sweep up cultural images that are significant in different countries

Harmful Stereotypes
Create a Global List

Option 3

Pros

- **Safety/Dignity**: Will help ensure the list is robust, up to date, and reflective of issues facing our global user base

Cons

- **Voice**: Removes more speech, and creates possibility of increased false positive
- **Operability**: Enforcement complexity may override benefits

Figure 26 Illustrates the principles that Policy makers use to decide over harmful stereotypes

The particular observation from Figure 26 and that this section was drawn upon is the fact that Voice is used as a pre-established reason that fundamentally influences decision-making. In particular, I would like to draw attention to Option 2 of this example, where the pros, or reasons to establish more protected categories, are provided by external stakeholders, while the cons, or reasons against expanding this list, are provided by Facebook, listing limitations to Voice as a key drawback. As the example illustrates, the use of Facebook's Principle of Voice has a primary place in decision-making. Even when it comes to providing safety and dignity to minority groups, Facebook shows that its tendency is to return to its own principle of Voice – or freedom of expression.

In sum, Facebook Principles have evolved throughout the years. Whereas the value that Facebook gives to Equality has never changed, the value that Facebook gives to Freedom has changed by specifically protecting freedom of expression or voice while, at the same time, limiting how content circulates across the platform, justified as a security matter. In considering this, the following sections explore how these principles are embedded in the different techniques employed by Facebook. First, the section will look at what principles influence policy officers, content standards and, within content standards, the specifics of hate speech policy. In addition, the section will also include results obtained through location analysis, especially results obtained through the analysis of the location of the community and content standards and the symbolic meaning of the elements that compose the community standards.

6.5 **Facebook Principles and Policy makers.**

According to our Key Informant 1, Siobhan Cummiskey, the Policy team operates across three principles, namely 'safety, voice and equity' (2019). ', those three pillars constantly inform our community standards' discussion, from when we create policies to when we refine them, etc. Those three policies are the ones that inform our decisions' (Key Informant 1 Cummiskey, 2019)

For Cummiskey, the three values are interdependent, as 'in order to have a voice on Facebook you have to feel safe. So we make sure FB is a safe place so everyone can use it and everyone can have a voice, and then safety voice' (2019). This implies that Facebook does not consider Voice or Freedom of Expression as an unqualified right for the user, but has to be exercised in a manner consistent with the rights of voice of others.

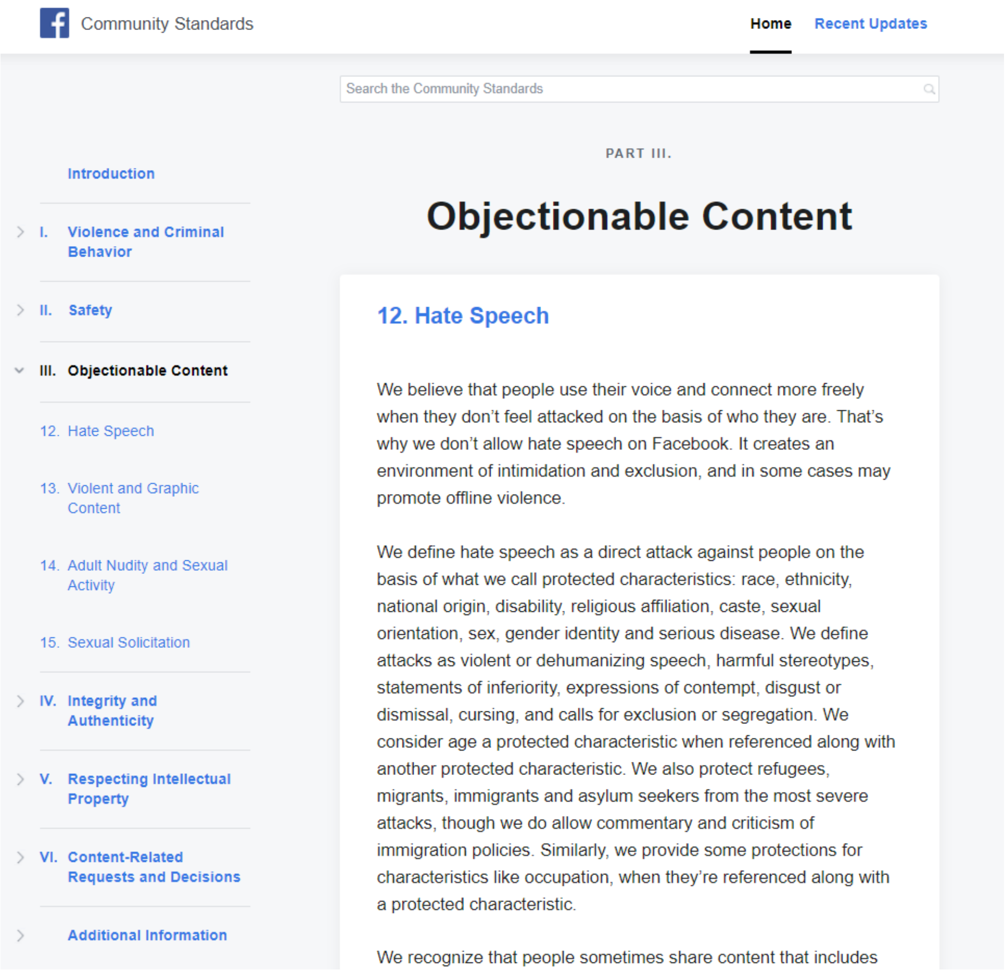
For Facebook policy makers, equity means that all users should be ruled under the same set of rules: 'the same decision on the same people on the same piece of content, no matter who has posted or who you are, we making the same and fair decision based on what we see and based on our policies, so equity is very important' (Cummiskey, 2019). That is, as Fundamental Equality or Article 4 indicates, 'There should be a single set of principles, rights, and responsibilities that should apply to all people using the Facebook Service' (Facebook Principles, 2019). Implications of this notion of equity are that: a) Equity on Facebook is tangled with Facebook principle and notion of connectivity and community, b) for policy makers, it implies that there is one single team in charge of designing policies, and those policies affect all users, regardless of their jurisdiction or geographical location.

Therefore, in considering that policy makers are influenced by Facebook Principles, and in considering that policy makers are in charge of designing Facebook Community Standards, the following section looks at how Facebook Principles are reflected in Facebook Community Standards.

6.6 Facebook Principles and the Community and Content Standards.

Policy operationalised all principles into policies, and all Facebook Policies are organised under Community Standards. Community Standards materialised Facebook Principles, particularly those in relation to Community and Equality. Community as the community standards regulates all Facebook users and equality, because it regulates all users equally.

Facebook's Hate Speech definition is nested within the Community Standards, under the specifics of Section III, 'Objectionable Content' (Facebook, 2021).


Community Standards
Home
Recent Updates
Search the Community Standards
Introduction
I. Violence and Criminal Behavior
II. Safety
III. Objectionable Content
12. Hate Speech
13. Violent and Graphic Content
14. Adult Nudity and Sexual Activity
15. Sexual Solicitation
IV. Integrity and Authenticity
V. Respecting Intellectual Property
VI. Content-Related Requests and Decisions
Additional Information

PART III.

Objectionable Content

12. Hate Speech

We believe that people use their voice and connect more freely when they don't feel attacked on the basis of who they are. That's why we don't allow hate speech on Facebook. It creates an environment of intimidation and exclusion, and in some cases may promote offline violence.

We define hate speech as a direct attack against people on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing, and calls for exclusion or segregation. We consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants and asylum seekers from the most severe attacks, though we do allow commentary and criticism of immigration policies. Similarly, we provide some protections for characteristics like occupation, when they're referenced along with a protected characteristic.

We recognize that people sometimes share content that includes

Figure 27 Illustrates the location of hate speech policies. They are under the Objectionable content section on the Facebook site

The word 'Objectionable' is defined as things that people dislike or oppose because they are unpleasant or wrong (Cambridge Dictionary, 2020). Objectionable is synonymous with contentious or controversial and it is, therefore, subject to scrutiny and further analysis (van Berkum et al., 2009). Facebook has adopted the term objectionable and popularised it, especially in automatic detection tools. Fundamentally, Facebook is developing algorithms to harvest, filter, or regulate large volumes of content that fall into diverse categories of a different nature, but that, together, constitute a problem for Facebook, such as false, deceptive, fraudulent, sexually explicit content, but also racist, homophobic, and misogynist content. The question is, why does Facebook consider hate speech objectionable? In the Community Standards, underneath section 12 Hate Speech, Facebook offers an active link that reads 'learn more about our approach to hate speech'. This link redirects the user to an article written by Richard Allan, Vice President of Public Policy, in 2017. This document reveals Facebook's set of arguments to explain its approach to hate speech. In the first place, the text presents Facebook as a platform 'for all ideas' (Allan, 2017), meaning all political ideas and beliefs from all individuals around the globe. As mentioned earlier, Facebook as a platform for 'all ideas' first appeared on Zuckerberg's post in February 2017, after the criticisms raised from the consequence of the US elections, and repeated on several occasions in 2017 (Zuckerberg, 27 September 2017, 12 July 2017). The notion of 'all ideas' permeates Facebook policy officers (Allan, 2017; Key Informant 2, 2017), and implies that all ideas are valid in relation to freedom of expression. Therefore, all ideas – aka content – are a priori valid. If all ideas are a priori valid, then this means that no ideas can be immediately excluded. Rather, they have to be checked one by one to examine if they may be in any way problematic or 'objectionable'. This idea is reinforced by Allan, when he argues that there 'is no universally accepted answer for when something crosses the line' (2017, unpaginated).

Since Facebook does not elaborate on its decision to classify hate speech as objectionable, we can turn our attention to the consequences of this decision. Despite the difference in nature of all these types of content, Facebook understands fraudulent content under the same category as racist content, neglecting therefore the entirety of all the body of work and theory around what racism is, the effects of racism in terms of creating inequalities, and its implications in society, ultimately leading to racism becoming 'debatable' (Titley, 2020). Titley defines debatability as the constant questioning of what is and what is not racism and who gets to define it. By constantly debating if certain content is racist, these debatability techniques contribute to the multiplying and perpetuating of oppressive narratives. In sum, this analysis fundamentally reveals that the notion of Equity influences the community standards by adopting a neutral viewpoint towards hate speech that translates into the word ‘objectionable’. In considering this, the following section looks into the specifics of how Facebook phrases, values and understands hate speech policy.

6.7 Facebook's Principles and Hate Speech policy

On 12 August 2017, in Charlottesville, white supremacist James Alex Fields Jr. deliberately drove his car into a crowd of people who had been peacefully assembling to protect against the Unite the Right rally, a right-wing supremacist group. The attack killed Heather Heyer and injured 28 other people. Five days later, Zuckerberg published a post stating that both himself and Facebook opposed hate. The importance of this post is that it reveals how Zuckerberg conceptualised hate, referring to it in behavioural terms instead of a structural problem:

> We aren't born hating each other. We aren't born with such extreme views. We may not be able to solve every problem, but we all have a responsibility to do what we can [...] [Facebook is a] platform for all ideas where people with different views can share their ideas. Debate is part of a healthy society… But when someone tries to silence others or attacks them based on who they are or what they believe, that hurts us all and is unacceptable. (Zuckerberg, 16 August 2016)

As can be noticed, in the above quote the notion of oppression fades or tends to vanish and, instead, hate is referred to as an individual problem of opinion, that can be ‘cured’ by talking and debating with those who oppose our own opinions. This is not a novelty. If we look, for example, at the specifics of racism, the individualist approach is manifested in a range of political positions, from the political discourses of UN delegates to national public institutions that, instead of addressing hatred as a structural phenomenon, explain racist hatred in pathological terms (Lentin, 2008). In these positions, racism is seen as an individual problem, associated with fear of the other and the protection of what is ‘ours’ and which could, therefore, be solved by increasing our knowledge of ‘other cultures’ (Lentin, 2008, p. 11). In line with Lentin's observation, Zuckerberg proposes that, in order to overcome the problem of ‘extreme views’, we should encounter them on the same platform. In the same quote, Zuckerberg then proceeds by emphasising that what is unacceptable is the 'act of attacking directly', in which case Facebook will take content down.

In considering all the above, the following paragraphs aim to analyse Facebook's official definition of hate speech, and aim to understand to what extent Zuckerberg's notion of hate and Facebook Principles are embedded in Facebook's official definition of hate speech:

In 2021, Facebook defines hate speech as follows:

> We define Hate Speech as a direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability. We also provide some protections for immigration status. We define 'attack' as violent or dehumanising speech, statements of inferiority, or calls for exclusion or segregation. (Facebook Hate Speech Definition. Last retrieved April 2021)

As the definition shows, Facebook does not use specific minority groups on its general definition of hate speech, but generic categories such as race, religious affiliation or gender, instead of black, Muslims or women. Also, there is emphasis on the definition of the word 'attack', but does not refer to expressions of hate that take forms of love (Back, 2002), e.g. I love my country pure and pristine. Neither considers problematic words accepted in society that are inherently racist, e.g. patriotism (see Siapera and Viejo Otero, 2015).

In considering this, it is clear that Facebook's definition of hate speech is influenced by Facebook's understanding of equity as it holds a neutral viewpoint of hate speech, and by its mission or principle of security, as it looks for expressions that 'attack' to be taken down. Indeed, this approach does not differ from the Neutral Viewpoint as seen in Chapter Three. It is, therefore, possible to argue that Facebook does not innovate on its approach to govern hate speech, rather it reproduces old ideas in the new digital environment.

The last question that remains is: what are the particular means by which Facebook enforces hate speech policy? How does neutrality manifest? How does it provide safety? How does it protect the voice? All these questions are tackled in the following chapter.

6.8 Conclusion

This chapter has explored the Facebook Principles, and their influence and relationship with different techniques of governance such as Mark Zuckerberg's posts, policy makers, and Community Standards. The chapter points out that the two main techniques that orchestrated Facebook's general direction are Facebook Principles and Mark Zuckerberg's posts. Together, these techniques confirmed what this dissertation refers to as a Dispositive of Ideology. It constitutes a dispositive because, as shown in this chapter, Mark Zuckerberg informs Facebook Principles, and Facebook Principles is a hierarchically superior document that represents the general direction of the platform. In this document are the Principles that shape hate speech, namely voice, fundamental equality and security, and the value that Facebook gives to these principles, where voice determines prevalence for freedom of expression, equity determines a neutral viewpoint, and security determines how hate speech policies are enforced.

Also, the chapter outlines the fundamental principles that Facebook has created, or re-created, and that shape its hate speech policy, being: Connectivity, Community, Equity, Voice and Safety. Within the Facebook site, all technological affordances, Zuckerberg's rhetoric, principles and governing documents are intertwined. Technological affordances cannot be separated from rhetoric, or from policies. Instead, we have to look at how they cooperate with and justify each other. In considering this, we can conclude that connectivity is facilitated by a series of technological affordances, which understand that all users are capable of meeting at the same time in the same place. In other words, technological affordances on social media do not reflect technologies and technological capabilities, but the ideology behind the platforms (Feenberg, 2017). This notion of connectivity has led Facebook to design centralised techniques of governance that Zuckerberg justifies as the need for a Facebook Community. As the analysis of Building a Global Community has illustrated, the central document that governs the user's behaviour is the Community Standards (Gillespie, 2018). However, whereas Community Standards is considered by Gillespie to be at the centre of Facebook operations, the findings of this chapter point out that Community Standards cannot be understood without the values and the visions that justify their existence – such as Article 10, One World or Fundamental Equality.

In the particular case of hate speech policies, this chapter shows evidence that Facebook's key informants define hate speech under notions of Fundamental Equality, as defined in Facebook Principles. Indeed, the analysis of Facebook's Principle of Fundamental Equality reveals that Facebook understands Equality as a notion that does not make distinctions among individuals, nor does it recognise minorities or groups. Rather, it punishes the act of discriminating. In returning to Chapter Three, and comparing previous approaches to hate speech regulation with Facebook's approach, I argue that Facebook has adopted a Neutral Viewpoint as regards hate speech regulation, by which Facebook does not take the side of any historically oppressed group. Instead, its approach is based on the 'act wrong' idea (Altman, 1993), by which Facebook takes content down when it constitutes a threat or aims to harm. That is, when the issue is safety. Finally, and considering that the Neutral Viewpoint was historically an approach designed to protect freedom of expression, the chapter analysed Facebook's notion of freedom. The analysis reveals that Facebook's notion of freedom has varied, as reflected through Facebook Principles. While, at first, Facebook embraced the idea of freedom to upload content and freedom of content to flow freely around the platform, its second version shows that safety replaces freedom of flow of information, indicating that a relatively new safety apparatus has been built to control content flows across the platform.

In particular, the following chapter aims to analyse how Facebook operationalised hate speech, which I argue is ruled by principles of freedom (Voice) and safety principles, presenting a notable parallelism with Foucault's notion of discipline and the notion of the dispositive of security (2007).

Chapter 7. Facebook and the Management of Content

7.1 Introduction

> '…that the establishment of apparatus of security to ensure the free circulation of grain *(content)* was not only a better source of profit, but also a much better mechanism of security against the scourge of scarcity'. (Michael Foucault 2010, 34. Parenthesis added)

Chapter Seven is the second empirical chapter of this dissertation and aims to respond to the following research questions: What techniques have Facebook developed to enforce its policies and what values underpin Facebook's operating system?

In responding to these questions, the chapter delves on the ways in which the operational system that Facebook has developed approaches the question of hate speech and what implications this may have.

Broadly speaking, content operation refers to the set of strategies and techniques used to detect, analyse, and categorise content and this chapter describes one by one how these function. In addition, this chapter further explain the relationship among techniques and how further they conform the apparatus of governance. Specifically, the chapter delves on Facebook dispositive of discipline and Facebook dispositive of security. Both dependant of dispositive of security as shown on chapter six and as the following image illustrate.

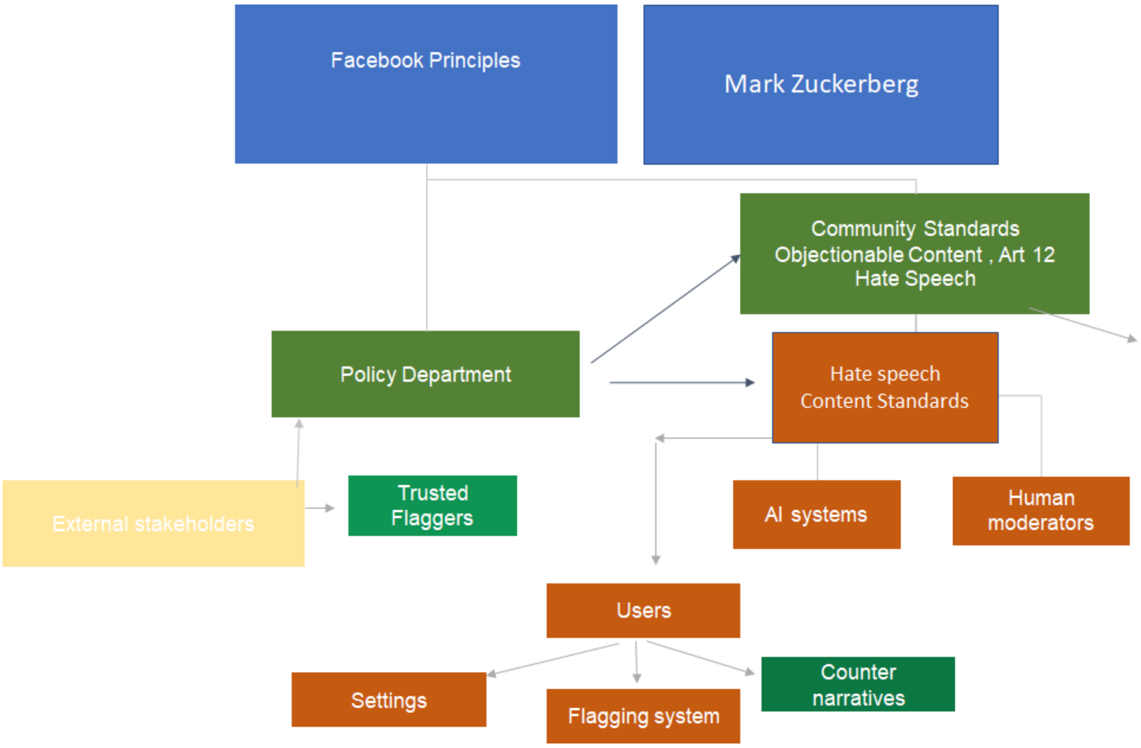


Figure 28 Facebook Governance apparatus according to techniques and dispositive. The ideological dispositive is in blue, the disciplinary dispositive is in green, and the security dispositive is in red

In order to explore the above, the chapter is organized as follows. First, the chapter argues that policy makers have a disciplinary role. Among all the tasks this department carries out, this chapter focuses on the work carried out on content standards. Content standards are a specific part of the Community Standards and relate to the specific forms of language and expression that are not allowed on the platform. The first section looks into how content standards are decided, the teams involved in the decision making process, and the role of external stakeholders. Here, the chapter argues that, while there is a well-structured process in place to decide content standards, evidence shows Facebook has taken top down decisions bypassing these processes.

Section two analyses in depth the series of techniques that Facebook has implemented to maximise user safety. The section analyses user's settings, such as privacy, visibility, Newsfeed relationships, and the flagging/reporting system. Next, the section analyses artificial intelligence, human reviewers, and Facebook's oversight board. The analysis of each of these techniques separately and the analysis of how they all relate to each other, reveals that their task is to, protect the principle of Voice, reduce the volume of hateful content, and maximise the possibilities for content to be on the platform. In light of this findings the four section argues that Facebook techniques to provide safety constitutes Facebook's 'dipositive of security' (Foucault 2007, p.379) The section concludes by arguing that how Facebook operationalised hate speech does not alter the conditions that allow for hateful content to be on the platform. In fact, the opposite arrangement is in place, that Facebook's apparatus of security is designed under the assumption that there will always be hateful content on the platform.

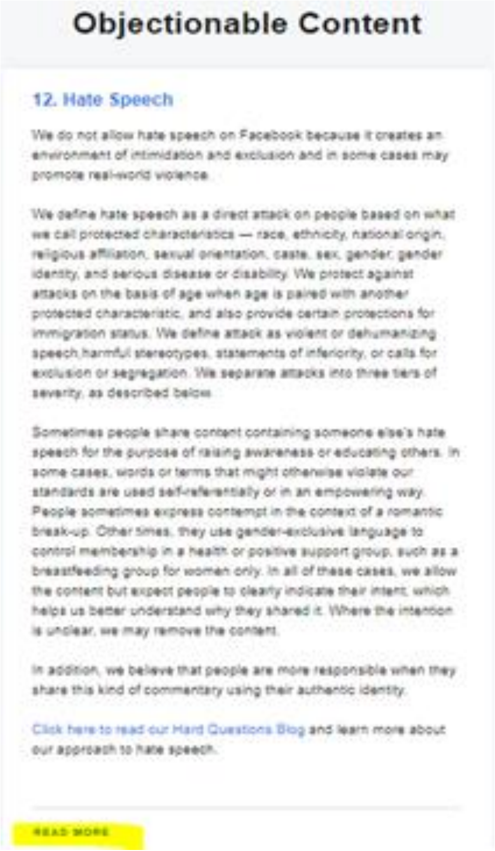
Overall, Chapter Seven shows how Facebook operates its policies and its principles and it does so by showing the different set of techniques that Facebook uses, the relationship between techniques (dispositives), and the principles that underpin each techniques and dispositive which as showing on chapter six, is the principle of Voice. The chapter argues that the mechanisms and systems to operationalise hateful content is a dispositive of security whose function is specifically dedicated to constantly delete content from the platform, not to alter the relations that favour discrimination.

7.2 Facebook's dispositive of discipline

According to Foucault (2007), discipline belongs to the order of organization. Discipline, he argues, works in an empty, artificial space that is to be completely constructed. The aim of discipline is simply a matter of maximizing the positive elements, for which one provides the best possible circulation, and of minimizing what is risky and inconvenient (Foucault, 2007, p. 17). In considering this, a dispositive of discipline on Facebook would refer to a series of interrelated mechanisms or created techniques that have the quality or function of directing, controlling, and disposing of something that have as aim to allow content to circulate with minimising risks. To a great extent, Facebook's Policy Department duty is optimizing how the platform functions, and minimise the risks that can emerge and to ensure that all Facebook Principles are implemented, which according to main Policy Department key informant, and as seen in Chapter Six, there are three Principles: 'Voice, equity and safety' (Cummiskey 2019).

Indeed, Facebook is a platform orientated to all forms of content (Waller and Anderson, 2019), meaning that the users are not restricted to talk about one specific topic. Facebook feed is open for all forms of content. This makes Facebook a constant producer of information, with an endless number of industries refining and using content for commercial purposes (Srnicek, 2017). As such, the mission of Facebook's policy officers is to ensure that this content remains on the platform, although with limitations, which are specified in the Content Standards.

According to a key informant, content standards changes are reviewed every two weeks (Cummiskey, 2019), or any time a trigger point (i.e. scandal, critique, event, impetus) puts pressure on Facebook. Facebook also takes into consideration issues raised by their internal research team. The meetings are specifically dedicated to review 'Content Standards' (2019), which refer to the specific expressions and kinds of content that are not allowed on the platform and which differ from policy definitions (i.e. what constitutes hate speech), as shown in the previous chapter. Content standards can be found under the 'read more' section of community standards, as Figure 29 illustrates.



Objectionable Content

12. Hate Speech

We do not allow hate speech on Facebook because it creates an environment of intimidation and exclusion and in some cases may promote real-world violence.

We define hate speech as a direct attack on people based on what we call protected characteristics — race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability. We protect against attacks on the basis of age when age is paired with another protected characteristic, and also provide certain protections for immigration status. We define attack as violent or dehumanizing speech,harmful stereotypes, statements of inferiority, or calls for exclusion or segregation. We separate attacks into three tiers of severity, as described below.

Sometimes people share content containing someone else's hate speech for the purpose of raising awareness or educating others. In some cases, words or terms that might otherwise violate our standards are used self-referentially or in an empowering way. People sometimes express contempt in the context of a romantic break-up. Other times, they use gender-exclusive language to control membership in a health or positive support group, such as a breastfeeding group for women only. In all of these cases, we allow the content but expect people to clearly indicate their intent, which helps us better understand why they shared it. Where the intention is unclear, we may remove the content.

In addition, we believe that people are more responsible when they share this kind of commentary using their authentic identity.

Click here to read our Hard Questions Blog and learn more about our approach to hate speech.

Figure 29 Illustrates where Content Standards are located in the site

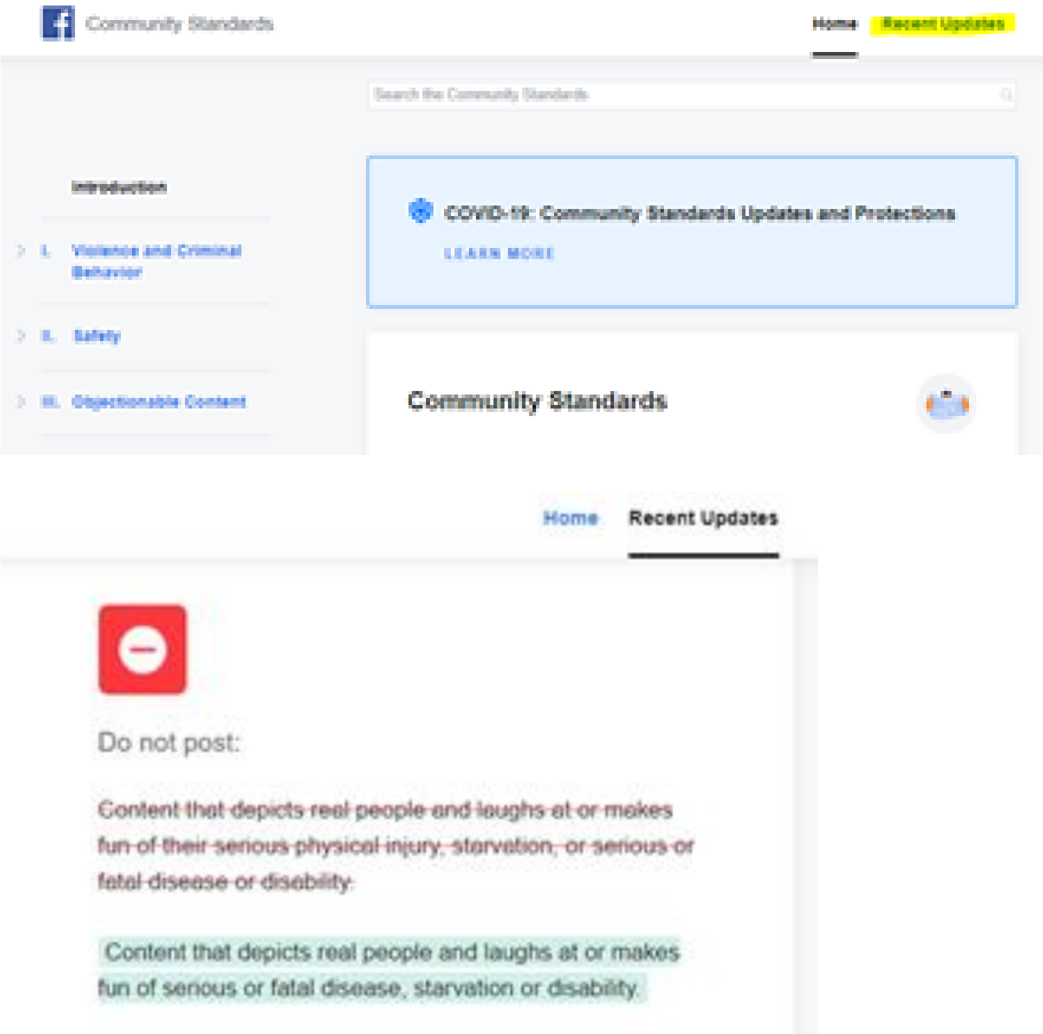
By clicking *'read more', Facebook* gives access to the specifics of what kind of content is not allowed on the platform, with clear examples such as content in which transgender or non-binary people are referred to as 'it', for example (Facebook Content Standards, 2021). Content standards are organised on three levels or tiers, which refer to levels of severity. Tier one, the most severe in terms of control, involves calls to violence or dehumanizing speech against other people based on their race, ethnicity, nationality, gender, or other protected characteristics (Facebook, 2021). Tier one includes the prohibition of specific expressions that dehumanise groups, such as comparisons of Black people with apes, Jews with rats, Muslims with pigs, or women with objects. This kind of categorisation of language that Facebook performs, echoes Matsuda argument by which hate speech regulation should define and limit specific forms of language that perpetuate subordination (1993). While this is a step forward toward the protection of minorities, it is important to consider as argued in chapter six that firstly, Facebook does not limit their prohibition to those particular groups, but to all groups subject of attack, giving therefore the case of what Delgado refers as ‘reverse enforcement’ (Delgado 1993, 2018) as argued in chapter three, secondly, that Content Standards change rapidly and are revised constantly, and therefore do not constitute a political statement on behalf of Facebook. Rather, they are operational formulas that are not fixed or determined politically or ideologically.

Tier two is the second term of severity and refers to inferiority or stereotyping statements. They are subject to more scrutiny, as such expressions may be a threat, but also a 'legitimate' expression by the user, as Facebook argues.

Finally, Tier three covers 'everything that excludes or segregates'. While these calls for exclusion can be considered free expression, they are part of Tier three because such content can potentially escalate towards more severe threats.

Content Standards meetings are attended by all staff members ‘all around the world’ (Cummiskey 2019). These meetings are referred to as the Product Policy Forum. Announcements are made by team members who are experts in a particular field and announce the need of changes or updates in particular content areas i.e. hate speech content or child protection policies content. Announcements are followed by the organisation of a working group which involves ‘all people across the company from engineering, to legal, to policy who engage with the particular working groups for ‘several weeks’ (Cummiskey, 2019). As a result, working groups will present recommendations that, once approved, become content standards. The condition is that all decisions taken in these meetings are ‘operational’ (Cummiskey, 2019), meaning that all recommendations are enforceable across Facebook and can be subject to mass processing by automatic systems and by an organised human review team.

Once recommendations are decided and approved, amendments or changes in content standards are then made public and can be found on Community Standards, under the section ‘recent updates’, which leads to Content Standards and shows what the exact changes in community content policies are that the team have introduced (see figure 30).


Community Standards
Home Recent Updates
Search the Community Standards
Introduction
I. Violence and Criminal Behavior
II. Safety
III. Objectionable Content
COVID-19: Community Standards Updates and Protections
LEARN MORE
Community Standards

Home Recent Updates

Do not post:

~~Content that depicts real people and laughs at or makes fun of their serious physical injury, starvation, or serious or fatal disease or disability~~

Content that depicts real people and laughs at or makes fun of serious or fatal disease, starvation or disability.

Figure 30 Illustrates the location on Facebook to access ‘recent updates’

Part of the task of the working groups involves engaging with external expertise for consultation purposes. As Cummiskey points out, Facebook relies on a network of experts gradually established through the years and an internal team specially dedicated to outreach experts (2019) known as ‘Stakeholder Engagement Associate’, regularly announced on Facebook Careers Page. Experts are, both people who represent or focus in areas such as anti-discrimination, human rights or anti-racism and subject experts, or members of academia. Facebook criteria to select their stakeholders are inclusiveness or diversity and expertise considered 'top subject matter experts' (Cummiskey, 2019).

Facebook communicates with stakeholders mostly through one to one informal conversations in person or video-conference, and they aim, not only to exchange information, but to build up relationships. By establishing these relationships, Facebook ultimately seeks to identify what the most persuasive and instructive views are, not necessarily with the aim of reconciling them, but to understand and obtain the full range of opinions regarding new Content Standard proposals (Siapera and Viejo Otero, 2021b). Even when the tone Facebook uses with stakeholders is casual and one to one conversation remains 'candid', it is evident that on Facebook, the systems and procedures to introduce the voice of experts on content standard decisions represent a well-established procedure embedded in their daily operations. This is supported by the fact that Facebook has included 'stakeholder engagement' within Facebook Community Standards, as below Figure 31 illustrates. Since stakeholder engagement is embedded within the text of Community Standards, it is considered an important process for Facebook. Reading the text under external stakeholders carefully shows that the role of this engagement is to justify and strengthen the decisions taken by Facebook. It represents a consultation process, but has no binding influence on Facebook. As such, its role is to legitimise and offer a patina of validity over Facebook's decisions on content.

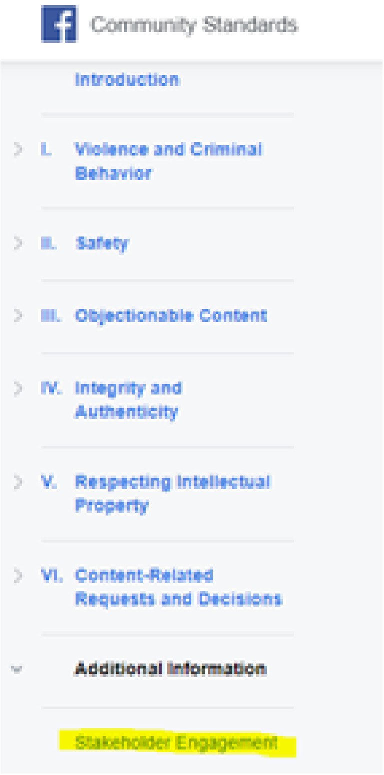

Community Standards
Introduction
I. Violence and Criminal Behavior
II. Safety
III. Objectionable Content
IV. Integrity and Authenticity
V. Respecting Intellectual Property
VI. Content-Related Requests and Decisions
Additional Information
Stakeholder Engagement

Figure 31 Illustrates where to locate experts within the community standards

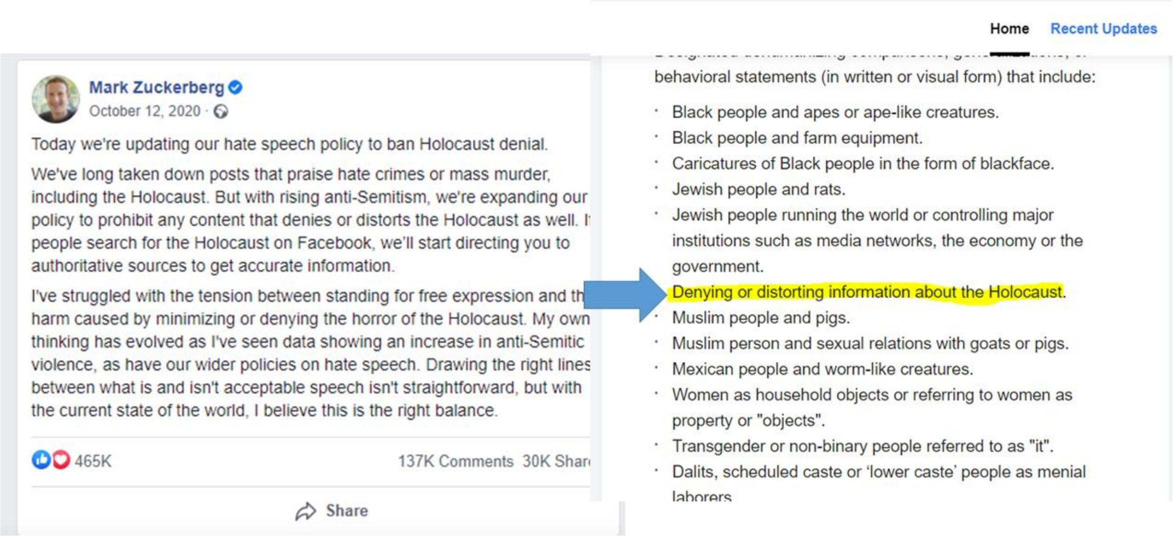
7.2.1 Top down decision making. The case of Holocaust denial

While it is clear the Facebook has gradually designed a process that brings external views to the table, it is also noticeable that this process can be altered by top down exceptions such as the Holocaust denial prohibition introduced on November 2020 by Mark Zuckerberg. On the 12 Of October of 2020, Mark Zuckerberg posted on his wall announcing that Facebook will ban Holocaust denial. As the case illustrates, the decisions did not follow the above mentioned standards and procedures. Instead, this policy decision revolved around Zuckerberg's personal struggles and decisions in relation to Holocaust denial and anti-Semitic content, as the following quote indicates.

> (…) with rising anti-Semitism, we're expanding our policy to prohibit any content that denies or distorts the Holocaust as well. If people search for the Holocaust on Facebook, we'll start directing you to authoritative sources to get accurate information. I've struggled with the tension between standing for free expression and the harm caused by minimizing or denying the horror of the Holocaust. **My own thinking has evolved as I've seen data** showing an increase in anti-Semitic violence, as have our wider policies on hate speech.

> Drawing the right lines between what is and isn't acceptable speech isn't straightforward, but with the current state of the world, I believe this is the right balance. (Emphasis added)
>
> (Zuckerberg, 12 October 2020)

Zuckerberg's post was immediately followed by an update posted on 'Recent Updates' of Facebook Content Standards (Figure 32)



Mark Zuckerberg

October 12, 2020 ·

Today we're updating our hate speech policy to ban Holocaust denial.

We've long taken down posts that praise hate crimes or mass murder, including the Holocaust. But with rising anti-Semitism, we're expanding our policy to prohibit any content that denies or distorts the Holocaust as well. I people search for the Holocaust on Facebook, we'll start directing you to authoritative sources to get accurate information.

I've struggled with the tension between standing for free expression and th harm caused by minimizing or denying the horror of the Holocaust. My own thinking has evolved as I've seen data showing an increase in anti-Semitic violence, as have our wider policies on hate speech. Drawing the right lines between what is and isn't acceptable speech isn't straightforward, but with the current state of the world, I believe this is the right balance.

465K 137K Comments 30K Shar

behavioral statements (in written or visual form) that include:

- Black people and apes or ape-like creatures.
- Black people and farm equipment.
- Caricatures of Black people in the form of blackface.
- Jewish people and rats.
- Jewish people running the world or controlling major institutions such as media networks, the economy or the government.
- Denying or distorting information about the Holocaust.
- Muslim people and pigs.
- Muslim person and sexual relations with goats or pigs.
- Mexican people and worm-like creatures.
- Women as household objects or referring to women as property or "objects".
- Transgender or non-binary people referred to as "it".
- Dalits, scheduled caste or 'lower caste' people as menial laborers

Figure 32 Illustrates How Zuckerberg's (12 October 2020) decision on banning Holocaust Denial was included on Recent Updates sections on Community Content Standards.

Facebook argues that policy decisions must be relevant, transparent, and legitimate. The stakeholder engagement process is part of this process of legitimacy. However, we can also see that the role of external stakeholders is part of a governmental technique associated with neoliberalism, which according to Miller and Rose (2008) has an 'intrinsic relation to the authority of expertise'. In this manner, through building relationships with experts (Miller and Rose, 2008), and/or in the process of setting up a neoliberal bureaucracy (Hibou, 2015), Facebook separates itself from authoritative forms of governing, however, in considering the procedures, and in considering the exception of Holocaust denial, it is possible to argue that whereas stakeholder expertise influences decision making on content standards, it does not yet substitute Facebook's central authority. Therefore, if policy officers do not have the power to substitute central authority, their role is mostly disciplinarian, maximizing the positive elements, for which they provide the best technique to allow the circulation of content, minimizing what is risky while still protecting Facebook principles.

In sum Facebook has put in place means to detect potential hateful content, and to analyse it and categorise it as acceptable or unacceptable. Facebook's operational activity is handled by policy officers and their task specifically is to define the content standards and to instruct users and human reviewers on what content is acceptable and what content is not (Cummiskey, 2019). To know what constitutes acceptable content, Policy officers engage with NGOs, academics, different experts in the field, as well as using feedback from users (Cummiskey, 2016). To enforce community standards, policy officers work with the product teams to design tools, design the guides that human reviewers will implement, and facilitate the work of the Oversight Board by selecting themes to be revised for further assessment. This thesis refers to all these activities and techniques as the Dispositive of Discipline. According to one of the key informants, all the work that policy officers carry out is inspired by three values: 'Voice, equity and safety' (Key Informant 1 2019). In light of this finding, it is possible to argue that Facebook's operational dispositive have been designed to discipline the space and to protect the principle of Voice while maximising the possibility of safety for the users, and it does so, not by limiting freedom of expression, but by limiting how information flows around the platform.

7.3 User settings

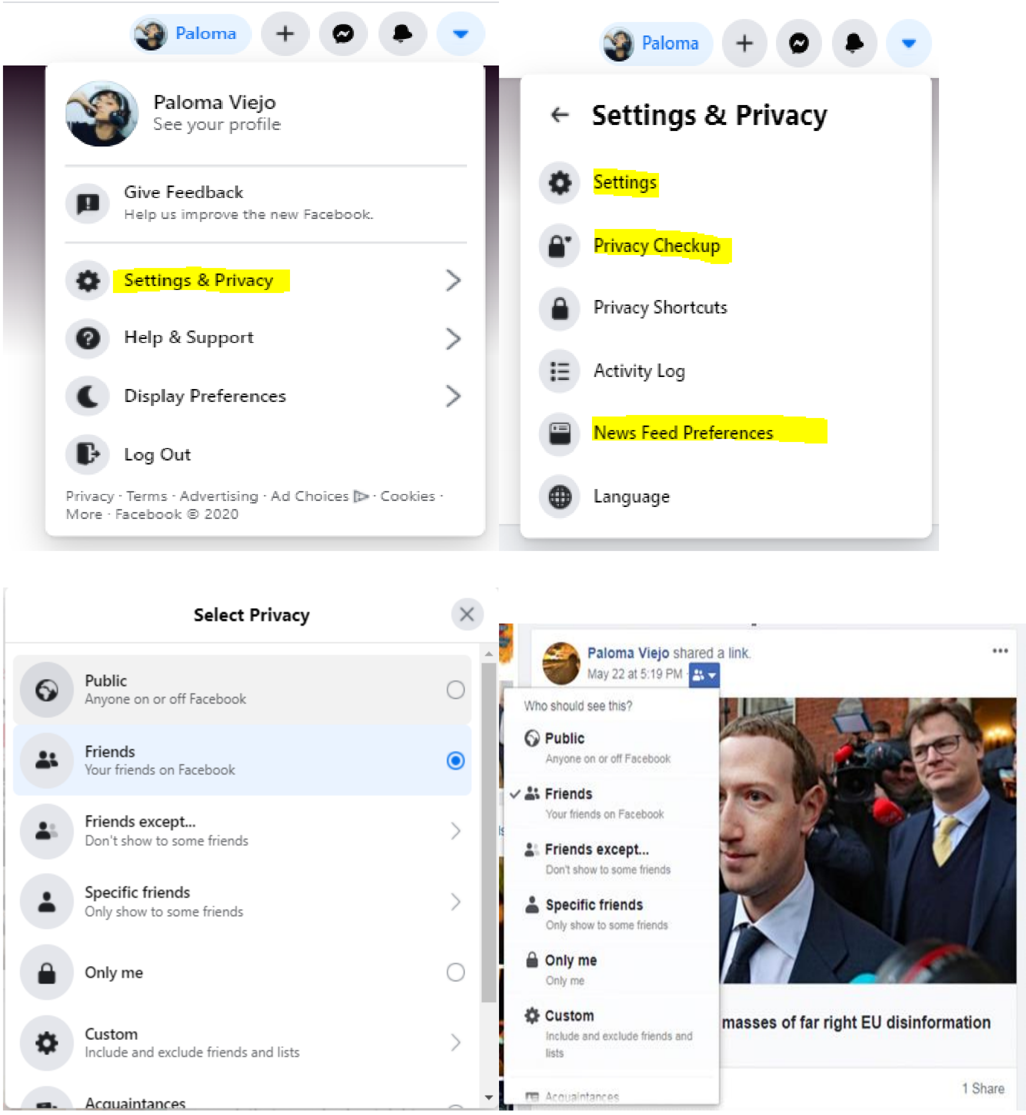
For a long time, Facebook users were at the centre of its decision making. Indeed, in the early stages of this research, the key informants refer to Facebook as a user centered company (Cartes, 2016; Cummiskey, 2016). However, the expansion of other forms and techniques of governance has gradually converted the user into a content provider, a flow of content controller, and a minor flagger or notifier of problematic content.

Facebook argues that users' freedom to express themselves on the platform is as important as user safety. The user is the key piece on Facebook because they are both consumers of the service and platform workers (Fuchs, 2014). In order to speak freely, the user must have the ability to enjoy the freedom to produce all kinds of content and, at the same time, they must have the guarantees that their freedom is not threatened or attacked. Therefore, the fundamental and more basic techniques created to control how information circulates on Facebook were directly given to the user, because, under Facebook's rationale, 'by giving people individual control, we can better balance our principles of free expression and safety for everyone' (Zuckerberg, 15th November 2018).

The ways Facebook has given control to the user have evolved from 2006 to 2021. Originally, if a user encountered an individual, a comment or considered that certain types of comments should not be present on Facebook, the company made a mailbox, 'Support Inbox', available to the user, which is still an active feature, by which the user could write to Facebook to express discontent regarding particular content (Cartes, 2016). Soon, the volume of Facebook users made Support Inbox unworkable. So Facebook began to establish a series of 'product solutions' (Cummiskey, 2016) or tools, by which the user, following their own discretion, could regulate what users they wanted to be in contact with, the content they wanted to be exposed to, and report what type of content they never want to be exposed to. These tools work simultaneously, they are settings of privacy and visibility, flagging report systems, and allowing for the customization of News Feed relationship databases.

7.3.1.1 Privacy and Visibility

To regulate content exposure, Facebook created settings to control privacy and visibility. The general settings are located in a drop-down window to the right of the screen, under an arrow (Figure 33 below) that leads the user to a series of Settings and Privacy mechanisms by which they can fundamentally control the visibility of their content and the sources of those who receive the content constituting the first content flow regulation filter (Figure 33 below). In addition, Facebook allows users to control the visibility of each piece of content they post (Figure 33 below).



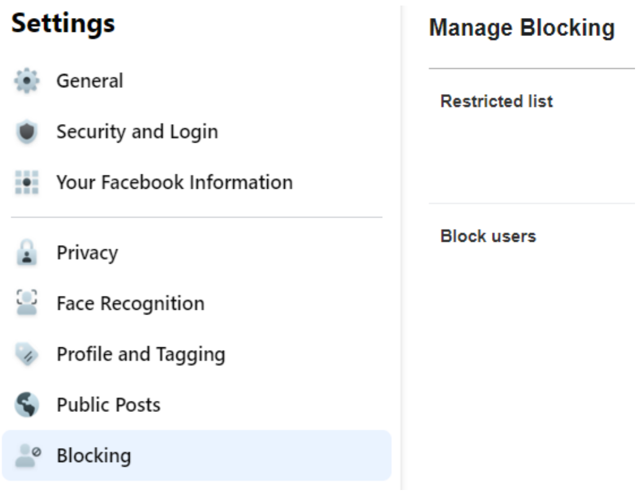
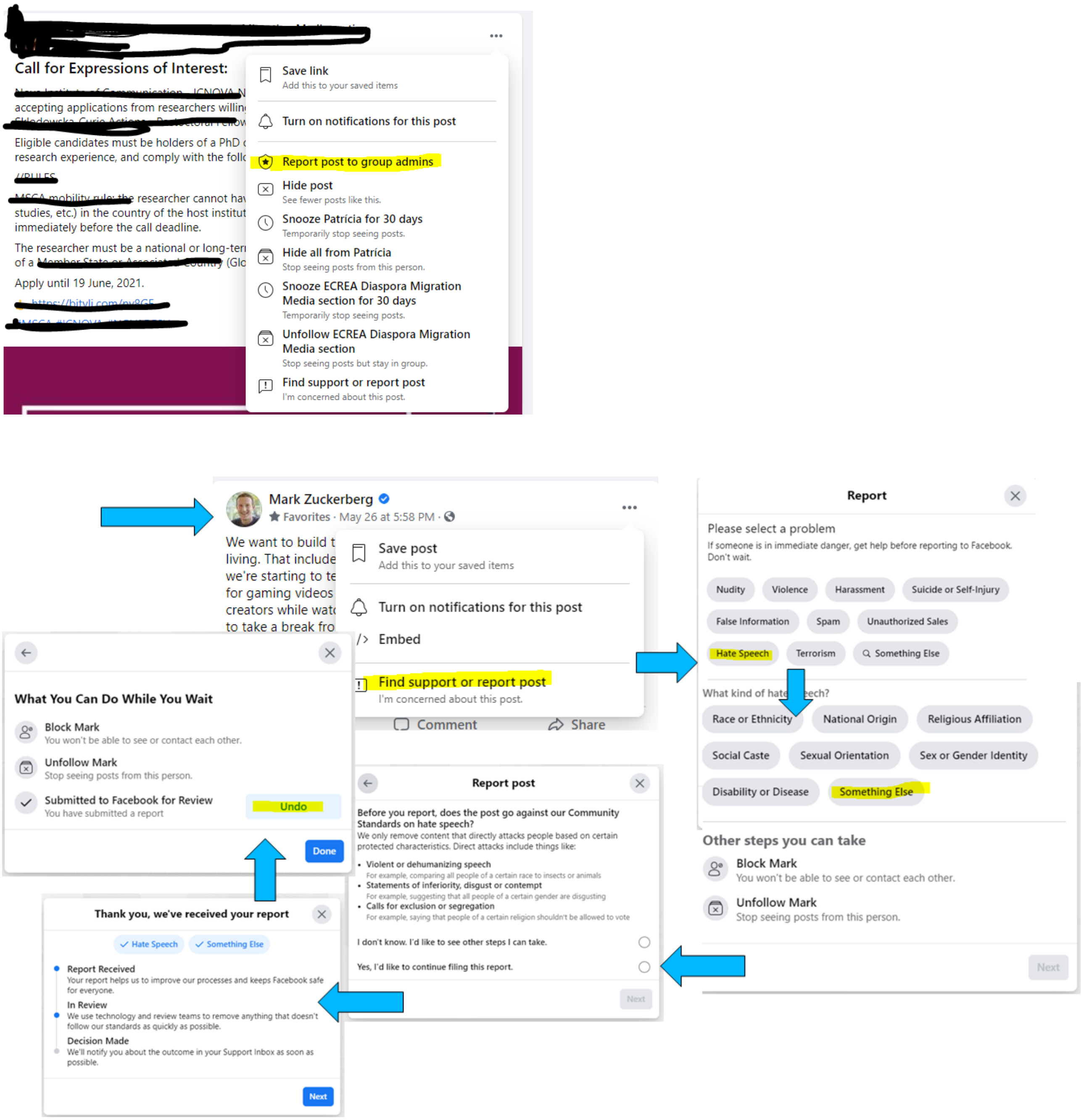


Figure 33 illustrates settings to illustrate how users can manipulate privacy settings and tools to control flow of content.

This set of settings are the first self-regulation technique of the user, who controls the audience for its content: the global community, their contact group, only a handful of their contact groups, or only the user himself. This series of techniques are designed to regulate what information the user wants to be exposed to and, ultimately, aims to cater to the user’s criteria on how information should flow on the platform and how content should circulate. The tools are rhetorically presented as privacy settings, but are ultimately designed to allow users to self-regulate the flow of content. They are the first recommended more accessible techniques for users which ultimately enables Facebook to avoid making decisions on behalf of the user, in terms of the information they want to see. For all hateful content that the user considers should not be on the platform, Facebook has made the flagging or reporting system available.

7.3.1.2 Flagging system

The flagging or report system is a mechanism that Facebook has implemented to systematize the activity of reporting. As Gillespie argues, 'only users are everywhere' (Gillespie, 2018, p. 129) and it is, therefore, only users who understand how hate expresses itself in different regions. The flagging technique is accessed through the specific settings and works as shown below.



[redacted]

Call for Expressions of Interest:

[redacted]

accepting applications from researchers willin

[redacted]

Eligible candidates must be holders of a PhD
research experience, and comply with the follo

[redacted]

researcher cannot hav
studies, etc.) in the country of the host institut
immediately before the call deadline.

The researcher must be a national or long-ter
of a [redacted] (Glo

Apply until 19 June, 2021.

[redacted]

Save link
Add this to your saved items

Turn on notifications for this post

Report post to group admins

Hide post
See fewer posts like this.

Snooze Patrícia for 30 days
Temporarily stop seeing posts.

Hide all from Patrícia
Stop seeing posts from this person.

Snooze ECREA Diaspora Migration Media section for 30 days
Temporarily stop seeing posts.

Unfollow ECREA Diaspora Migration Media section
Stop seeing posts but stay in group.

Find support or report post
I'm concerned about this post.

Mark Zuckerberg
Favorites · May 26 at 5:58 PM

We want to build t
living. That include
we're starting to te
for gaming videos
creators while wat
to take a break fro

Save post
Add this to your saved items

Turn on notifications for this post

Embed

Find support or report post
I'm concerned about this post.

Comment
Share

Report

Please select a problem
If someone is in immediate danger, get help before reporting to Facebook. Don't wait.

Nudity · Violence · Harassment · Suicide or Self-Injury · False Information · Spam · Unauthorized Sales · Hate Speech · Terrorism · Something Else

What kind of hate speech?

Race or Ethnicity · National Origin · Religious Affiliation · Social Caste · Sexual Orientation · Sex or Gender Identity · Disability or Disease · Something Else

Other steps you can take

Block Mark
You won't be able to see or contact each other.

Unfollow Mark
Stop seeing posts from this person.

Next

Report post

Before you report, does the post go against our Community Standards on hate speech?
We only remove content that directly attacks people based on certain protected characteristics. Direct attacks include things like:

- Violent or dehumanizing speech
  For example, comparing all people of a certain race to insects or animals
- Statements of inferiority, disgust or contempt
  For example, suggesting that all people of a certain gender are disgusting
- Calls for exclusion or segregation
  For example, saying that people of a certain religion shouldn't be allowed to vote

I don't know. I'd like to see other steps I can take. ○

Yes, I'd like to continue filing this report. ○

Next

Thank you, we've received your report

✓ Hate Speech ✓ Something Else

Report Received
Your report helps us to improve our processes and keeps Facebook safe for everyone.

In Review
We use technology and review teams to remove anything that doesn't follow our standards as quickly as possible.

Decision Made
We'll notify you about the outcome in your Support Inbox as soon as possible.

Next

What You Can Do While You Wait

Block Mark
You won't be able to see or contact each other.

Unfollow Mark
Stop seeing posts from this person.

Submitted to Facebook for Review
You have submitted a report

Undo

Done

Figure 34 Illustrates Facebook's reporting system for users (from Siapera and Viejo Otero, 2021)

As seen in the Figure above, each piece of content, whether a post or a comment under a post, gives users a number of possible actions. This is accessed through three points placed to the right of each comment or post. To report a post, users must follow the sequence of actions, as detailed on Figure 34. The first option that Facebook offers is to hide the comment, which means that the users will not see the content again. The second option is to temporarily stop viewing posts from the page or profile that has produced hateful content, then unfollow the page/profile. Finally, as a last resort Facebook gives the user the possibility to report the post.

Flagging inappropriate content is a laborious and detailed activity whose study reveals that: first, users are tasked with applying their own judgment to make the decision to report a piece of content and determine the basis on which this report is justified. Second, the options are sequenced in an order that goes from ignoring messages to reporting. The order of options shows evidence of Facebook's relative resistance to prioritizing content deletion. And, thirdly, self-control of the flow of content and individual reporting shows the individualization of the process for tackling hateful content. Indeed, the reporting system does not work for profiles, only for comments; this means that a page that hosts nationalist content and whose initiative is de facto segregationist would not be affected by the reporting system, which expects that each individual piece of content has to be actively reported. Removing a page or disabling an account does not fall into the hate speech category. It, instead, comes under a different category within Community Standards, known as violent and criminal behaviours and dangerous individuals and organisations.

On Facebook, there are two sets of flaggers: normal flaggers, aka users, and members of the so-called 'Trusted Partner Programme' (Cummiskey, 2019). The Trusted Partner Programme refers to the organizations and institutions that collaborate with Facebook and who provide contextual and nuanced information of particular abuses carried on through Facebook. Trusted Partners can become trusted flaggers and have special channels to communicate with Facebook. However, they also use the standardised flagging reporting system, as using both channels inform Facebook of the adequacy of its content standards (Cummisky, 2019).

7.4.1.3 News Feed Settings

Facebook also permits the user to manipulate News Feeds. The News Feed is a proprietary technology registered on United States Patent under the ownership of 'Zuckerberg et al' (2006) and explained by Zuckerberg as 'the most advanced system that considers everything your friends are posting, and all of the media content you might be interested in to then show you what you'll find most important' (Zuckerberg, 5th of September 2016). The News Feed system and how it functions has been under scrutiny since its incorporation in 2007. Back then, News Feed was perceived by early Facebook users as a system that violated users' privacy (Hoadley et al., 2010). More recent critiques concern the role of the News Feed in creating 'Echo chambers'. Echo chambers are a phenomenon that occurs when favoured information among communities reinforces selective exposure and generates group polarization (Quattrociocchi, Walter and Scalam Sunstain, 2016). To overcome the issue of echo chambers, but also to give the user the capacity to select their information sources, Facebook incorporated 'News Feed Settings'. This is fundamentally a self-governing technique that follows a Foucauldian governmentality rationale by which users can voluntarily and manually access the Relationships Database and self–organize their own personal sources of information; that is, to select the users whose information will appear on their News Feed (see Figure 35 and 36)

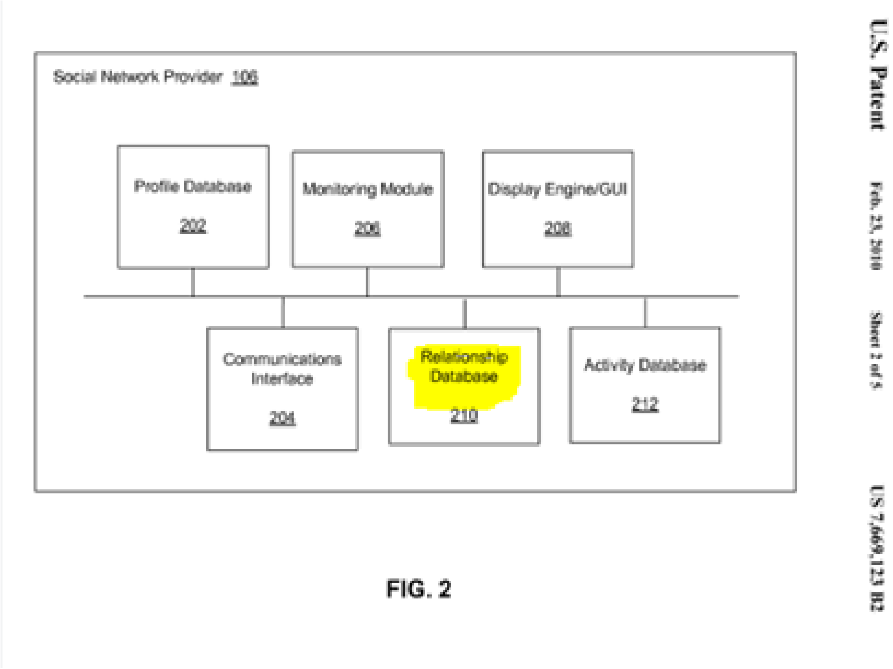
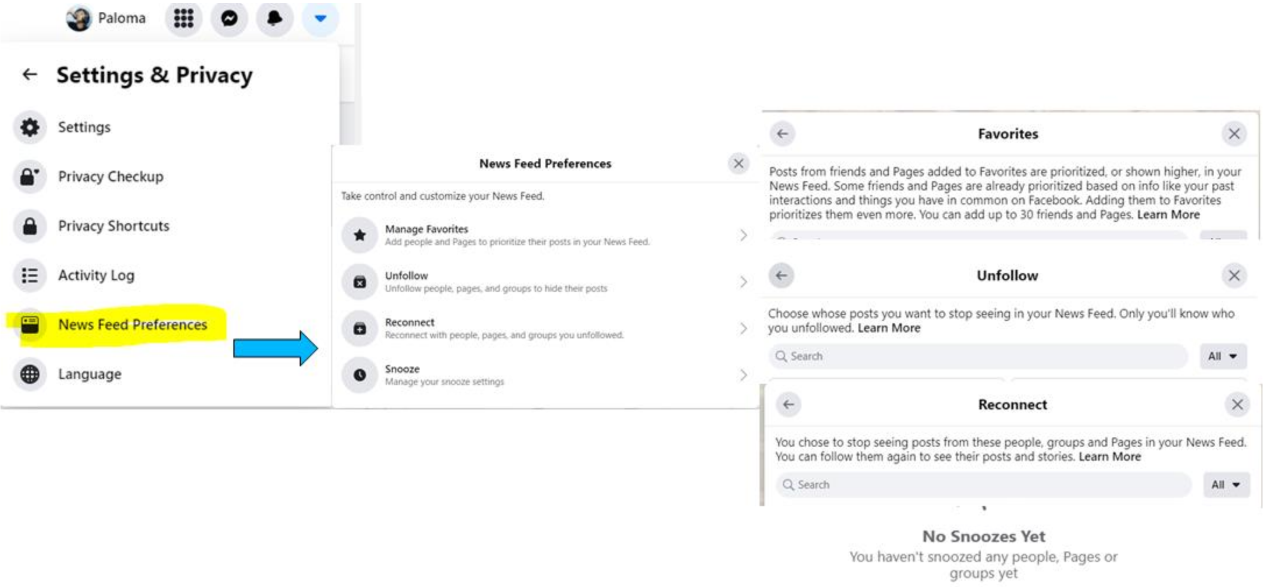


Figure 35 Illustrates News Feed Patent US7669123B2 Facebook Inc. Illustrates the different modules that News Feeds has and highlights the module or database that Facebook has given access to users to manipulate their source of information.



Take control and customize your News Feed.

Add people and Pages to prioritize their posts in your News Feed.

Unfollow people, pages, and groups to hide their posts

Reconnect with people, pages, and groups you unfollowed.

Manage your snooze settings

Posts from friends and Pages added to Favorites are prioritized, or shown higher, in your News Feed. Some friends and Pages are already prioritized based on info like your past interactions and things you have in common on Facebook. Adding them to Favorites prioritizes them even more. You can add up to 30 friends and Pages. Learn More

Choose whose posts you want to stop seeing in your News Feed. Only you'll know who you unfollowed. Learn More

You chose to stop seeing posts from these people, groups and Pages in your News Feed. You can follow them again to see their posts and stories. Learn More

No Snoozes Yet

You haven't snoozed any people, Pages or groups yet

Figure 36 Illustrates Walkthrough on how users can manipulate News Feed settings

In sum, privacy and visibility settings, reporting systems, and control of News Feed are the three product solutions that are under users control and which allows them to customise their user experience by controlling the flow of information. All mechanisms are ruled by the principle of freedom of information, in the sense that they are optional and definitely speak of Facebook's idea of governance, which revolves around and builds upon the willingness of users to participate in the control of hateful content. It is also noticeable that all those mechanisms are tightly related to controlling how content flows on the platform, which does not function as censorship but as regulation of the flow of information. Privacy and visibility settings and News feed manipulation are the most basic product solutions, it is manipulated directly by users and that facilitates Facebook minimum direct intervention. The reporting system is different, in that it constitutes a step further and a call on Facebook for intervention. However, while users are indeed everywhere, the reality is that not all users report abuse on Facebook, as not all users consider it is their obligation to trust the system in place (Siapera, Moreo, Zhu, 2018). Facebook, obliged by external pressure and under the Code of Conduct, has invested in more sophisticated systems and technology to detect harmful content, which includes artificial intelligence for automatic detection.

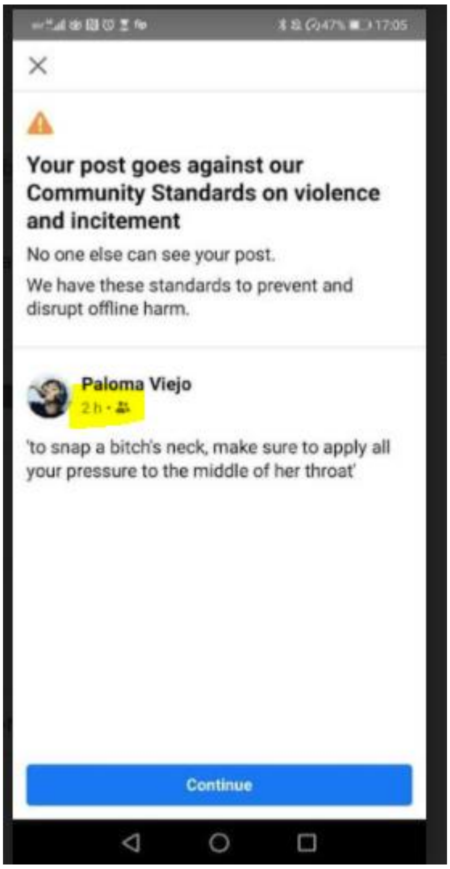
7.4. Artificial Intelligence

The second technique that Facebook has implemented to control the flow of information is based on Artificial Intelligence methods used for the automatic detection of content. These techniques consist of the development, learning, and application of algorithms that detect and filter hateful content, sometimes even before it reaches the user's News feed (Cummiskey, 2019). The use of automation began to be applied to stop pornographic material and nudity (Cartes, 2016). Once Facebook signed the EC Code of Conduct, the company had to invest more effort in complying with the European agreement, which requires companies to review and delete content in less than 24 hours.

In 2017, Mark Zuckerberg announced the investment and advancement of automatic detention systems for hateful content. The role of automatic detention is to stop and eliminate material faster that repeatedly appears on the platform and that clearly breaches community content. Whereas it is claimed that some of that material is deleted before it reaches the user (Cummiskey, 2019), a short experiment run for the present dissertation showed evidence that clear misogynistic material was deleted, not immediately, but two hours after it was posted (see Figure 37 below), showing that, for two hours, explicitly misogynistic content has freely flowed across the platform.

7.4. 1 Experimenting with automatic detection

The experiment consisted of posting the following line: ‘to snap a bitch’s neck, make sure to apply all your pressure to the middle of her throat’.



Your post goes against our Community Standards on violence and incitement

No one else can see your post.

We have these standards to prevent and disrupt offline harm.

Paloma Viejo
2 h ·

'to snap a bitch's neck, make sure to apply all your pressure to the middle of her throat'

Figure 37 illustrates Facebook warning after a misogynistic post was posted by the author

The line for this example was copied from Facebook’s moderation guidelines, leaked to The Guardian and archived as Facebook files. According to The Guardian (2017), this exact line was permitted on Facebook. I posted this at 15:00 on the 3rd of December 2020, but it was not until 17:05 hours that I received a Facebook notification, as seen above. This brief experiment showed that explicit and recognised misogynistic and violent content was on the platform for two hours but also points out that policies on content have become more severe, when compared to 2017.

The EC Code of Conduct stipulates removal in 24 hours, so it can be said that 2 hours is a success. However, 2 hours is also a long window for misogynistic content to be on a platform, as it can be reproduced several times, screenshot, and sent by private encrypted messengers or external platforms that Facebook will not have access to.

Facebook advertises the percentage of content they delete from the platform as positive. Its transparency reports point out that automatic detection has eliminated 98% of what Facebook has established as content that breaches its community standards (Facebook Transparency Report, 2020). However, as Chris Guilliard (2020) points out, Transparency Reports hide the fact that this 98% represents what Facebook has categorized internally as hateful content, and that there is actually the possibility for hateful content to be on the platform.

In addition, it should be noted that not all automatic detection is designed to eliminate content. It is also oriented towards reducing its visibility and limiting its circulation. Some kinds of content are still on the platform, even when its circulation or capacity of circulation is reduced directly by Facebook. This is known as borderline content.

7.4.2 Understanding borderline content

In Zuckerberg's Blueprint for Facebook Content Governance (15 November 2018), he announced the introduction of a mechanism to 'discourage border-line content'. Zuckerberg argues that:

> people get involved disproportionately with more exciting and provocative content' as such it is important to discourage the accumulation of content and reduce its visibility to deescalate a potential threat. According to Zuckerberg, these phenomena can undermine the quality of public discourse and lead to polarization. It also can degrade the quality of our services.

Facebook's solution to all these problems is a 'penalization tool', by which content gets 'less distribution and therefore interrupts engagement' (Zuckerberg, 15 November 2018). In practice, this means that Facebook intervenes regarding content that does not breach the standards but that is close to the line. The intervention consists in a reduction of visibility to disable engagement on topics that can potentially escalate to breaching Facebook policies.

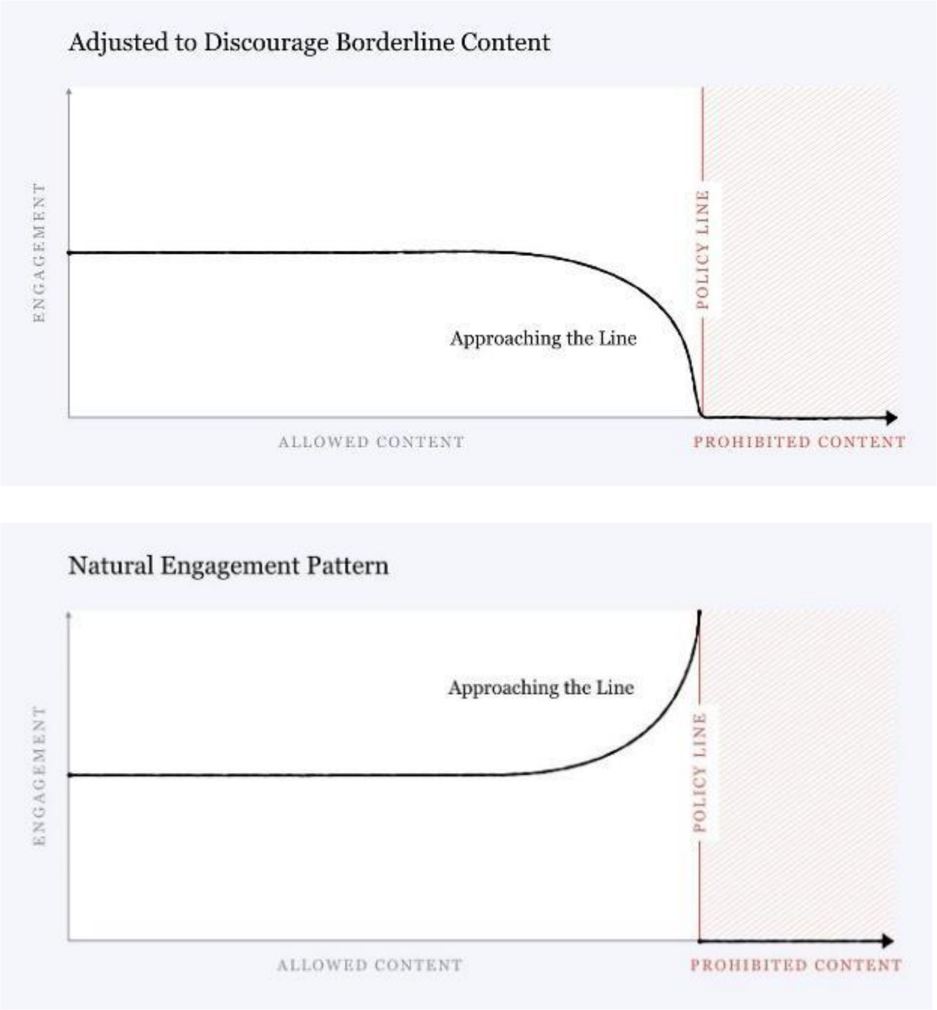


Figure 38 Borderline content. Source: Blueprint for Governance enforcement (Zuckerberg, 15th November 2018).

By applying the construct of borderline content, Facebook does not limit freedom of expression, as content is not removed from the platform. Instead, it limits its capacity of circulation. This shows that Facebook designs safety solutions based on controlling information flow.

In sum, Facebook uses automatic detection to eliminate content that is registered in its database as hateful (Roberts, 2019). In addition, it has developed a third category of content, border line content, that is neither left free to circulate nor is it removed, but rather its visibility is decreased. These techniques provide safety by cutting or limiting the flow of information faster than it would be done manually. However, despite advances in AI and automatic systems, they do not pick up all such content but queue up unclear cases for human reviewing (Cummiskey, 2019).

7.5 Human Reviewers

This section discusses the role of human reviewers, through the research developed by Roberts (2016, 2019), the interviews with key informants, and Zuckerberg's public posts.

According to Roberts, content moderation is where:

> What is the content and what it depicts, its intent or what is meant to do when consumed or circulated, its unintended consequences (that is, what else it might do, beyond its first order intent) and its meaning (which can be highly specific, culturally, regionally or otherwise) all intersect (Roberts, 2019, p. 34)

The job of human reviewers is guided by content standards and their task is to review all posts and publications flagged as potentially harmful. There are over 15 thousand people reviewing content for Facebook (Cummiskey, 2019). These teams are distributed by languages and cultures so that moderators are people with necessary cultural and linguistic skills. Since Facebook's policies are universal and all users are on the same platform, instead of applying local rules, it hires moderators whose knowledge of language and culture allows them to decide whether a piece of content is hateful or not according to Facebook's content standards which are universally applied.

Reviewers have limited access to user information. Teams are divided into internal teams, which are in Facebook Headquarters and external teams or outsourced workers (Gillespie, 2018; Roberts, 2019). Human reviewers specifically look at content flagged by Facebook users or detected by automatic systems. The reviewer would either 'confirm' the report and delete the content, 'un-confirmed' it and leave the content, 'confirm its difficulty' and lower its visibility, or 'escalate it', whereby it is passed up to Facebook internal teams for further review (Cummiskey, 2019). They first look at the piece of content and, only when there is a level of severity, is the reviewer allowed to access more information. In order to avoid conflicts of interest, Facebook uses separation and alarm systems. Using the Facebook system that human reviewers have at their desk, no user has close relations to the reviewer. When there is the case of possible conflict of interest, the system alerts if the reviewer and the user have common connections (Cummiskey, 2019).

On Facebook human reviewers are part of the safety team. In 2017, Zuckerberg announced the hiring of ten thousand employees, 'for safety and security, with the possibility of extending that number to twenty thousand' (Zuckerberg, 1 November 2017). In January 2018, Zuckerberg reiterated that 'to prevent hate speech and ensure the security of the platform, Facebook invests in staff and technology in equal parts so that Facebook invests around 14,000 people working across community ops, online ops, and our security efforts' (Zuckerberg, 31st January 2018). By 2019, Cummiskey pointed out that there are a total of '30,000 people working at Facebook just for safety and security where half of those are content reviewers', revealing that Facebook has invested heavily on safety that since 2017 the company have gone from 7 and half thousand reviewers to more to 13 thousand (Cummiskey, 2019).

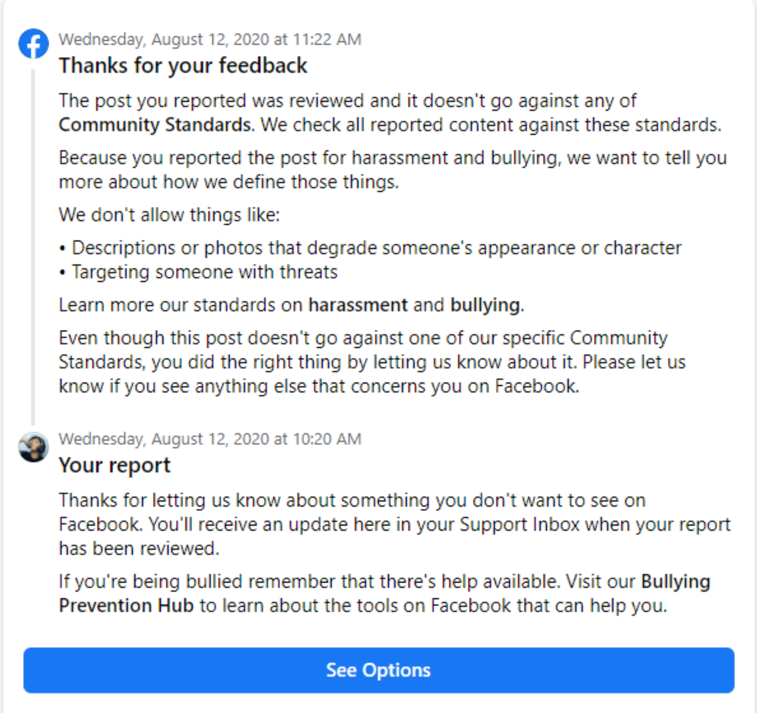
In sum, human reviewers' task is to analyse and decide if a piece is a risk and should be taken out of circulation or if it is a free expression that might contribute to further conversations. While their decision is fundamental for the future circulation of content, in 2019 Facebook created an external board, the Oversight Board, with the capacity to overrule human review decisions made by Facebook moderators. The following section looks at the oversight board.

7.6 Oversight Board

The act of appealing a moderation decision is a system that Facebook incorporated already in 2018 for individual posts. It was specifically conceived for users to express their disagreement with Facebook

> Starting this week, we're also establishing a new appeals process so you'll be able to appeal decisions about individual posts if you think we made a mistake' (Zuckerberg, 24th April 2018) .

The user had the possibility to request a second review for content that the user has reported or posted.



Wednesday, August 12, 2020 at 11:22 AM

Thanks for your feedback

The post you reported was reviewed and it doesn't go against any of **Community Standards**. We check all reported content against these standards.

Because you reported the post for harassment and bullying, we want to tell you more about how we define those things.

We don't allow things like:

- Descriptions or photos that degrade someone's appearance or character
- Targeting someone with threats

Learn more our standards on **harassment** and **bullying**.

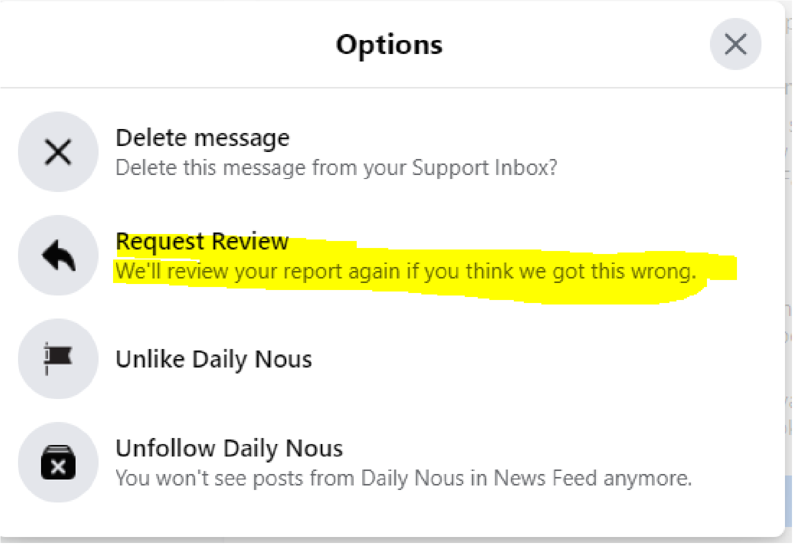
Even though this post doesn't go against one of our specific Community Standards, you did the right thing by letting us know about it. Please let us know if you see anything else that concerns you on Facebook.

Wednesday, August 12, 2020 at 10:20 AM

Your report

Thanks for letting us know about something you don't want to see on Facebook. You'll receive an update here in your Support Inbox when your report has been reviewed.

If you're being bullied remember that there's help available. Visit our **Bullying Prevention Hub** to learn about the tools on Facebook that can help you.



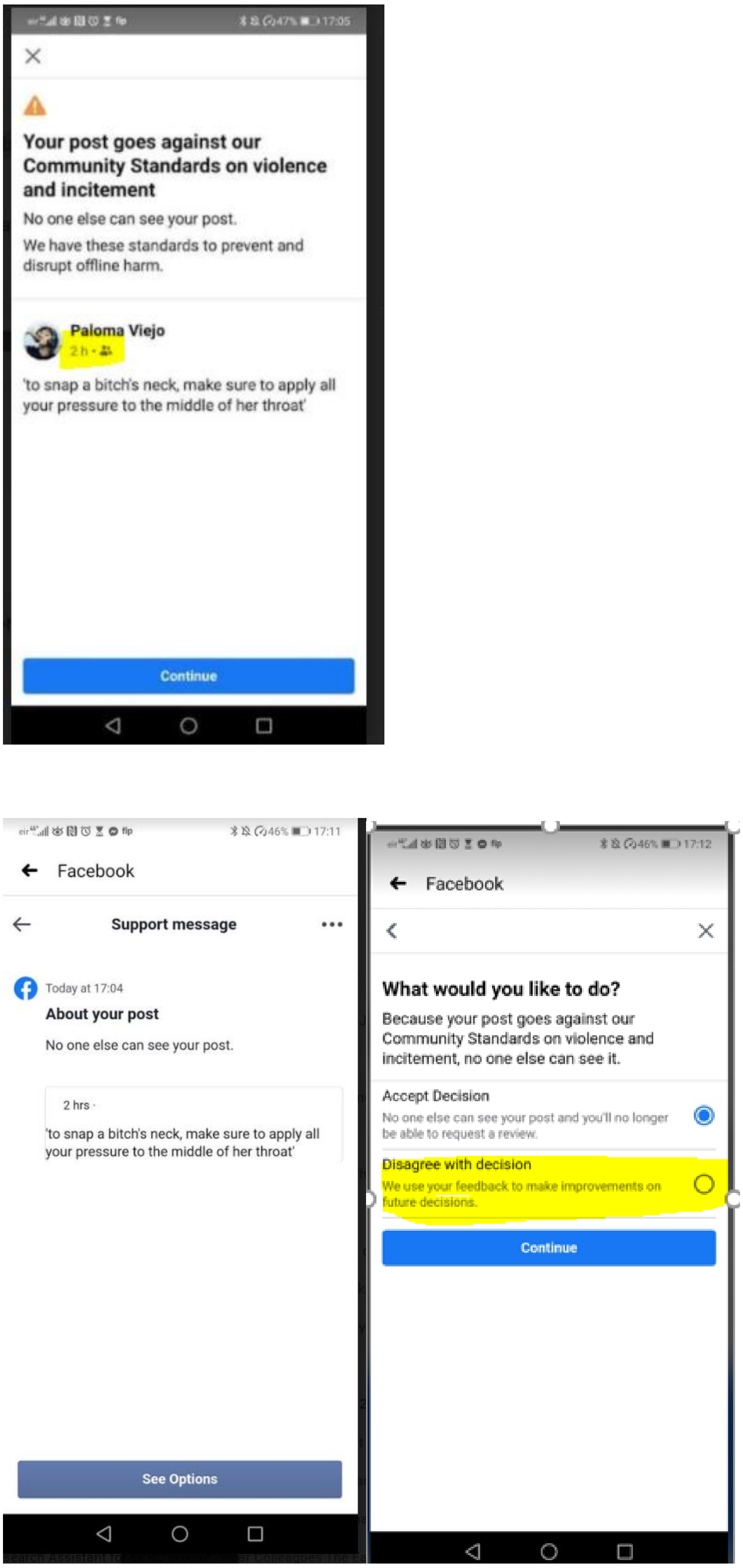
Options

Delete message
Delete this message from your Support Inbox?

Request Review
We'll review your report again if you think we got this wrong.

Unlike Daily Nous

Unfollow Daily Nous
You won't see posts from Daily Nous in News Feed anymore.



Your post goes against our Community Standards on violence and incitement

No one else can see your post.

We have these standards to prevent and disrupt offline harm.

Paloma Viejo

2 h ·

'to snap a bitch's neck, make sure to apply all your pressure to the middle of her throat'

Today at 17:04

About your post

No one else can see your post.

2 hrs ·

'to snap a bitch's neck, make sure to apply all your pressure to the middle of her throat'

What would you like to do?

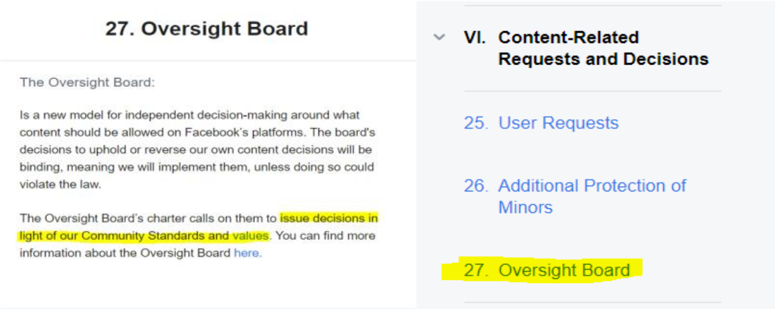
Because your post goes against our Community Standards on violence and incitement, no one else can see it.

Accept Decision ●
No one else can see your post and you'll no longer be able to request a review.

Disagree with decision ○
We use your feedback to make improvements on future decisions.

Figure 39 illustrates how to disagree with decision to the moderation team.

To request a review or disagree with a decision to the moderation team Facebook has made some tools available. The problem with this option is that it puts Facebook at the centre of the decision making. In order to overcome this, Zuckerberg announced the creation of an external panel. On ‘A Blueprint for Content Governance and Enforcement’ (15 November 2019), Zuckerberg announced the creation of an external panel of experts. This announcement led to the creation of the Oversight Board, which functions as an external enforcement technique that is mobilised in case of disputes, i.e. when a piece of content is taken down and the owner appeals. In addition, Oversight Board Members provide Facebook policy recommendations. Facebook commenced to organised the Board in 2019 (Cummiskey, 2019), justifying its creation on the basis of three reasons: to prevent the concentration of power on Facebook teams, which were tasked with setting the de facto global standard of free expression; to create accountability; and to ensure that these decisions are made in the interest of Facebook Community and not for commercial reasons (Zuckerberg, 15 November 2018). Building the Oversight Board was a task that involved a series of public consultations, focus groups in several countries of all continents, and consultations with experts. The board was formalized in December 2019 and its functions were included as part of the Community Standards to make it available to the users (see Figure 40).



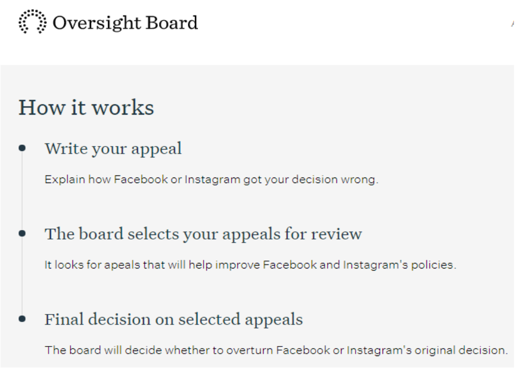
27. Oversight Board

The Oversight Board:

Is a new model for independent decision-making around what content should be allowed on Facebook’s platforms. The board's decisions to uphold or reverse our own content decisions will be binding, meaning we will implement them, unless doing so could violate the law.

The Oversight Board’s charter calls on them to issue decisions in light of our Community Standards and values. You can find more information about the Oversight Board here.

VI. Content-Related Requests and Decisions
25. User Requests
26. Additional Protection of Minors
27. Oversight Board



Oversight Board

How it works

- Write your appeal
  Explain how Facebook or Instagram got your decision wrong.
- The board selects your appeals for review
  It looks for apeals that will help improve Facebook and Instagram's policies.
- Final decision on selected appeals
  The board will decide whether to overturn Facebook or Instagram's original decision.

Figure 40 Shows where the Oversight Board is located in Community Standards and how to appeal to the Oversight Board. Source Facebook site and Oversight Board site

Once users opt (voluntarily and on an individual basis) for a review of the decision, the appeal first reaches Facebook, which identifies the cases and prioritizes them based on the level of a) significance and b) difficulty (Facebook, Oversight Board, 2020). By significance, Facebook refers to cases involving actual violence, having severe consequences, or far-reaching public discourse. By difficulty, Facebook refers to cases that call into question existing policies. Facebook first classifies them, and passes it to the members of Oversight Board who have the responsibility to specifically deliberate if those pieces of content will be on the platform or if the single piece of content will be taken down. In addition, the Oversight Board can make recommendations on content standards.

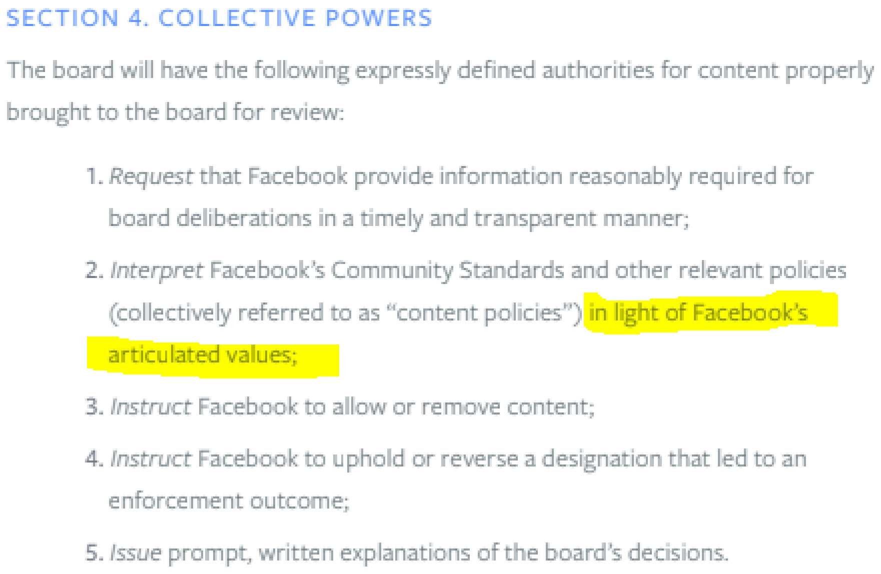
7.6.1 Oversight Board Chart

While the Oversight Board is promoted as an external independent panel, further analysis reveals that the Oversight Board is regulated by ‘The Oversight Board Charter’. This Charter is composed of seven articles and it acts as the foundational governing document for the Oversight Board. The Chart stipulates the board structures, responsibilities, and relationship with Facebook. It also specifies the Principles and Values that underpin the Oversight Board.

A closer reading of the Charter reveals that the principles that rule its decision making are the Facebook Principles. Evidence of this is found in Article 1: Members, Section 4, Collective powers. Point 2 Interpret. By virtue of this article, the board has to interpret content and community standards in light of Facebook articulated values.

> 2. *Interpret* Facebook's Community Standards and other relevant policies (Collectively referred to as 'content policies') in light of Facebook's articulated values (emphasis added)

(Facebook, 2021 https://www.oversightboard.com/governance/)



SECTION 4. COLLECTIVE POWERS

The board will have the following expressly defined authorities for content properly brought to the board for review:

1. *Request* that Facebook provide information reasonably required for board deliberations in a timely and transparent manner;
2. *Interpret* Facebook's Community Standards and other relevant policies (collectively referred to as "content policies") in light of Facebook's articulated values;
3. *Instruct* Facebook to allow or remove content;
4. *Instruct* Facebook to uphold or reverse a designation that led to an enforcement outcome;
5. *Issue* prompt, written explanations of the board's decisions.

Figure 41 illustrates Article 1: Members, Section 4, Collective powers. Points 1 to 5

In addition, Article 2 Bases of decision making, stipulates that 'the board will review content enforcement decisions and determine whether they were consistent with Facebook's content policies and values', adding: 'When reviewing decisions, the board will pay particular attention to the impact of removing content in light of human rights norms protecting free expression' (Facebook, 2020, unpaginated). In other words, their decisions are not based on what they, as experts, would consider appropriate judgements on whether content is hateful and must be removed, but rather on the extent to which Facebook's decisions were correct, given Facebook's own guiding principles and content standards.

The commonality among user settings, Artificial Intelligence, content reviewers, and the Oversight Board is that they are techniques to provide safety. Together, they do constitute a dispositive of security (Foucault, 2010, p.278) whose mission is to provide a safe environment to protect the principle of Voice. As it will be argued below, this dispositive has more similarities with a factory controlling the quality of its products (Gillard, 2020) than with a measure implemented to combat discrimination. As such, Facebook's idea of 'operational' measures refers to the capacity of enforcement measures to control and possibly take down large quantities of content. In considering this, the following section aims to look further into, Facebook's dispositive of security, its function, and significance for the governance of hate speech.

7.7 Facebook's dispositive of security

In order to provide Safety, Facebook has set up a series of enforcement measures, namely: user settings, automatic detection, human reviewers, and the Oversight Board. The commonality of these techniques is that they only function for content that is circulating on the platform. They all cooperate with each other, forming a network and an assemblage (Cummiskey, 2019), laid out to constantly measure if content should or should not be deleted. Their arrangement is designed as follows: under the principle of Voice, Facebook allows users to upload any form of content. If the content is harmful it can be blocked, its visibility reduced, or reported by another user or detected by automatic systems. The content that is flagged by other users goes in a queue for human reviewers, whereas hateful content that is detected automatically is deleted without reaching anyone's newsfeed (Facebook, 2021). This is the case if it is clear that it breaches content standards. If there is any doubt, for example a video with ironic content, it gets sent to human reviewers. Human reviewers then decide according to content standards if the piece of content should remain on the platform or not. If deleted, the user who originally posted has the right to appeal to Facebook, where it is then considered by the Oversight Board (Facebook Oversight Board site, 2021, unpaginated). Content therefore passes through a series of techniques that constantly reassess if the piece should be left or taken down. The arrangement of the techniques implies that, if anything, taken together they favour the possibility for content to remain on the platform. That is, once content is removed, Facebook offers two opportunities to reverse this removal, firstly by appealing to Facebook and, secondly, by appealing to the Oversight Board. The Oversight Board's mission is to protect freedom of expression (2021), and it is there to assess Facebook's decision to remove content but not its decision to leave potentially harmful content online. In light of this safety scheme that Facebook has implemented, it is possible to argue then that Facebook's notion of safety consists of protecting users' voices while maximising the possibility of safety using techniques that fundamentally control the circulation of content. While this can be seen as a reasonable compromise, Facebook's emphasis or strategy is to look after the safety of its structures, and to guarantee that users upload content. Thus, their safety in terms of their actual wellbeing and physical or mental integrity is not deemed important. We can therefore argue that Facebook ultimately is not concerned with justice, oppression, or emancipation of the group that is targeted, but with the capacity of the user to keep uploading content (Terranova, 2015).

At the core of Facebook apparatus of security lies the tension between safety and freedom. Since securing freedom for Facebook does not accept coercive measures, the result is a display of mechanisms and techniques that control and secure the freedom of circulation of content. Indeed, Facebook's solution has been building enforcement techniques upon the structures that facilitate the circulation of content. This is what Foucault (2007) refers to as a dispositive of security. As pointed out in chapter four, security is understood as the need that emerges in liberalism to secure freedom (Foucault, 2007). Since securing freedom for liberalism does not accept coercive measures, the result is a display of mechanisms and techniques that control and secure the freedom of circulation of people and goods (Foucault, 2007), or of content in the present case.

Zuckerberg has made it clear that the key to protecting voice and ensuring safety relies on controlling the flow of content. According to Zuckerberg, 'freedom means you don't have to ask permission first, and that by default you can say what you want. If you break our community standards or the law, then you're going to face consequences afterwards. We won't catch everyone immediately, but we can make it harder to try to interfere' (Zuckerberg, 21 September 2017). What this quote reveals is that Facebook will protect the users' right to upload any form of content, but in turn, Facebook controls their structures and how they are used by setting up a series of mechanisms of control and detection that will 'catch' hateful content later, once it is flowing on the platform. In implementing the above discussed enforcement techniques, Facebook controls and secures the circulation of content and benefits from it.

The dispositive of security serves the principle of Voice, which commences -technologically speaking- with a tab that can be found on all user profiles, at the top of each user's Facebook wall, and under a simple question: "What's on your mind?" Through this almost naïve question, the user is enticed to spell out their thoughts, providing an avenue for hateful content to be on the platform.

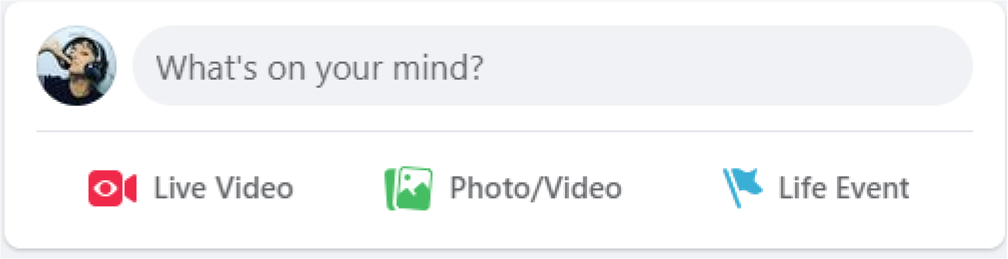


Figure 42 Illustrates Facebook users tab to upload content.

Since hate speech content is not desirable, what technically follows from the Voice tab is an entire 'dispositive of security' made of users settings, artificial intelligence, automatic detection, human reviewers, and the oversight board that Facebook has gradually implemented on Facebook to control the flow of content. Any of the techniques that are part this dipositive do not limit freedom of expression (voice) but the freedom of information. Facebook's safety mission, or intervention in content, has been criticized as 'censorship', as it challenges notions of Facebook as a free speech zone. This statement is not uncommon from organisations that defend Freedom of Expression, such as Article 19.org, which often presents Facebook as a freedom of speech censor. However, the nuance or technicality that we need to return to, is the fact that Facebook does not limit the voices of its users; rather, it controls the flow of information. As such, Facebook does not limit freedom of expression, but freedom of information. This differentiation is not a casualty, but the fruit of Facebook balancing business with ideology, as reflected by Facebook Principles, particularly when Facebook eliminated Article 3 Free flow of information and substituted it with the mission of Safety.

Circulating content is related with the economy of data. According to Foucault to actively promote the circulation of peoples and goods, avoids scarcity and prevents accumulation. If we start reading Facebook or any social media platform under this light we can actually see that to promote the circulation of content, any type of content is at the heart of Facebook. So hateful content has a role. Hateful content is necessary as it enables the creation of more content, but also to allow too much accumulation of content would give as results bottlenecks in terms of diversity of content. Therefore, to eliminate hateful content, is not related with the pursue of equality as previous actors aim to achieved during the human rights, but to simple continue to promote the circulation of content.

The question that follows then is: Is controlling information and the flow of information a pertinent measure to combat hate speech or discrimination? In adopting a social justice viewpoint, the answer would be no. Facebook's dispositive of security is the fruit of a larger ideological apparatus of governance whose eminent liberal views do not have the inherent capacity to fight against structural discrimination (Pincus, 1994). Instead, Facebook's security only functions as a constant and vigilant machine that systematically reduces hateful content, but that, in itself, does not have the power or capacity to alter the conditions of possibility for hateful content be on the platform. In fact, this thesis argues that only transforming the ideological position towards voice and towards equality would create those possibilities.

7.8 Conclusion

This chapter has explored the techniques that composed both disciplinarian and security dispositive of Facebook and has looked at the values that underpin these dispositives, which are voice and safety. In making sense of this finding, the present chapter argued that, while the principle of Safety justified the creation of content standards and the implementation of enforcing techniques, the principle that underpins the actual practice and decision making is the Principle of Voice. The principle of Voice operates in the first instance as a tab with a question for the user: “What’s on your mind?”. This tab offers the fundamental condition of possibility for Facebook to obtain its main economic source, content, and it is protected both for economic reasons and for ideological reasons.

Facebook defends freedom of expression with limitations. Limitations do not limit users' freedom of expression, but the circulation of content, which Facebook justified under the rationale of safety. With this aim, Facebook has designed a series of security techniques, known as content moderation (Gillespie, 2018). This chapter argues that all those techniques constitute an apparatus of security whose aim is, not to moderate, but to control how content circulates, to maximise user safety, to maximise the analysis of content, and to maximise the possibility for content to remain on the platform. Facebook apparatus of security is a vigilant, preventative system of control, set up to be constantly on the lookout for the appearance of a security threat. This apparatus of security does not function alone, but in coherence with other activities carried out by the Policy department and leadership, which the next chapter will refer to as dispositives of discipline and ideology, respectively.

Based on findings in chapter six and chapter seven, the next chapter, Chapter eight, aims to conclude this thesis by illustrating how Facebook governs hate speech with a hierarchically organised apparatus of governance that is led by Facebook's dispositive ideology.

Chapter 8. Facebook's Hate Speech Apparatus of Governance

8.1 Introduction

This chapter aims to conclude the thesis and illustrate how Facebook governs Hate Speech. As it was pointed out in chapter four, social media has added to previous ethical, philosophical and legal conversations, a conversation about management and operability, adding a new layer of complexity to the concept of hate speech. Governing hate speech in social media is a framework made of ideological principles and technological affordances that informed and reinforce each other. Therefore, in order to explain how Facebook, governs hate speech, this thesis has aimed to present the techniques that Facebook has designed and the values and principles that underpin each techniques.

Looking at how Facebook governs hate speech from an ideological angle is research that, to the best of my knowledge, has not been undertaken before. This research angle was inspired by Miller and Rose who indicated that in order to research governance we should look at the knowledge and ideology that underpins all governance techniques (2008, p.30). This thesis, has not only look at the principles and values behind each technique, but also the source of Facebook ideology, namely Facebook Principles and Mark Zuckerberg as covered in chapter six. From Foucault's lectures on Security, Territory and Population (2007), I drew upon the notion of discipline and disciplinary techniques, by which I refer to all the techniques with the capacity of managing Facebook, creating regulations, and the authority to implement them. Finally, there is the Enforcement category, a term previously used by Tijana Milosevic (2018), which refers to all techniques by which the governance of hate speech is executed within the platform. In this case, both discipline and enforcement have been covered in chapter seven.

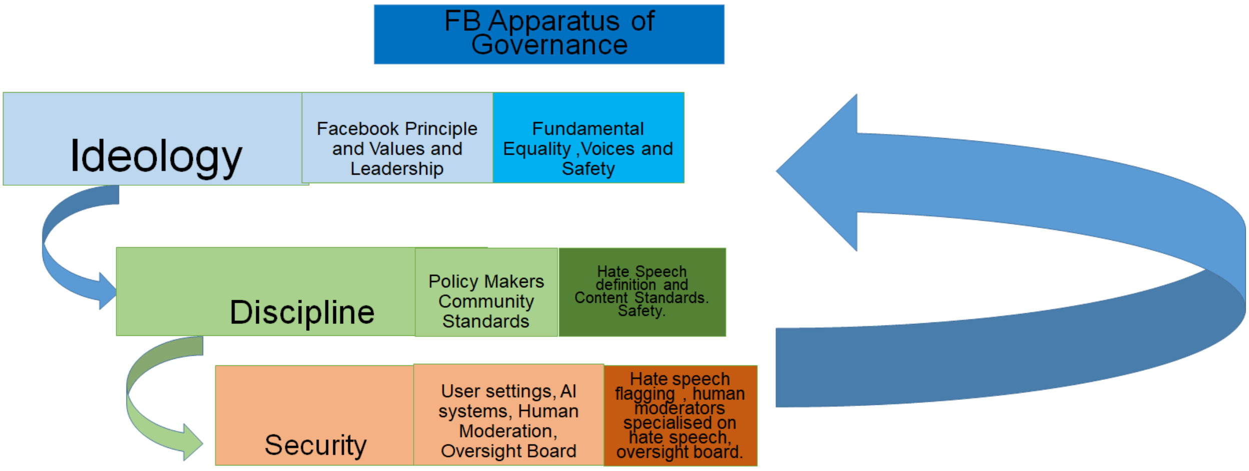
In considering all the above, this chapter returns to Foucault and the notion of dispositive to explain how Facebook governs hate speech both ideological and technologically. The chapter argues that there are three dispositive namely, ideology, discipline, and security, that together they conform Facebook Apparatus to Govern Hate speech. Facebook apparatus to govern hate speech is a top down structure lead by an ideological setting that clearly define Facebooks position in relation to equality and freedom. Facebook's definition of equality and freedom are crucial to understanding Facebooks understanding of hate speech and the technological affordances design to operate hateful content. Therefore, in analysing Facebook's ideological settings, it is possible to conclude that Facebook does not aim to fight oppression, but to reduce accumulation of hateful content to protect its structures and to operate efficiently and profitably as a digital platform.

In explaining Facebook apparatus to govern hate speech, then, the chapter returns to the historical argument of Hate speech, contextualising Facebook in the history of hate speech regulation, and highlighting how this thesis seeks to make an original contribution to knowledge on Hate speech and Platform Governance. Finally, the chapter presents implications for future research.

8.2 Facebook's Apparatus to Govern

The structures of power can be understood in an inductive manner, ascending from the 'most infinitesimal mechanisms' to ever more general forms (Finlay, 2015). The present chapter reorganises all the dispositive identified in chapter six and seven, namely, the dispositive of ideology, the dispositive of discipline, and the dispositive of security, looking at how they interact with each other, how they depend on each other, and how they influence each other.

In chapter six and seven, I argued that Facebook has three dispositives: the ideological dispositive, which is comprised of Facebook Principles and Mark Zuckerberg's posts; the disciplinary dispositive, which is comprised of the Product Policy team, Community and the Content Standards; and the dispositive of security, made up of user's settings, AI systems, human reviewers, and the Oversight Board. Together, they form Facebook's Hate speech apparatus of governance as the following figures 43 illustrate.



FB Apparatus of Governance
Ideology
Facebook Principle and Values and Leadership
Fundamental Equality ,Voices and Safety
Discipline
Policy Makers Community Standards
Hate Speech definition and Content Standards. Safety.
Security
User settings, AI systems, Human Moderation, Oversight Board
Hate speech flagging , human moderators specialised on hate speech, oversight board.

Figure 43 Facebook Hate Speech apparatus of Governance

8.2.1 Dispositive of Ideology

As chapter six pointed out, Facebook's ideological dispositive is composed by Facebook Principles and by Mark Zuckerberg.

Gillespie argues that Facebook, of all possibilities, would prefer not to govern (2018). However, Facebook has always cultivated a will to govern. Zuckerberg argued that 'one does not create technology to build a company, but to change the world' (16th August 2016, Video, min 4.50). Inevitably, aiming to change the world leads to governance. In fact, it has always been Zuckerberg's idea to use Facebook as a vehicle to contribute to changes in society.

> There is going to be some big picture or thing that you are moving towards in the future, that is the ideology behind what you are doing, you know? You know we think that social networks should be platforms, we think the world should just generally more open and transparent, that people should share more stuff, but those things aren't products on itself, so in fact there has to be a very tactical use-case that drives what users are going be doing on a daily basis… you probably don't have enough empathy for what people who have a use case are doing, and what one application is going to do, that the world is becoming more open...Facebook is one of those companies that is going to push those things forward' (Zuckerberg Video transcript, 2009)

It would be different, however, to say that Facebook did not always want to intervene on the type of content that users upload. As chapter six illustrated, before 2017 Facebook was user centred and invested in settings and product solutions for users to make their own individual decisions around content (Cummiskey, 2016). It was from 2017 onwards that Facebook commenced to directly intervene by strengthening its content standards. In order to do so, it first had to make a structural change, or change its DNA (Zuckerberg, 2018), which meant that it had to change its Principles, an action that, as Chapter Six argues, consisted of removing its commitment to the free flow of information in favour of a newly formed mission of Safety.

On Facebook, no one can change its Principles and Values except leaders, and in particular Facebook creator and CEO, Mark Zuckerberg. Part of the analysis conducted for this thesis with policy officers and the Oversight Board was with the aim of examining to what extent policy officers or external independent members have the capacity, task, or power to alter Facebook's approach to hate speech. No evidence points to the existence of any other department on Facebook or outside Facebook with the power or capacity to change Facebook Principles. As Chapter seven illustrates, policy officers can transform content standards, suggest changes, but they do not have the power to change Facebook Principles and missions. When Facebook created the Oversight Board, the thesis examined its capacity to change Facebook structures. Our reading of the Foundational Chart reveals that Facebook's newest technique of governance is bound by Facebook Principles. The Oversight Board, therefore, does not have the authority, nor the capacity, to transform Facebook Principles.

In considering all the above, it is possible to argue that the Facebook's Governance apparatus of hate speech depends on the ideological dispositive aim of determining the direction of the platform. This ideological dispositive protects the idea of Voice, Fundamental Equality, and drives the mission of security, which directly influences how Facebook defines and operates hate speech adopting an Neutral Viewpoint with emphasis in making Facebook platform a safe environment for users to keep uploading content.

8.2.2 Dispositive of Discipline

Echoing Foucault's arguments, the function of Facebook's policy officers is to discipline the online space that constitutes Facebook. This is undertaken, not so much in order to establish limits on hateful content, but in order to ensure the production and circulation of all kinds of content, and to provide some form of stability to Facebook. Ultimately, as Terranova argues, it aims to 'secure the social' (2017) by ensuring that users do not disengage, as our informant Siobhan Cummiskey put it in 2019.

The Product Policy team are the designers of what is and what is not acceptable on the platform and their role is to discipline the space, to ensure that the platform functions according to the order set by Facebook Principles: voice, equity, and safety (Cummiskey, 2019). The policy department is in charge of selection and modification of the content standards, the training of human reviewers, the relationship with trusted flaggers, creating and implementing programmes, and engaging with external stakeholders. Content Standards are the main disciplinary document on Facebook and regulates what all users can or cannot say and what human reviewers can or cannot delete. Therefore, the Product Policy team and the content standards ensure that Facebook principles are reflected in the platform practices and dynamics and, together, they constitute Facebook's dispositive of discipline.

8.2.3 Dispositive of Security

As it is argued in chapter seven, the dipositive of security is arranged to enforce the content standards and it is supervised by policy officers. Facebook's security aims to protect the activity of uploading content –or principle of voice- by functioning as a constant and vigilant machine that systematically reduces hateful content to create a less hostile environment, but which, in itself, does not have the power or capacity to alter the fundamental conditions of possibility for hateful content to be on the platform, which once again, is based upon the principle of Voice and that technology that Facebook makes available to upload content.

In sum, Facebook's apparatus to govern hate speech is a top down structure, ideologically determined to facilitate the production of data, and to avoid the accumulation of data that might prevent the user from producing more content. The apparatus that Facebook has implemented does not aim to fight oppression but to allow Facebook to operate efficiently as a digital platform. In considering this, it is possible to argue that this analysis has indeed allowed us to test the extent to which Foucault's ideas on security, territory, and freedom are valid in the new digital territories, as Terranova argued (2015). In conclusion, the rationale behind Facebook's definition and approach to hate speech is far removed from social justice demands or the philosophical motivations and drivers of the positions we identified so far in the previous chapters two and three. Rather, it is a matter of organising Facebook's digital space to 'protect the social' (Terranova, 2015) and the production of content by ensuring that Facebook users do not encounter large volumes of hateful content.

8.3 Facebook and the history of hate speech regulation

It was the Soviet bloc in 1947 who persisted in regulating discrimination through a proposal to Article 7 for the Human Rights Chart. In its draft, the Soviet bloc proposed regulating privileges, protecting minorities, and criminalizing expressions of superiority. After a series of tensions and amendments on behalf of the Western Bloc, protection against discrimination was approved. However, the final version no longer protected minorities, nor did it regulate privileges. Rather, it gave protection to all individuals equally and recognized them as equal before the law. The Western and liberal amendment to the Soviet proposal on how to regulate discrimination survived and triumphed, and what we call the Neutral Viewpoint became the widest spread form of regulation for hate speech.

Zuckerberg's discourse is rooted in liberal ideology, and Facebook is a continuum of the trajectory that the Western Bloc set for regulating hate speech. Zuckerberg's definition of hate (16th August 2017) speaks of individual behaviour and he strongly believes in the power of the individual voice. An analysis of Zuckerberg’s words can be related to those of Californian ideology, ‘where all individuals could express themselves freely within cyberspace’ (Barbrook & Cameron, 1996, p. 45). Facebook aimed to contribute to this idea which also enclosed the notion of connectivity and global community, as McLuhan stated (1967). However, hate is also an idea freely expressed by many, and it did penetrate the platform, forcing Facebook to renounce its ideal of freedom in order to protect its structures. Regulating hate speech and hateful content was one of the measures that Facebook undertook, and it did so by adopting the Neutral Viewpoint , which was developed and defended by the liberal left at US Campus Universities in the 1980’s. In this sense, Facebook has not innovated on how to define hate speech.

However, if we look at how Facebook enforces its hate speech policy for which the platform categorised hateful content as a safety problem, then we can see where the innovation lies.

Indeed, the Neutral Viewpoint only punishes the act of speaking hatefully, a notion that emanates from Stuart Mill, for whom actions should only be limited to prevent harm to other individuals (1859). This position discards the possibility of paying attention to the reasons why some groups harm others and only encloses two possibilities with respect to hate speech and discrimination: preventing and punishing the specific act of discriminating. Therefore, to focus on the action, all possibilities to intervene get intertwined with notions of safety and security.

As a result, and by focusing on the notion of safety, Facebook implements a coherent set of rules and enforcement, but - in the absence of courts, judges or superior bodies that could take larger considerations - it has displaced the intention held by previous actors. If for previous actors Hate Speech regulation was intended to be a resource to achieve equality, Facebook operationalised hateful content as a matter of security. Hence, this is how Facebook contributes and transforms the notion of hate speech, by no longer associating hate speech with equality but, instead, to a safe environment.

Facebook has certainly contributed to social change, but it does not contribute to fighting against oppression or emancipation. As Koretsky argued, we must seek equality, not only according to the law, or by arguing we all are equal before the law, but 'according to fact and substance' (1947). There is no fact and substance in Facebook machinery and the approach they have adopted is inherently incapable of fighting discrimination. Facebook approach to hate speech has been set up to only reduce hateful content with the singularity that their problem with scale of hateful content is a self-created problem and their remedy is what they can control, so it is possible to argue that Facebook, whose protection of voice and defence of fundamental equality seems unmovable, will always allow the production of hate speech in consequence stripping its users of agency to fight against discrimination.

8.4 Contribution to existing literature

One of the key theoretical findings of this dissertation has been to revitalize Morsink's argument that the Soviet Bloc created the foundation to establish hate speech regulation (1999). Revitalising this argument has contributed to the idea that the regulation of hatred and the fight against discrimination has a long history, with a meaning and scope greater than the simple limitation of freedom of expression. The Soviet perspective for regulating hatred falls far from most of the perspectives developed during Hate Speech Codes on American campuses in the 1980s. During this period, the prominence of allusions to freedom of expression, understood in the American sense, and endowed with a tone and a liberal style, were hard to ignore and this debate has permeated all the literature on this topic. So much so, that the concept of fighting expressions of hatred has lost meaning in a lot of literature, as it is understood only in relation to freedom of expression.

Consecutively, a second contribution is the organization and elaboration of the four different approaches to hate speech. The Historical Materialist approach, the Neutral Viewpoint , the Freedom of expression absolutism approach, and the European approach. Discerning, systematizing, and organizing the different approaches to hate regulation has been at the centre of the current thesis. This task was inspired by Delgado (1993) and Altman (1993), who tried to separate the arguments that were given during Hate Speech on US Campus debates. The authors found two arguments: one made up of conservative and neoliberal groups, who advocated Freedom of Expression absolutism and a second prepared by Liberal Left groups that advocated regulating the act of discrimination. This separation was necessary and enlightening, since in the literature these were mixed together, without offering clarity about the different positions and their implications. As a result, this dissertation has not sought to make comparisons, but has helped to organize the different arguments that exist while providing two new visions that are not American, such as the Historical Materialist approach and the European approach.

Thirdly, and fundamentally, the research and empirical results of this dissertation contribute to the literature on Platform Governance in the following manner. This dissertation contributes to understanding Facebook's ideological settings to approach hate speech, research that had not been previously carried out. This complements José Van Dijck's (2013) work, who observed that Facebook Principals was a hierarchically higher document on Facebook's structure. The author does not explore the meaning of this observation, as her focus was on the Terms of Service of various media platforms. Therefore, for this research I followed up on her observation, explored it and provided evidence that, as van Dijck observed, Facebook Principles is indeed a hierarchically superior document that influences, not only the Terms of Service, but all the company's activity. It is the DNA of Facebook. In addition, and as van Dijck points out, to know the policies of companies you have to follow the track of how the data is managed. Facebook Principles may not reveal how data is managed commercially, but it shows empirically how Facebook's principles and values contribute to justifying the creation of operational mechanisms that are highly compatible with the financial needs of the company, which is to ensure safety and to ensure that the user continues to create content freely.

In addition to van Dijck's work, this dissertation aims to contribute to the work developed by Gillespie (2010, 2015 and 2018), Roberts (2019), Suzor (2018), DeNardis (2016), Owen (2019), Gorwa (2019), and Klonick (2018) on platform governance. Platform governance has focused the majority of its efforts on understanding the operational part of platforms, the different layers that make up the governance of digital platforms, and how they influence the social fabric. As such, the present dissertation contributes to the field of platform governance by uncovering the ideological dispositive on which Facebook has built all the operational techniques to govern hateful content. The empirical analysis showed that Facebook's dispositive in its entirety is a solid and ideologically coherent apparatus, where both the definition of hate speech and the techniques to operationalise and act on hateful content respond to the principles of Voices, Equity, and the mission of Security. Crucial for this has been the interpretation of what Facebook understands by safety, which is not related to users' integrity, but to the safety and integrity of Facebook structures. To elaborate this contribution, this research has focused on analysing Facebook's ideological body by reviewing Facebook Principles and by systematically analysing Mark Zuckerberg's discourse. The analysis, which has never been carried out before, confirms that Facebook governance is clearly influenced by its creator, Mark Zuckerberg, and that Facebook's foundational value is Voice or Freedom of expression.

In sum, this dissertation has revitalised Morsink's argument, has contributed to organising different approaches to hate speech, has contributed to identifying two perspectives that emerged outside the US dominance of this area, and it has contributed to platform governance by systematically analysing the principles and values underneath Facebook's operational techniques and mechanisms. Ultimately, this dissertation confirms Terranova's observation, that Foucault's studies on liberalism and neoliberalism can be applied to the digital realm (2015).

8.5 Implications for future research

Following the Black Lives Matter movement in 2020, Mark Zuckerberg announced a series of measures to meet the needs of black people in the United States. The information that exists and that Facebook has been willing to give regarding this new possible tools is limited. When I, as a researcher, contacted some of the key informants I had no success in eliciting further information. Despite this limitation, it is possible that, considering the current atmosphere of division and polarization in society, Facebook will develop projects or product solutions to manage racist content. Therefore, and if Facebook continues with this project, the question will be to analyse if these techniques will affect Facebook's ideological settings, their disciplinarian mechanism, and if they are added to the general dispositive of security.

It is also essential to investigate the work of automatic detection and artificial intelligence in the detection of hateful content. At the time of writing, automatic detection techniques are integrated into the present dispositive of security. As a result, Facebook algorithms align with Facebook principles and values as the rest of Facebook's mechanisms, therefore automatic detention does not constitute any kind of innovation, as far as the management of hate speech is concerned. It simply does what humans would do, faster and at scale. Even so, the area is already being investigated by authors related to platform governance and it would be pertinent to look at automatic detection of hateful content from an ideological point of view, in line with the research Benjamin Ruha has developed (2019).

A third avenue for future research is the study of Human Moderators as the new recipients of hate. The Soviet bloc pinpoints that women and people from the colonies were the subject of oppression and whose labour was appropriated, enabling the first accumulation of capital. As capitalism keeps advancing, the groups that are exploited vary. In the case of Facebook, the dispositive of security is built on the shoulders of thousands of anonymous reviewers who for minimum wage are exposed to hatred that is connected to them by language, origin, background, and so on. Based on the findings of this thesis, and supported by the work developed by Roberts, this thesis proposes a future study that explores how media platforms are creating new but invisible categories of workers on whose labour platforms accumulate profit.

Finally, I am interested in continuing contributing to the field of Platform Governance and Social Media studies in general by carrying out the study of platform alternatives and how they would govern hateful content, that is the main norms, values, and ideologies that might shape alternative platforms that are not limited to private interests. What are the conditions that an alternative system might employ to avoid hateful content? What historical approach, if any, would those platforms adopt? Whereas private companies like Facebook might have limited life, the technology and forms of communication are here to stay. Social media platforms are channels for expression and, because of this, the study of how they govern and how they govern hate speech in particular is still of interest. The present dissertation argues that two aspects have become pivotal to understanding the ideology and form of governance of private social media corporations: the first is their vision of equality and the second is what private corporations understand as common, i.e. what is created and shared among and between social groups. These two elements are significantly embedded in private social media policies, reflected in their flagging technique, revealing the ideology behind their model of governance, which ultimately contributes to explaining the dynamics of the circulation of hate within social media. Based on these findings, future research could undertake a similar kind of analysis on alternative platforms, proposing an examination of how alternative initiatives reformulate and redress the main issues involved in the governance of the activities of platform users, and the underlying ideology therein.

Bibliography

Ahmed, Sarah (2004) *The cultural politics of emotion*. Edinburgh University Press. Edinburgh

Adler, A. and Seligman, M. E. P; (2016). Using wellbeing for public policy: Theory, measurement, and recommendations. International Journal of Wellbeing, 6(1), 1-35. doi:10.5502/ijw. v6i1.429

Alkiviadou, Natalie (2016) Regulating Internet Hate. *JIPITEC* 216.

Allan, Richard (2017) Hard Questions: *Who Should Decide What Is Hate Speech in an Online Global Community?* Available at: https://about.fb.com/news/2017/06/hard-questions-hate-speech/. Last retrieved May 2021

Almagor, R. Cohen (1993) 'Harm Principle, Offence Principle, and the Skokie Affair', *Political Studies*, XLI(3), pp. 453-470

Altman, Andrew (1993) 'Liberalism and Campus Hate Speech: a philosophical examination'. *Ethics* 103(2), pp. 302-317

Alston, William P. (1988) 'The Deontological Conception of Epistemic Justification, Philosophical Perspectives', *Epistemology*, 2, pp. 257-299

Anne Linke, Ansgar Zerfass, (2013) 'Social media governance: regulatory frameworks for successful online communications', *Journal of Communication Management*, 17(3), pp.270-286

Ariadna Matamoros-Fernández (2017) 'Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube', *Information, Communication & Society*, 20(6), pp.930-946

Aristotle (1955) *The Ethics of Aristotle: The Nicomachean Ethics.* (rev. ed.) (J. K. Thomson, trans.). New York: Viking.

Atton Chris (2006) 'Far right media on the internet: culture, discourse and power', *New media society* 8(4), pp. 573-587

Atton Chris. (2004) Chapter *1 The internet, Power and Transgression.* An alternative Internet. Radical Media, Politics and Creativity. Edinburgh University Press.

Back, Les. (2002) 'Aryans Reading Adorno: cyber culture and twenty-first- century racism', *Ethnic and Racial Studies*, 25(4), pp. 628-651

Baez, Benjamin (2002) *Affirmative Action, Hate Speech, and Tenure: Narratives About Race and Law*. Routledge, USA

Baldwin, Sandy (2014) *The Internet Unconscious. On the Subject of Electronic Literature*. Bloomsbury Academic.

Balibar, Etienne and Ingram James (2014) *Equaliberty: Political Essays*. Duke University Press Books.

Bangstad, S. (2014) 'The weight of words: the freedom of expression debate in Norway', *Race & Class*, 55(4), pp. 8-28.

Banks, James (2010) 'Regulating Hate Speech online, International Review of Law', *Computers and Technology*, 24(3), pp. 233-239

Banks, James (2011) 'European Regulation of Cross-Border Hate Speech in Cyberspace: the limits of legislation', *European Journal of Crime, Criminal Law and Criminal Justice*, 19, pp. 1-13

Barbrook, Richard & Cameron, Andy (1996) 'The Californian ideology', *Science as Culture*, 61, pp.44-72

Bargh, John (2002) *Beyond Simple Truths: The human-internet Interaction*. The society for the Psychological Study of Social Issues

Berg, Chris (2012) *The Soviet Origins of Hate Speech*. IPA Review.

Bergami (1981) Gramsci: comunista critico [Gramsci: A Critical Communist]. *Angel* Milan,

Berger, P. L. and T. Luckmann (1966) *The Social Construction of Reality: A Treatise in the Sociology of Knowledge,* Garden City, NY: Anchor Books

Bleich, Erik (2014) 'Freedom of Expression versus Racist Hate Speech: Explaining Differences Between High Court Regulations in the USA and Europe', *Journal of Ethnic and Migration Studies*, 40(2), pp.283-300.

Bogots,Ian.,Monfort,Nick (2007). *Platform Studies: Frequently questioned answers. In Proceedings of the digital arts and culture conference*. Irvine.

Bogost, Ian and Monlfort, Nick (2009) *Racing the Beam: The Atari Video Computer System.* MIT Press.

Bory, Paolo (2020) *The Internet Myth. From the Internet Imaginary to Network Ideologies* University of Westminster Press.

Boyd, D (2004) *The social lives of networked teens*. New Haven: Yale University Press.

Boyle, Kevin (2001) 'Hate Speech-The United States Versus the Rest of the World', *Maine Law Review*, 53, pp. 487-502

Brown, Wendy (2006) 'Neoliberalis, Neoconservatism, and De-democratization', *Political Theory*, 34(6), pp. 690-714

Brown, Alexander (2017) 'What is hate speech? Part 1: The Myth of Hate', *Law and Philos* 36, pp. 419–468.

Bryman A., (2008). Social Research Methods, 3rd edition, Oxford University Press.

Bryman, A., (2012). Social research methods, 4 th edition, Oxford University Press, Oxford

Bucher, T., & Helmond, A. (2018). The Affordances of Social Media Platforms. In J. Burgess, A. Marwick, & T. Poell (Eds.), The SAGE Handbook of Social Media (pp. 233–253). Sage Publications.

Bucher, Taina (2012) 'The Friendship Assemblage: Investigating Programmed Sociality on Facebook', *Television and New Media* 14(6), pp. 479-493.

Burgess & T. Poell (Eds.) (2016) The Sage handbook of Social Media. Sage. Pre-Publication Copy. Available at: http://www.annehelmond.nl/wordpress/wp-content/uploads/2016/07/BucherHelmond_ SocialMediaAffordances-preprint.pdf Last time retrieved 30th May 2021

Butler, Judith (1997) *Excitable Speech: A Politics of the Performative*. New York: Routledge

Calleros, Charls (1995) 'Peternalism, counterspeech and Campus Hate Speech Codes: a Reply to Degado and Yun', *Arizona State Law Journal* 27, pp. 1249-128

Cambridge Dictionary (2020) Principle definition. Available at: https://dictionary.cambridge.org/dictionary/english/principle?q=Principle

Cambridge Dictionary (2021) Hate Speech definition. Available at: https://dictionary.cambridge.org/dictionary/english/hate-speech.

Cambridge Dictionary (2021) Objectionable definition. Available at: https://dictionary.cambridge.org/dictionary/english/objectionable

Cammaerts, Bart (2009) 'Radical pluralism and free speech in online public spaces: the case of North Belgian extreme right discourses', *International journal of cultural studies*, 12(6), pp. 555- 575.

Cao,Meadows, Wong, Xia (2021) 'Understanding consumers' social media engagement behaviour: An examination of the moderation effect of social media context', *Journal of Business Research*, 122, pp. 835-846

Cohen, Jodi H. & Raymond, Jennifer M. (2011) 'How the internet is giving birth to a new social order', *Information, Communication & Society*, 14(6), pp.937-957

Coleman, Beth (2009) 'Race as technology', *Camera Obscura* 24(1), p.203

Conaway, R. & Wardrope, W. (2012) 'Do their words really matter? thematic analysis of U.S. and latin american ceo letters'. In Goodwin, J. (Ed.), *SAGE biographical research* (pp. v3-61-v3-84)

Conquest, R. (1971) *The Great Terror. Stalin's Purge of the Thirties*. Harmondsworth: Penguin.

Corlet, Angelo (2005) 'Race, Racism and Reparations', *Journal of Social Philosophy*, 36(4)

Corlett, J.; Francescotti, R. (2002-2003) 'Foundations of Theory of Hate Speech', *Wayne Law Review* 48(3), pp.1071-1100

Council of Europe (1997) Recommendation No R. (97) 20 on Hate Speech. Available at: https://rm.coe.int/1680505d5b

Council of Europe (2007) Recommendation 1805. Blasphemy, religious insults and hate speech against persons on grounds of their religion. Available at: http://www.assembly.coe.int/nw/xml/XRef/Xref-XML2HTML-EN.asp?fileid=17569&lang=en

Council of Europe (2015) ECRI 15th Recommendation on combating Hate Speech. Available at: https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01

Council of Europe (2021) Values. Available at: https://www.coe.int/en/web/about-us/values. Last time retrieved 23 April 2021.

Council of Europe (2021) Toolkit Definitions. Available at: https://www.coe.int/en/web/echr-toolkit/home

Cuthbertson, Anthony. 2017. “Who controls the Internet? Facebook and Google dominance could cause the ‘Death of the Web.’” Newsweek, November 2. Available at: www.newsweek.com/facebook-google-internet-traffic-net-neutrality-monopoly-699286.

Cheney-Lippold, John (2017) *We are data. Algorithms and the making of our digital selves.* New York University Press

Daniels, Jessie (2008) Race, Civil Rights, and Hate Speech in the Digital Era. *Learning Race and Ethnicity: Youth and Digital Media*. Edited by Anna Everett. The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning. Cambridge, MA: The MIT Press, 129–154

Darley, J. M., & Latane, B. (1968) ‘Bystander intervention in emergencies: Diffusion of responsibility’, *Journal of Personality and Social Psychology*, 8(4, Pt.1), pp.377–383.

Darley, J. M., & Latane, B. (1970) *The unresponsive bystander: why doesn't he help?* New York, NY: Appleton Century Crofts.

Delgado & Yun (1994) 'The Neo conservative Case Against Hate-speech Regulation-Lively D'Souza, Gates, Carter and the Thought Love Crowd', *Venderbilt Law Review* 47, pp 1808-1825

Delgado, Richard & Stefancic Jean (2018) *Must we defend Nazis?* New York University Press.

Demchenko, C. de Laat and P. Membrey (2014) 'Defining architecture components of the Big Data Ecosystem', *2014* International Conference on Collaboration Technologies and Systems (CTS), Minneapolis, MN, 2014, pp. 104-112

DeNardis, L. (2014) *The Global War on Internet Governance*. New Haven: Yale University Press.

DeNardis, L & Hackl, A. (2015). Internet Governance by Social Media Platforms. *Telecommunications Policy.* Advance online publication. Available at: http://www.sciencedirect.com/science/article/pii/S0308596115000592

Deutche Welle (2015) Facebook's Zuckerberg promises Merkel action on hate speech, 27 September. Available at: http://www.dw.com/en/facebooks-zuckerberg-promises-merkel-action-on-hate-speech/a-18744036

Dickinson, J.J., Poole, D.A (2000) 'Efficient coding of eyewitness narratives: A comparison of syntactic unit and word count procedures', *Behaviour Research Methods, Instruments, & Computers* 32, pp.537–545

Dreyfus, Hubert and Rabinow, Paul (1982) Michel Foucault: Beyond Structuralism and Hermeneutics. Chicago: The University of Chicago Press

Elkin-Koren, Perel Maayan (2020) *Guarding the Guardians: Content Moderation by Online Intermediaries and the Rule of Law*. Oxford Handbook of Online Intermediary Liability.

Ehrlich, Howard. (1962) 'The Swastika Epidemic of 1959-1960: Anti-Semitism and Community Characteristics', *Social Problems, 9*(3), pp.264-272.

Ehrlich Howard and Scimecca, Joseph (1991) Offensive Speech on Campus: Punitive or Educational Solutions? Educational Record N72 pp26-29

Essed, Philomena (1991) *Understanding Everyday Racism*. SAGE

European Commission (2020) Available at: The EU Code of conduct on countering illegal hate speech online. Last time retrieved 28 May 2021.

European Commission (1999) 'Action plan of the Council and the Commission on how best to implement the provisions of the Treaty of Amsterdam on an area of freedom, security and justice', *Official Journal of the European Communities*, 42.

European Union. (2020) 5th evaluation of the Code of Conduct. Fact sheet. Available at: https://ec.europa.eu/info/sites/default/files/codeofconduct_2020_factsheet_12. Last accessed 28/05/2021

European Commission (2021) Questions and Answers on Tobacco Advertisement. Available at: https://ec.europa.eu/commission/presscorner/detail/en/MEMO_01_205.

European Commission (2016) European Code of Conduct. Available at: https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en.

European Union (2020) Hate speech and hate crime in the EU and the evaluation of online content regulation approaches. Available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/655135/IPOL_STU(2020)655135_EN.pdf.

Facebook (2020) Hate speech definition. Available at: https://www.facebook.com/communitystandards/hate_speech.

Facebook (2020) Hateful groups Policy______Available at: https://www.facebook.com/communitystandards/hate_speech]

Facebook (2021) Product Policy Forum Minutes____ Available at: https://about.fb.com/news/2018/11/content-standards-forum-minutes/ Last Time retrieved 31st March 2021.

Facebook (2021) Recommendation Harmful Stereotypes. Available at: https://about.fb.com/wp-content/uploads/2018/11/PPF_08.11.2020_Harmful-Stereotypes.pdf

Facebook Design Principles (2021) Available at: https://m.facebook.com/nt/screen/?params=%7B%22note_id%22%3A693790757917796%7D&path=%2Fnotes%2F%7Bnote_id%7D&refsrc=http%3A%2F%2Fwww.google.com%2F&_rdr

Facebook (2021) Oversight Board. Available at: https://oversightboard.com/

Facebook (2021) Oversight Board Chart. Available at: https://oversightboard.com/governance/

Facebook (2021) Investor Report. Available at: https://investor.fb.com/investor-news/press-release-details/2021/Facebook-Reports-First-Quarter-2021-Results/default.aspx

Federicci, Silvia (2004) *Caliban and the Witch*. UK. Autonomedia.

Feenberg Andrew (2005) 'Teoría crítica de la tecnología', *Revista CTS*, 5(2), pp.109-123

Feenberg, Andrew (2017) A Critical Theory of Technology. *Handbook of Science and Technology Studies*. Ed. Ulrike Felt, Rayvon Fouché, Clark A. Miller, Laurel Smith-Doerr, eds., MIT Press, pp. 635-663.

Feengber, Andrew (2003) *Modernity Theory and Technology Studies. Reflections on Bridging the gap*. MIT Press, Massachusetts.

Feenberg, Andrew (2010) The Critical Theory of Technology *and Values*. Blackwell. UK

Feldman, Steven P (1999) 'The levelling of organizational culture: Egalitarianism in critical postmodern organization theory', *The Journal of Applied Behavioural Science*, 35(2), pp. 228-244

Finlay, Andrew (2015) Liberal intervention, anthropology and the ethnicity machine, *Peacebuilding* , 3, (3)

Fish, Stanley (1994) *There is not such a thing as Free Speech, and it is a good thing too.* Oxford University Press. NY

Fisher, Eran and Christian Fuchs (2015) *Reconsidering Value and Labour in the Digital Age*. Basingstoke: Palgrave Macmillan. UK.

Foucault, Michele (2007) *What is Critique? The Politics of Truth*. Semiotext(e), Los Angeles.

Foucault, Michele (2002) *The archaeology of knowledge.* Routledge. New York

Foucault, Michel (2007) *Security, Territory , Population. Lectures at the collage of France 1977-1978.* Palgrave Mc Millan. New York.

Froio, Catherina & Bharath Ganesh (2019) 'The trans nationalisation of far right discourse on Twitter', *European Societies,* 21(4), pp. 513-539

Fuchs, Christian (2014) *Social Media a critical introduction*. SAGE. UK

Fukuyama, Francis (1992) *The End of History and the Last Man*. New Press.

García Gloria (2004) 25 años de libertad de expresión. VII Congreso de la Asociación de Historiadores de la Comunicación. Barcelona, 18-19 de noviembre de 2004

Genova, Nicholas De (2015) In the land of the Setting Sun. Reflections on Islamization and Patriotic Europeanism. Available at: http://movements-journal.org/issues/02.kaempfe/15.de-genova--pegida-islamization-patriotic-europeanism.html (Last Accessed on the 26th of February 2016.

Gerrard, Ysabel (2018) 'Beyond the hashtag: Circumventing content moderation on social media', *New Media & Society* 20.

Gibbs, Annette (1992) Reconciling Rights and Responsibilities of Colleges and Students: Offensive Speech, Assembly, Drug Testing, and Safety ASHE-ERIC Higher Education Report No. 5. Available at G. https://files.eric.ed.gov/fulltext/ED354837.pdf

Gillespie, Tarleton (2010) 'The politics of "platforms."' *New Media & Society*, 12(3), pp.347–364.

Gillespie, Tarleton (2015) Facebook's improved "Community Standards" still can't resolve the central paradox. *Social Media Collective*. Available at:

https://socialmediacollective.org/2015/03/18/facebooks-improved-community-standards-still-cant-resolve-the-central-paradox/

Gillespie, Tarleton (2015) 'Platforms intervene', *Social Media + Society*, 1, pp.1–2.

Gillespie, Tarleton (2018) *Custodians of the Internet, platforms, content moderation, and the hidden decisions that shape social media*. Yale University. USA.

Gillespie, Tarletion. (2020) 'Content moderation, AI, and the question of scale', *Big Data & Society*.p 1-5

Gilliard, Chris (27th August 2020) Facebook's Technocratic Reports Hide Its Failures on Abuse. One Zero. Available at: https://onezero.medium.com/facebook-is-hiding-its-failure-to-keep-abuse-off-its-platform-behind-technocratic-reports-682d871ef1ca

Ging, Debbie and Siapera, Eugenia (2019) *Gender Hate Online: Understanding the New Anti-Feminism*, Palgrave Macmillan

Glassman, Jim (2006) 'Primitive accumulation, accumulation by dispossession, accumulation by extra-economic means', *Progress in Human Geography*, 30(5), pp. 608-625

Gomberg Paul (1990) 'Patriotism is like Racism', *Ethics*, 101(1), pp. 144-150

Gorwa, Robert (2019) 'What is platform governance?', *Information, Communication & Society*, 22(6), pp. 854-871

Gorwa R, Binns R, Katzenbach C. (2020) 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance', *Big Data & Society*

Griffin, Roger (1991) *The Nature of Fascism*. Routledge, London.

Grimmelmann J. (2015) 'The virtues of moderation', *Yale Journal of Law* 17: 42

Guardian The (2017) Facebook Files. Available at: https://www.theguardian.com/news/series/facebook-files

Guba, E. G., & Lincoln, Y. S. (1994) *Competing paradigms in qualitative research.* In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (p. 105–117). Sage Publications, Inc

Giuseppe Gabrielli & Roberto Impicciatore (2021) Breaking down the barriers: educational paths, labour market outcomes and wellbeing of children of immigrants, Journal of Ethnic and Migration Studies, https://doi.org/10.1080/1369183X.2021.1935655

Harvey, David (2016) Neoliberalism is a political project. Jacobin magazine. Available at: https://www.jacobinmag.com/2016/07/david-harvey-neoliberalism-capitalism-labor-crisis-resistance/

Haupt, Joachim (2021) 'Facebook Futures: Mark Zuckerberg's discursive construction of a better world', *New Media & society* 23(2), pp.237-257

Heinze, Eric (2006) *Viewpoint Absolutism and Hate Speech. The Modern Law Review* Blackwell Oxford.

Heinze, Eric (2013) 'Hate Speech and the normative foundations of Regulation', *International Journal of Law in Context*, 4, pp. 590-617.

Heinze, Eric (2016) *Hate Speech and Democratic Citizenship*. Oxford University Press. UK

Hestres, L. E. (2013) 'App Neutrality: Apple's App Store and Freedom of Expression Online', *International Journal of Communication*, 7, pp.1265–1280

Hibou, Béatrice. (2015). The Bureaucratization of the World in the Neoliberal Era. Palgrave.

Hirschmann, Albert O (1978) 'Exit, Voice and The State', *World Politics* 31(1), pp.90-107

Hoadley, Christopher M, Xu Christopher M, Lee Joey J, Rosson, Mary Beth (2010) 'Privacy as information access and illusory control: The case of the Facebook News Feed privacy outcry', *Electronic Commerce Research and Applications*, 9(1), pp. 50-60

Hobsbawm, Eric (1996) Identity Politics and The left. NLR I/127, May-June

Hoffmann AL, Proferes N and Zimmer M (2018) '"Making the world more open and connected": Mark Zuckerberg and the discursive construction of Facebook and its users', *New Media & Society* 20(1), pp. 199-218

Independent (2019) Forres, Adam. New Zealand attack: Facebook says it has removed 1.5 million videos of Christchurch massacre. Available at: https://www.independent.co.uk/news/world/australasia/new-zealand-attack-facebook-videos-footage-removed-christchurch-a8826921.html.

Iversen, Stefan (2017). Narratives and Online Decorum: The Rhetoric of Mark Zuckerberg's Personal Storytelling on Facebook. *Style, 51*(3), 374-390

Jackson, B. F. (2014) 'Censorship and freedom of expression in the age of Facebook', *New Mexico Law Review*, 44(1), pp.121-168.

Jane, Emma A. (2018) 'Gendered cyberhate as workplace harassment and economic vandalism', *Feminist Media Studies*, 18(4), pp. 575-591

Jager, Siegfried and Maier, Florentine (2009) *Theoretical and Methodological aspects of Foucauldian critical discourse analysis and dispositive analysis.* Ed. Ruth Wodack and Michael Meyer. *Methods for Critical Discourse Analysis:* 34 SGAE. UK

Jinkyung Na, Michal Kosinski, and David J. Stillwell (2014) 'When a New Tool Is Introduced in Different Cultural Contexts: Individualism–Collectivism and Social Network on Facebook', *Journal of Cross-Cultural Psychology*, 46(3), pp. 355–370

Jinkyung Na, Michal Kosinski, and David J. Stillwell (2014) 'When a New Tool Is Introduced in Different Cultural Contexts: Individualism–Collectivism and Social Network on Facebook', *Journal of Cross-Cultural Psychology*, 46(3), pp. 355–370

Jubany Olga, Roiha Malin (2018) *Las Palabras son armas. Discurso de odio en la red*. Universidad de Barcelona

Keller R (2011) 'The sociology of knowledge approach to discourse (SKAD)', *Human Studies* 34(1), pp. 43–65.

Keller R, Hornidge A and Schunemann WJ (2018) *The Sociology of Knowledge Approach to Discourse: Investigating the Politics of Knowledge and Meaning-Making*. New York: Taylor & Francis

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

Kotljarchuk, Andrej and Sundström, Olle (2017) *Ethnic and Religious Minorities in Stalin's Soviet Union*. New Dimensions of Research. Sodertorn University

Kotljarchuk, Andrej (2017) Propaganda of Hatred and the Great Terror. A Nordic Approach. in Ethnic and Religious Minorities in Stalin's Soviet Union: New Dimensions of Research / [ed] Andrej Kotljarchuk; Olle Sundström, Huddinge: Södertörns högskola p. 91-121

Kreindler, Isabelle (1977) 'A Neglected Source of Lenin's Nationality Policy', *Slavic Review*, 36(1), pp. 86-100.

Keyes CL.M.;. (1998). Social Well-Being. *Social Psychology Quarterly*, *61*(2), 121–140. https://doi.org/10.2307/2787065

Langlois,Ganaele. (2013) 'Participatory Culture and the New Governance of Communication: The Paradox of Participatory Media', *Television & New Media*, 14(2), pp. 91–105

Lawyers Co-operative publishing Company. (1998). *American jurisprudence, second edition. a modern comprehensive text statement of American law, state and federal ; historical and legal documents, facts, tables, charts, and statistics of special interest to attorneys*. Rochester, N.Y., Lawyers Co-operative Pub.

Le Marche, Gara Ed. (1996) *Speech & Equality: Do We Really Have to Choose?* New York: New York University Press.

Livingstone, S and Brake, D.R (2010) 'On the rapid rise of social networking sites: new findings and policy implications', *Children and Society*, 24(1), pp. 75-58

Livingstone, Sonia (2008) 'Internet literacy: young people's negotiation of new online opportunities' in T.McPherson (ed) *Digital youth, innovation and the unexpected,* Cambridge: MIT Press. pp.101-122

McKinnon, Catherine (1987) Feminism Unmodified: Discourses on Life and Law. Cambridge, MA: Harvard University Press.

Magalhães João Carlos and Katzenbach, Christian (2020) Emerging Structures of Platform Governance and Copyright. Methods and Challenges in Studying Content Policies

Public and private regulatory framework of online intermediaries. Presentation. University of Szeged, Szeged, Hungary: 05.05.2020

McGonagle, Tarlach and Donders, Yvonne (2015) *The United Nations and Freedom of Expression and Information: Critical Perspectives*. Cambridge University Press.

Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto (2017) A Measurement Study of Hate Speech in Social Media. In Proceedings of the 28th ACM Conference on Hypertext and Social Media (HT '17). Association for Computing Machinery, New York, NY, USA, 85–94

Marcuse, Herbert (1993) *El hombre unidimensional. Ensayo sobre la ideología de la sociedad industrial avanzada.* Planeta-Agostini. Spain

Martin, T. (2001) *The Affirmative Action Empire. Nations and Nationalism in the Soviet Union, 1923–1939*. Ithaca: Cornell University Press.

Marx, Karl (1909) *Capital Volume One*.

Matamoros-Fernández, Ariadna (2017) 'Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube', *Information, Communication & Society*, 20(6), pp.930-946

Matsuda, M. J. (1993) 'Public response to racist speech: Considering the victim's story'. In M. J. Matsuda, C. R. Lawrence, R. Delgado, & K. W. Crenshaw (Eds.), Words that wound: Critical race theory, assaultive speech, and the first amendment . Westview Press San Francisco. pp. 17-52.

M. J. Matsuda, C. R. Lawrence, R. Delgado, & K. W. Crenshaw (1993) *Words that wound: Critical race theory, assaultive speech, and the first amendment*. Westview Press. San Francisco.

McGonagle, Tarlach . (2012). Minorities and Online 'Hate Speech': A Parsing of Selected Complexities, *European Yearbook of Minority Issues Online, 9*(1), 419-440.Mchagama Jacob (2015) The problem with hate speech laws. The review of Faith & International Affairs 13:1, 75-82 Routledge.

Mchangama, Jacob (1st Decemeber 2011) The Sordid Origin of Hate Speech Laws. Available at https://www.hoover.org/research/sordid-origin-hate-speech-laws

Mies, Maria. (1986) *Patriarchy & Accumulation on a World Scale. Women in the international division of labour.* Zed Books Ltd. New York.

Miller, Peter and Ross, Nicolas (2008) *Governing the present.* Polity Press. UK

Milosevic, Tijana. (2016) 'Social Media Companies' Cyberbullying Policies', *International Journal Of Communication, 10*, 22.

Milosevic, Tijana. (Spring, 2018) *Cyberbullying Policies of Social Media Companies*. MIT Press

Mir R, Watson A. (2000). Strategic management and thephilosophy of science: the case for a constructivistmethodology. *Strategic Management Journal* 21(9):941–953

Monasta, A. Antonio Gramsci. Prospects 23, 597–612 (1993). https://doi.org/10.1007/BF02195137

Morozov, Evegeny (2013) *To save everything, click here*. Penguin books. USA.

Morsink, Johanes (1999) *The universal Declaration of Human Rights. Origins, Drafting and Intent*. Pennsylvania Press. US

Myers West S (2018) 'Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms', *New Media & Society* 20(11), pp.4366-4383.

Newman Stephen L (2010) 'Should Hate Speech be allowed on the Internet? A reply to Raphael Cohen-Almagor', *Amsterdam Law Forum*, 1(2), pp.119-124

O'Flaherty, Michael (2012) 'Freedom of Expression: Article 19 of the International Covenant on Civil and Political Rights and the Human Rights Committee's General Comment No 34', *Human Rights Law Review*, 12(4), pp. 627–654

O'Flaherty, Michael (2020) Freedom of Expression: Article 19 of the International Covenant on Civil and Political Rights and the Human Rights Committee's General Comment

No. 34, 12 Hu,.Rights L.Rev HUM. RIGHTS L. REV. 627 *Communication Law and Policy,* 25:4, 483-486

O' Reilly, T (2005) 'Web 2.0: compact definition? 1 October, O'Reilly radar. Available at: http://radar.oreilly.com/2005/10/web-20-compact-definition.html

O'Neill Cathy (2017) *Weapons of math destruction.* Penguin. UK

Oksanen, Atte & Hawdon, James & Holkeri, Emma & Näsi, Matti & Räsänen, Pekka. (2014). Exposure to Online Hate among Young Social Media Users.in M Nicole Werehime (ed) *Sould of Society: a Focus on the Loves of Children & youth (Sociological Studies of Children and Youth, Volum 18)* Emerald Group Publishing Limited, pp.253-273.

Óscar Pérez de la Fuente (2010) Libertad de Expresión y el caso del Lenguaje del odio. Una aproximación desde la perspectiva norteamericana y perspectiva alemana. Cuadernos Electrónicos de Filosofía del Derecho. Espania. Available at http://e-archivo.uc3m.es/handle/10016/12085.

Óscar Pérez de la Fuente (2010) Libertad de Expresión y el caso del Lenguaje del odio. Una

Owen, Tylor (2019) The Case of Platform Governance. CIGI Paper No. 231 Available at: https://www.cigionline.org/publications/case-platform-governance.

Paslawsky, Alexandra. (2012) 'The Growth of Social Media Norms and Government's Attempts at Regulation', *Fordham International Law Journal,* 35, pp. 1485-1542.

Patelis, Korinna (2013) 'Facebook.com text: Industrialising personal data production', *First Monday,* 18(3-4). Available at: http://journals.uic.edu/ojs/index.php/fm/article/view/4615/3424doi:10.5210/fm.v18i3.4615 Last access 11/01

Pincus L. Fred and Ehrlich, J Howard (1994) *Race and Ethnic Conflict. Contending Views on Prejudice, Discrimination and Ethnoviolence.* Routledge

Pincus, F. L. (1996) 'Discrimination Comes in Many Forms: Individual, Institutional, and Structural', *American Behavioral Scientist*, 40(2), pp. 186–194

Platform Governance Archive (2021). Available at: https://pga.hiig.de/explore?fbclid=IwAR2ezo9WTamMgPAKuMhZLfU2IyBM7Kj3akEd_TgH52trTU-ytn0wwgx02sc

Post,Robert (1991) 'Racist Speech, Democracy, and the First Amendment, *William and Mary Law Review* 32, pp. 290-91.

Prior, Lindsay (2008) 'Repositioning Documents in Social Research', *Sociology*, 42, pp.821-836.

Prilleltensky (2003) 'Understanding, Resisting, and Overcoming Oppression: Toward psychopolitical validity', *Am J Community Psychol* 31, pp.195-201.

Quattrociocchi, Walter and Scala, Antonio and Sunstein, Cass R., Echo Chambers on Facebook (June 13, 2016). Available at: http://dx.doi.org/10.2139/ssrn.2795110

Rabban, David. (1994). The IWW Free Speech Fights and Popular Conceptions of Free Expression before World War I. *Virginia Law Review*, *80*(5), 1055-1158.

Raffnsøe, Sverre & Thaning, Morten & Gudmand-Høyer, Marius. (2014). What is a dispositive? Foucault's historical mappings of the networks of social reality. Organization. 10.1177/1350508414549885).

Roberts, Sarah T (2016) Commercial Content Moderation: Digital Labourers' Dirty Work Media Studies Publications. P 12.

Roberts, S. T (2019) B*ehind the Screen*. Yale University Press

Ruha, Benjamin (2019) *Race after technology*. Polity Press

Saavedra, Modesto (2005) El Lenguaje del Odio en la Jurisprudencia del Tribunal Constitucional Español. *Persona y Derecho vl55 547-576*

Sessen, S. (2002) 'Towards a Sociology of Information Technology', *Current Sociology*, 50(3), pp. 365–388.

Shannon, Claude E. 1948 *A Mathematical Theory of Communication*, Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656

Shiell , Timothy C. (2009). *Campus Hate speech on Trial.* University Press of Kansas

Siapera, Eugenia (2018). Understanding New Media. SAGE. UK

Siapera, Moreo, Zhu, Hate Speech Track. HateTrack-Tracking-and-Monitoring-Racist-Hate-Speech-Online. Retrieve from https://researchrepository.ucd.ie/handle/10197/9916 Last time retreive 13 Jannuary 2021.

Siapera, Eugenia; Viejo Otero, Paloma (2021) Governance hate. Facebook and Digital Racism. Television and New Media. SAGE

Siapera, Eugenia; Viejo Otero, Paloma (2021) Content Moderation and External Stakeholders: Platform Governance and the Role of Experts. First Platform Governance Conference.

Skinner, Christina Parajon "Unprofessional Sides of Social Media and Social Networking: How Current Standards Fall Short," South Carolina Law Review 63, no. 2 (Winter 2011): 241-284

Sorial, Sarah, (2015) Hate speech and Distorted Communication: Rethinking the limits of incitement. Law and Philosophy 34: 299-324

Springer, Simon 2007 A Brief History of Neoliberalism. Journal of Peace Research 44; 126

Srnicek Nick (2017) Platform Capitalism. Polity Press. UK

Staub Ervin. (2004) The origins and Evolution of Hate, With Notes on Prevention. In Steinberg Robert J ed. *The psychology of hate.* UK American Psychology Association.

Sternberg and Sternberg (2008) The nature of hate. Cambridge University Press.

Sterne Jonathan.(1999) *Thinking the Internet. Cultural Studies Vs Milennium* on Steves Jones Ed Doing Internet Research . Critical Issues and methods for Examining the Net. Thousands Oaks, CA and London SAGE Publications

Strossen, (1990) Nadine"Regulating Racist Speech on Campus: A Modest Proposal?" Duke Law Journal pp. 557-58.

Stuart Mill, John (1859) *On Liberty* . Batoche books. Canada

Sunstein , Cass (2000) Ideas yes, Assaults No. The First Amendment protects the exchange of ideas, not verbal assaults. Available at https://prospect.org/culture/ideas-yes-assaults/ (last accessed 26 February 2016)

Suzor, Nicholai (2018) Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms. Social Media and Society. Volume 4 Issue 3

Terranova Tiziana. (2015)Securing the Social: Foucault and Social Networks. In Fuggle S., Lanci Y., Tazzioli M. eds Foucault and the History of Our Present. Palgrave Macmillan, London

Terranova, Tizianav(2017) *Platform Capitalism and the Government of the Social. Facebook's 'Global Community'* Retrieved from http://www.technoculture.it/en/2017/02/21/platform-capitalism-and-the-government-of-the-social/ Last access 28/06. No longer available. See Appendix IV.

Thornberry, P. (1989). Self-Determination, Minorities, Human Rights: A Review of International Instruments. *The International and Comparative Law Quarterly, 38*(4), 867-889

Titley, Gavan (2016) *The debatability of Racism* . Networked pariticipative media and postracialsim Retrieve from http://raster.fi/2016/02/17/the-debatability-of-racism-networked-participative-media-and-postracialism/ Last Access 02/2016.

Titley, Gavan (2020) *Is free speech racist?* Polity

Tiwana et al (2010) Research Commentary: Coevolution of Platform Architecture, Governance, and Environmental Dynamics Information Systems Research 21(4), pp. 675–687.

Tuhiwai Smith, Linda (2012) *Decolonising methodologies: research and indigenous peoples,* second edition, Zed Books. London.

Tussey, Deborah (2014). Facebook: The New Town Square, 44 Sw. L. Rev. p 385

Tynes (Eds.), The intersectional Internet. *Race, sex, class and culture online* (pp. 147–160). New York, NY: Peter Lang.

Tynes, B. (2006). 'Children, adolescents, and the culture of *online* hate' in N.Dowd, D.Singer y R.F. Wilson (eds) *Handbook of children, culture, and violence. Pp 267-289.* California. *SAGE.*

UN General Assembly (1946) . Resolution 59 (1) *Calling of an International Conference on Freedom of Information* . Retrieved from *https://undocs.org/en/A/RES/59(I)*. Last time retrieve 10 May 2021

UN General Assembly (21 December 1965) *International Convention on the Elimination of All Forms of Racial Discrimination*, , United Nations, Treaty Series, vol. 660, p. 195, available at: https://www.refworld.org/docid/3ae6b3940.html . Last accessed 2 June 2021

UN. (1998). *The Universal Declaration of Human Rights, 1948-1998*. New York. United Nations Dept. of Public Information.

U.S. Patents. (2010) US 7,669,123 B2 by Zuckerberg, Sanghvi, Bosworth Cox Sittig Hughes Geminder Corson inventors. News Feed. Retrieved from https://patentimages.storage.googleapis.com/ba/9d/0c/10890aa849eff0/US7669123.pdf . Last time I retrieved 30th May 2021.

United States. (2021)230: Protection for private blocking and screening of offensive material. Retrieved from https://uscode.house.gov/view.xhtml?req=(title:47%20section:230%20edition:prelim), last Time retrieve 28/05/2021

Van Berkum, Jos J.A. Holleman Bregje, Nieuwland Mante, Otten Marte, Murre, Jaap (2009) Right or Wrong?: The Brain's Fast Response to Morally Objectionable Statements. Psychological Science Vol 20, Issue 9,

van Dijk TA (1998) Editorial: Discourse and Ideology. *Discourse & Society*. 1998;9(3):307-308.

van Dijck, Jose and Poell (2013) Understanding Social Media Logic Media and Communication, Vol. 1, Issue 1, pp. 2-14, van Dijck, Jose. (2013). *The culture of connectivity: A critical history of social media*. Oxford University Press.

Vasiuc A; (2019) Sharing the Insecure Sensible: The Circulation of Images of Roma on Social Media. In: van Baar H., Ivasiuc A., Kreide R. (eds) The Securitization of the Roma in Europe. Human Rights Interventions. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-319-77035-2_11

Verge The (2019) The trauma Floor. The secret live of FAcebook moderators in America. by Newton, Casey. Retrieve from. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizon.Last Time retrieve 28/05/2021

Venturiello, María Pía; Gómez Bueno, Carmuca; Martín Palomo, María Teresa (2020). Entramados de interdependencias, cuidados y autonomía en situaciones de diversidad funcional. *Papeles del CEIC*, *vol. 2020/2*, papel 234, 1-18. http://dx.doi.org/10.1387/pceic.20940.

Viejo Otero, Paloma (2020) Guide to understand Hate Speech. Eskura.

Viejo Otero, Paloma (2021) How the Donald Trump years have changed Facebook. Onhttps://www.rte.ie/brainstorm/2021/0119/1190648-donald-trump-mark-zuckerberg-facebook/ Last time retrieve 19/01/2021

Waller and Anderson (2019) Generalists and Specialists: Using Community Embeddings to Quantify Activity Diversity in Online Platforms. on WWW '19, May 13–17, 2019, San Francisco, CA, USA.

Walker, Samuel (1991) *Hate Speech. The History of an American Controversy. University Nebraska Press. USA*

Weil, Simone (2005) *An Anthology*. Penguin Classics

Werth, N. (2003). *The mechanism of a mass crime. The Great Terror in the Soviet Union, 1937–38," in The Specter of Genocide*. Mass Murder in Historical Perspective, eds. R. Gellately & B. Kiernan, Cambridge: Cambridge University Press, pp. 215–239

White, Edward G. (1996) The First Amendment Comes of Age: The Emergence of Free Speech in Twentieth-Century America. Michigan Law Review, Vol. 95, No. 2 (Nov., 1996), pp. 299-392

Willson, Robert E, Gosling, Samuel D, Graham, Lindsay T. (2012) A review of Facebook Research in the Social Sciences APS . Perspectives on Psychological Science 7(3) 203-220

Wodack, Ruth and Meyer, Michael (2008) Methods of Critical Discourse Analysis. SAGE

Zuckerberg et al (2006) News Feed Patent, Available at: https://patentimages.storage.googleapis.com/ba/9d/0c/10890aa849eff0/US7669123.pdf last time retrieve 17 January 2021.

Zuckerberg, Mark, (19 November2016) "Mark Zuckerberg at Apec CEO Summit. Available at: https://www.facebook.com/zuck/videos/10103270382349351. Last access 15/05/2021

Zuckerberg, Mark, (18 November 2016) " Heading Out to Peru. 'Our community " (2016). Available at: https://www.facebook.com/zuck/posts/10103268839201831

Zuckerberg, Mark (1st February 2017) Founders Letter. Missions and Values. Available at: https://www.facebook.com/notes/261129471966151/. Last access 15/05/2021

Zuckerberg, Mark (17 February 2017) Building a Global Community. Available at: https://www.facebook.com/notes/mark-zuckerberg/building-global-community/10154544292806634/ Last access15/05/2021

Zuckerberg, Mark (15 November 2018)Blueprint for Governance and enforcement. Available at: https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634 Last access 15/05/2021

Zuckerberg, Mark (5ht of May 2020) binary system of governance:__ Available at: https://www.facebook.com/zuck/posts/10111985969467901

Zuckerberg, Mark, "MZ post announcing a new Facebook logo" (2019). Transcripts of Zuckerberg. 1064. Available at: https://epublications.marquette.edu/zuckerberg_files_transcripts/1064

Zuckerberg, Mark. (17 October 2019) Mark Zuckerberg Stands for Voice and Free Expression, Available at: https://about.fb.com/news/2019/10/mark-zuckerberg-stands-for-voice-and-free-expression/. Last retrieve 19/01/2021

Zuckerberg, Mark (28th April 2021) on Business solutions. Available at: https://www.facebook.com/zuck/posts/10112931311844201. Last time retrieved 31st May 2021.

Zuckerberg, Mark (4th November 2019) "MZ post announcing new Facebook logo". Available at: https://www.facebook.com/zuck/videos/10110362239999771/https://www.facebook.com/zuck/videos/10110362239999771/ Last retrieve 18/05/2021

Zuckerberg, Mark (1st April 2021) On Covid Vaccine Available at: https://www.facebook.com/zuck/posts/10112878091987221. Last time retrieve 31/05/2021

Zuckerbrg, Mark (19th Novemebr2016) APEC Peru Summit. On Connecting. Available at: https://www.facebook.com/zuck/videos/10103270382349351/.Last time retrieve 31/05/2021

Zuckerberg, Mark (6 November 2016) Connectivity Lab. Available at: https://www.facebook.com/zuck/posts/10103264282762971.Last time retrieve 31/05/2021

Zuckerberg, Mark (2017) "Announcing Facebook Communities Summit" . Available at: https://www.facebook.com/zuck/posts/10103777403435031.Last time retrieve 31/05/2021

Zuckerberg, Mark (22nd June 2017) Bring the World together. Available at: https://www.youtube.com/watch?v=RYC7nAcZqn0.Last time retrieve 31/05/2021

Zuckerberg,Mark, (27th September 2017) Responding to Trump. Available at: https://www.facebook.com/zuck/posts/10104067130714241.Last time retrieve 31/05/2021

Zucerberg Mark, Founders Letter 2012, Five years late (1st February 2017) Available at: https://www.facebook.com/zuck/posts/10154500412571634.Last time retrieve 31/05/2021

Zuckerberg, Mark (28th December 2018) 2018 Personal Challenges. Available at: https://www.facebook.com/zuck/posts/10105865715850211

Zuckerberg, Mark, (6ht May 2020) "MZ post about launching independent Oversight Board" Available at: https://www.facebook.com/zuck/posts/10111886600788531

Appendices

Appendix I: **Mark Zuckerberg post titles**

TITLE	2016	LINK
MZ's Favorite Moments from 2016	31-12-2016	https://www.facebook.com/zuck/posts/10103377702962981
Jarvis: AI's Perspective	21-12-2016	https://www.facebook.com/zuck/videos/10103353413344571/
Video Sheryl reflecting on 2016.	21-12-2016	https://www.facebook.com/zuck/videos/10103353645165001/
FB's projects aimed at fighting misinformation	12_12_2016	https://www.facebook.com/zuck/posts/10103338789106661
MZ's version of the FB feature, Year in Review	12-12-2016	https://www.facebook.com/zuck/videos/10103330600217261/
Partnership FB and "community groups". Homlessness	2-12-2016	https://www.facebook.com/zuck/posts/10103297997238891
Peru discussing connectivity programs. Political leaders.	19-11-2016	https://www.facebook.com/zuck/posts/10103271832263711
Apec CEOs Summit Peru. Video.	19-11-2016	https://www.facebook.com/zuck/videos/10103270382349351/
Heading Out . 'Our community'	18-11-2016	https://www.facebook.com/zuck/posts/10103268839201831
Combating misinformation	18-11-2016	https://www.facebook.com/zuck/posts/10103269806149061
Social Good Forum	18-11-2016	https://newsroom.fb.com/news/2016/11/facebooks-social-good-forum/
Connectivity Lab	17-11-2016	https://www.facebook.com/zuck/posts/10103264282762971
MZ's thoughts on the 2016 elections, post-results.	16-11-2016	https://www.facebook.com/zuck/posts/10103253901916271
Elections, News science, AI and Techonomy	11-11-2016	https://www.facebook.com/zuck/videos/10103248351713921/
Election 2016. Feeling Hopeful	9-11-2016	https://www.facebook.com/photo.php?fbid=10103245877492281&set=a.529237706231.2

Voting registration at the 2016 Elections	6-11-2016	https://www.facebook.com/zuck/posts/10103235600517421
2016 Earning Calls	2-11-2016	http://edge.media-server.com/m/p/ho5sxhxp
community's progress and video	2-11-2016	https://www.facebook.com/zuck/videos/10103225611545401/
Woman in Technology	21-10-2016	https://www.facebook.com/photo.php?fbid=10103193021491131&set=a.529237706231.2
FAcebook in 100 Lenguage	30-9-2016	https://www.facebook.com/zuck/videos/10103141435699471/?permPage=1
Video with News Feed original Team	06-07-2016	https://www.facebook.com/zuck/videos/10103087013971051/
Post on Social feeds	06-07-2016	https://www.facebook.com/zuck/posts/10103087138471551
Post on Social Feeds 10 anniversary	02-07-2016	https://www.facebook.com/zuck/posts/10103084921703971
Conference University of Rome	29-08-2016	https://www.youtube.com/watch?v=UohGRvkzT9Y
How to build a future. Interview	16-08-2016	https://www.youtube.com/watch?v=Lb4IcGF5iTQ
FB's newest advanced hardware lab	03-08-2016	https://www.facebook.com/zuck/posts/10103010090805691
Earning's call 2016	27-07-2016	http://edge.media-server.com/m/p/s4g5rjyo
Facebook users increasing	27_02_2016	https://www.facebook.com/zuck/posts/10102994874459371
Facebook messangers users	20-07-2016	https://www.facebook.com/zuck/videos/10102977526035751/
LGBT celebration	24-06-2016	https://www.facebook.com/photo.php?fbid=10102925513703881&set=a.529237706231.2
Obama, Zuckererbg and other tech innitiatives	24-06-2016	https://www.youtube.com/watch?v=QrAxspymjMM
Annual Stackholder meetings	21-06-2021	https://investor.fb.com/investor-events/default.aspx

Instagram Miilion users	14-06-2016	https://www.facebook.com/photo.php?fbid=10102910644965951&set=a.612287952871.2
FAcebook Q&A First.	14-06-2016	https://www.facebook.com/zuck/videos/vb.4/10102895343490231
Safety Check in Orlando Florida Nightclub shooting	12-06-2016	https://www.facebook.com/zuck/posts/10102890573589151
Benefits of Live video	08-06-2016	https://www.facebook.com/zuck/videos/10102880633988191/
Facebook and Hosting Conservatives	18-06-2016	https://www.facebook.com/photo.php?fbid=10102840575485751&set=a.529237706231.2
TITLE	**2017**	**LINK**
Technology to prevetn suicide	27-11-2017	https://www.facebook.com/zuck/posts/10104242660091961
A year travelling. Communities	16-11-2017	https://www.facebook.com/notes/mark-zuckerberg/wrapping-up-a-year-of-travel/1015544
A year travelling. Video	10-11-2017	https://www.facebook.com/zuck/videos/10104181411255121
Small bussines. Communities	9-11-2017	https://www.facebook.com/zuck/posts/10104177999317671
Dream Act. Inmigration	8-11-2017	https://www.facebook.com/zuck/posts/10104174814021031
Oil. Comminities	8-11-2017	https://www.facebook.com/zuck/posts/10104174467739981
Leo Varadkar	2-11-2017	https://www.facebook.com/zuck/posts/10104148941579611
Earnins Call 2017	1-11-2017	https://investor.fb.com/investor-events/event-details/2017/Facebook-Q3-2017-Earnings/de
Security and integrity in Facebook	1-11-2017	https://www.facebook.com/zuck/posts/10104146268321841
Transparency Policies	27-10-2017	https://www.facebook.com/zuck/posts/10104133053040371
Visiting desaster, Comminities	9-10-2017	https://www.facebook.com/zuck/videos/10104094186863501
11th Data Centre	2-10-2017	https://www.facebook.com/zuck/posts/10104084311553701

On La Vegas Shouting LGBT.SafetyCheck	2-10-2017	ttps://www.facebook.com/zuck/posts/10104077579948891
Responding to Trump	27-09-2017	https://www.facebook.com/zuck/posts/10104067130714241
Russiian Interference	21-09-2017	https://www.facebook.com/zuck/videos/vb.4/10104052858820231/
Protecting Elections integrity	21-09-2017	https://www.facebook.com/zuck/posts/10104052907253171
Crisis Responds Tools	14-09-2017	https://www.facebook.com/zuck/posts/10104036412758271
Dream Act. DACA Vicdeo. Migration	06-09-2017	https://www.facebook.com/zuck/videos/vb.4/10104018747978631/
Dream Act. DACA interview . Migration	06-09-2017	https://www.facebook.com/zuck/posts/10104018638038951
End of DACA.	05-07-2017	https://www.facebook.com/zuck/posts/10104016069261801 (no longer available)
Dream Act. DACA Migration	31-08-2017	https://www.facebook.com/zuck/posts/10104005521334931
DACA. Pro DACA	25-08-0217	https://www.facebook.com/zuck/posts/10103990563874801
Safety Check Texas	25-08-0217	https://www.facebook.com/zuck/posts/10103991565557421
Facebook Jounalism Project	23-8-2017	https://www.facebook.com/zuck/posts/10103985905610001
No place for Hate	16-08-2017	https://www.facebook.com/zuck/posts/10103969849282011
Watch Tab	15 -08-2017	https://www.facebook.com/zuck/videos/10103953494202721
1 billion Connections	26-07-2017	https://www.facebook.com/zuck/posts/10103921337100811
Gender in Military	26-07-2017	https://www.facebook.com/zuck/posts/10103919587377271
Progress Report	26-07-2017	https://www.facebook.com/photo.php?fbid=10103920078253551&set=a.529237706231.2
Facebok Communities. FB Groups	19-07-2017	https://www.facebook.com/photo.php?fbid=10103899863723631&set=a.529237706231.2

Blackfeet Reservation visits. Unfair reservations	16-07-2017	https://www.facebook.com/zuck/posts/10103892215949831
save Net Neutrality	12-07-2017	https://www.facebook.com/zuck/posts/10103878724831141
DACA Scholarships	29-06-2017	https://www.facebook.com/zuck/videos/10103842351214161/
2 Billion Users	27-06-2017	https://www.facebook.com/photo.php?fbid=10103832396388711&set=a.941146602501.2
2017 Earning Calls	26-06-2017	https://investor.fb.com/investor-events/event-details/2017/Facebook-Q2-2017-Earnings/de
LGBT Pride	24-06-2017	https://www.facebook.com/photo.php?fbid=10103823689108171&set=a.529237706231.2
Bringing the world Closer together. Communtiy	22-06-2017	https://www.facebook.com/photo.php?fbid=10103818114983761&set=a.612287952871.2
Facebook Communitues Summit	22-06-2017	https://www.facebook.com/zuck/videos/vb.4/10103817960742861/
New Mission. Bring the Workd Closer Together	22-06-2017	https://www.youtube.com/watch?v=RYC7nAcZqn0
First ever Community Summit	21-06-2017	https://www.facebook.com/zuck/posts/10103816093894041
Sharing on Instragram	20-06-2017	https://www.facebook.com/photo.php?fbid=10103812997848541&set=a.529237706231.2
Hard Questions	15-06-2017	https://www.facebook.com/zuck/posts/10103800085669651
Safety Check	14-06-2017	https://www.facebook.com/photo.php?fbid=10103796872833201&set=a.529237706231.2
Facebook Community Summit	06-06-2017	https://www.facebook.com/photo.php?fbid=10103777403365171&set=a.529237706231.2
Annual Stakeholder meeting	01-06-2017	https://investor.fb.com/investor-events/event-details/2017/Facebook-2017-Annual-Stockh
FB member connections,. Connectivty	01-06-2017	https://www.facebook.com/photo.php?fbid=10103764965670411&set=a.529237706231.2
Hardvard Key Note	25-05-2017	https://www.facebook.com/notes/mark-zuckerberg/harvard-commencement-2017/1015485

New FB feature for india	26-04-2017	https://www.facebook.com/photo.php?fbid=10103677793444251&set=a.612287952871.2
VR to induce empathy in users	23-04-2017	https://www.facebook.com/zuck/videos/10103671105741461/
F8	21-04-2017	https://www.facebook.com/zuck/posts/10103665651406991
camera as platform for augmented reality.	19-04-2017	https://www.facebook.com/zuck/posts/10103661342566941
Tipying directly from your Brain	19-04-2017	https://www.facebook.com/zuck/videos/10103661167577621/
F8 2017 Key Note on VR	18-04-2017	https://www.facebook.com/zuck/videos/10103658908469891/
F8 2017 Key Note	18-04-2018	https://www.facebook.com/zuck/videos/10103658355917211/
5 years of Facebook and Instragram	16-04-2017	https://www.facebook.com/photo.php?fbid=10103653939113521&set=a.529237706231.2
Life after death community	13 April 2017	https://www.facebook.com/zuck/posts/10103647917336211
200 million use instragrams	13-04-2017	https://www.facebook.com/zuck/posts/10103645289592231
Fake news and Free Speech	11-04-2017	https://www.facebook.com/zuck/posts/10103640012188191
100 Anniversary US WWI	07-04-2017	https://www.facebook.com/photo.php?fbid=10103625699575781&set=a.529237706231.2
AI tools for FB, Messenger, and IG to prevent revenge porn	05-04-2017	https://www.facebook.com/zuck/posts/1010362053427708
Data Centre	04-04-2017	https://www.facebook.com/photo.php?fbid=10103617898858481&set=a.529237706231.2
Community Help. New features	24-03-2017	https://www.facebook.com/photo.php?fbid=10103593792882051&set=a.529237706231.2
Congress Rep Video	15-03-2017	https://www.facebook.com/zuck/posts/10103573965601081
Hardvard. New Marshal Plan	17-05-2017	https://www.facebook.com/zuck/posts/10103553693446641

Tools and suicide	01-03-2017	https://www.facebook.com/photo.php?fbid=10103537991633201&set=a.529237706231.2
Communities	25-02-2017	https://www.facebook.com/photo.php?fbid=10103529978371841&set=a.529237706231.2
Oil Rig. Communities	25-02-2017	https://www.facebook.com/photo.php?fbid=10103529363334381&set=a.529237706231.2
Can we all do better?	18-02-2017	https://www.facebook.com/zuck/posts/10103515344528181
Building a Global Community notes	17-02-2017	https://www.facebook.com/zuck/posts/10103511653584861
Building a Global Community notes 2	17-02-2017	https://www.facebook.com/photo.php?fbid=10103512981493721&set=a.529237706231.2
Building a Global Community	17-02-2017	https://www.facebook.com/zuck/posts/10154544292806634
200 million Users.FB Lowband. Connectivity	08-02-2017	https://www.facebook.com/photo.php?fbid=10103488405279651&set=a.529237706231.2
FB Friends. Connectivity	01-02-2017	https://www.youtube.com/watch?v=P0ssJ6Mf9kw
FB Earings Call 2016	01-02-2017	https://investor.fb.com/investor-events/event-details/2017/Facebook-Q4-2016-Earnings/de
Founders Letter	01-02-2017	https://www.facebook.com/zuck/posts/10154500412571634
Migration. Communities	27-01-2017	https://www.facebook.com/zuck/posts/10103460278231481
Facebook Products	09-01-2017	https://www.facebook.com/photo.php?fbid=10103398844570031&set=a.529237706231.2
TITLE	**2018**	**LINK**
2018 MZ personal Challenge	28-12-2018	https://www.facebook.com/zuck/posts/10105865715850211
Facebook Bussines Model	11-12-2018	https://www.facebook.com/zuck/posts/10106257635829431
British Parliament published FB internal emails	11-12-2018	https://www.facebook.com/zuck/posts/10105559172610321
CNN I'm not stepping down as Facebook chairman	27-11-2018	https://www.youtube.com/watch?v=dkceJq1EsfQ

Blue Print For Content Governance and Enforcement	15-11-2018	https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-an
full Community and Quartley update via post	20-10-2018	https://www.facebook.com/zuck/posts/10105349847863791
Third Quarter 2018	20-10-2018	https://investor.fb.com/investor-events/event-details/2018/Facebook-Q3-2018-Earnings/de
new video device	08-10-2018	https://www.facebook.com/photo.php?fbid=10105301461340771&set=a.529237706231&
Facebook Security Cyberattack	28-09-2018	https://www.facebook.com/zuck/posts/10105274505136221
Facebook dating service	20-09-2018	https://www.facebook.com/zuck/posts/10105254744057581?__xts__%5B0%5D=68.ARC8l7uizogGFg3lio2Ga0qHwm44UkarJjJr2APhz5A3zvKzmGL_bHnp6eEwpBC7CopJqUY
Evan Osnos article interview with MZ	17-09-2018	https://www.newyorker.com/magazine/2018/09/17/can-mark-zuckerberg-fix-facebook-be
FB note preparing electins	12-09-2018	https://www.facebook.com/notes/mark-zuckerberg/preparing-for-elections/101563000476
MZ discoussing his personal challenges	12-09-2018	https://www.facebook.com/zuck/posts/10105224999156601?__xts__%5B0%5D=68.ARBaXjkjNdS21XzaAThg9PfkrzJ_dTLszwUZ3H6b3Q4biIc&__tn__=-R
FB help in protecting democracy	04-09-2018	https://www.washingtonpost.com/opinions/mark-zuckerberg-protecting-democracy-is-an-
Community Boost programme	29-08-2018	https://www.facebook.com/photo.php?fbid=10105205803200451&set=a.612287952871&y5RwHtdAr_Fv575BsA7YWTGG8oP7AE38B4wQO2H1X5apetex8SiedkEdN4GRdEEK
efforts to improve safety, security, and privacy on FB.	21-08-2018	https://www.facebook.com/zuck/posts/10105188590724391?__tn__=K-R
missinformation. deleting accounts	31-07-2018	https://www.facebook.com/zuck/posts/10105140110214721?__xts__%5B0%5D=68.ARDR
Facebook Q2 2018 Earnings	25-07-2018	https://s21.q4cdn.com/399680738/files/doc_financials/2018/Q2/Q218-earnings-call-transc
Community Update	25-07-2018	https://www.facebook.com/zuck/posts/10105127808736981?__xts__%5B0%5D=68.ARA
On Holocoust Denial	18-07-2018	https://www.vox.com/2018/7/18/17588116/mark-zuckerberg-clarifies-holocaust-denial-of

Interview with Kara Swisher	12-07-2018	https://www.recode.net/2018/7/18/17575156/mark-zuckerberg-interview-facebook-recode
LGBT Celebration	24-06-2018	https://www.facebook.com/zuck/posts/10105044271137001?__xts__%5B0%5D=68.ARD WKVv8sKYb61X6omk_Ot1c5ZxVJ0DFvl7NPhcYOSuLFJKKuRf7Kd_M3b22O5kgqjHl
Video Phone	20-06-2018	https://www.facebook.com/zuck/videos/10105035195075501/
Donations tools	19-06-2018	https://www.facebook.com/zuck/posts/10105029863405211?__xts__%5B0%5D=68.ARD oc5UKEJuW_4QY9Eiq4W8mKbVxImYUMlqtiwGHZKVOua_B0QCWG0qfmMqx4Kw
Changes online political Ads	24-05-2018	https://www.facebook.com/photo.php?fbid=10104995803386761&set=a.529237706231&
Meeting Presidente Macron	23-05-2018	https://www.facebook.com/photo.php?fbid=10104971005072781&set=a.529237706231&
Meetings in Europe	22-05-2018	https://www.facebook.com/photo.php?fbid=10104969495512951&set=a.529237706231&
MZ Testifies before the EU Parlament	22-05-2018	https://www.facebook.com/photo.php?fbid=10104971005072781&set=a.529237706231&
FB Priorities Safe Community	15-05-2018	https://www.facebook.com/zuck/posts/10104954056777301?__xts__%5B0%5D=68.ARD R
2018 F* MZ Key note	01-05-2018	https://www.facebook.com/zuck/videos/10104900382520941/
2018 1Q resoults	25-04-2018	https://s21.q4cdn.com/399680738/files/doc_financials/2018/Q1/Q1-18-Earnings-call-trans
Building Community Standards	24-04-2018	https://www.facebook.com/zuck/posts/10104874769784071?__xts__%5B0%5D=68.ARD pt3hDB&__tn__=-R
House Committee appereance	11-04-2018	https://www.washingtonpost.com/news/the-switch/wp/2018/04/11/transcript-of-zuckerber
MZ Senate Hearing	10-04-2018	https://www.washingtonpost.com/news/the-switch/wp/2018/04/10/transcript-of-mark-zuck
MZ on his Senate Hearing	10-04-2018	https://www.facebook.com/photo.php?fbid=10104804714685051&set=a.529237706231&
Build new AI tools to moniter fake accounts.	09-04-2018	https://www.facebook.com/zuck/posts/10104797374385071?__xts__%5B0%5D=68.ARD k0JMEGsQQ19MWCLDZTfQVs9FxAPYPeogmcPH3G85rAETW7SAaauxjZQp5AaHvz
Interview with Robinson Meyer	09-04-2018	https://www.theatlantic.com/technology/archive/2018/04/mark-zuckerberg-atlantic-exclus
Verification Steps for advertisers	06-04-2018	https://www.facebook.com/zuck/posts/10104784125525891

Hard Questions. Prtecting peoples information	03-04-2018	https://newsroom.fb.com/news/2018/04/hard-questions-protecting-peoples-information/
Removal of 270 IRA accounts	03-04-2018	https://www.facebook.com/zuck/posts/10104771321644971
Facebook Hardest year. Interview	02-04-2018	https://www.vox.com/2018/4/2/17185052/mark-zuckerberg-facebook-interview-fake-new
Interview with Stephen Dubmer	01-04-2018	http://freakonomics.com/podcast/mark-zuckerberg/
Privacy shortcuts	28-03-2018	https://www.facebook.com/zuck/posts/10104747087565261
Interview on Cambridge Analyitica	21-03-2018	https://www.youtube.com/watch?v=G6DOhioBfyY
Interview on Cambridge Analyitica	21-03-2018	https://www.nytimes.com/2018/03/21/technology/mark-zuckerberg-q-and-a.html
Post on Cambridge Analytica	21-03-2018	https://www.facebook.com/zuck/posts/10104712037900071
Facebook Privacy article	21-03-2018	https://www.wired.com/story/mark-zuckerberg-talks-to-wired-about-facebooks-privacy-pr
MZ on the Enlightiment	11-02-2018	https://www.facebook.com/zuck/posts/10104551772588041
Leadership in Communities	09-02-2018	https://www.facebook.com/zuck/posts/10104544657960811
2017 full year resoults	31-01-2018	https://investor.fb.com/investor-events/default.aspx
MZ Goals 2018	31-01-2018	https://www.facebook.com/zuck/posts/10104501954164561
News Feeds and local news	29-01-2018	https://www.facebook.com/zuck/posts/10104493997365051
News Feed users manipulation	19-01-2018	https://www.facebook.com/zuck/posts/10104445245963251
FB Board of Directors	18-01-2018	https://www.facebook.com/zuck/posts/10104440501581031
Dreamers ACT migration	17-01-2018	https://www.facebook.com/zuck/posts/10104437004723761

MLK quote, "The arc of the moral universe is long, but it bends towards justice."	15-01-2018	https://www.facebook.com/zuck/posts/10104430853346171
NewsFeeds update	11-01-2018	https://www.facebook.com/zuck/posts/10104413015393571
MZ Personal Challenges	04-01-2018	https://www.facebook.com/zuck/posts/10104380170714571
TITLE	**2019**	**LINK**
On individualism	02-12-2019	https://chanzuckerberg.com/about/annual-letter/
New FB logo	04-11-2019	https://www.facebook.com/zuck/videos/10110362239999771/
Quoting Aaron Sorking on Freedom	21-10-2019	https://www.facebook.com/zuck/posts/10110290602526841
On Quartely resoults	30-10-2019	https://www.facebook.com/zuck/posts/10110264733792991
3Q 2019 Earnings	20-10-2019	https://investor.fb.com/financials/default.aspx
FB News	25-10-2019	https://www.facebook.com/zuck/videos/vb.4/10109952468829481
FB on bussines	25-10-2019	https://www.nytimes.com/2019/10/25/opinion/mark-zuckerberg-facebook-news.html
FB rypocurrency	23-10-2019	https://financialservices.house.gov/calendar/eventsingle.aspx?EventID=404487
FB and affordable housing	22-10-2019	https://www.facebook.com/zuck/posts/10109900954369871
MZ on NBC Personal relfections	22-10-2019	https://www.facebook.com/zuck/posts/10109900083924251
Changes at FB	21-10-2019	https://www.nbcnews.com/nightly-news/video/extended-interview-mark-zuckerberg-on-e
MZ on FB responsability	21-10-2019	https://www.nbcnews.com/nightly-news/video/mark-zuckerberg-i-feel-responsible-for-ho
MZ link to NBC interview	21-10-2019	https://www.facebook.com/zuck/posts/10109887266784901
MZ and Dana Perino	18-10-2018	https://www.youtube.com/watch?v=LJDANzTzI0o
MZ on Free Speech	17-10-2019	https://www.washingtonpost.com/technology/2019/10/17/facebook-ceo-mark-zuckerberg-

Standing ofr Voice and Freedom of expression	17-10-2019	https://www.facebook.com/notes/mark-zuckerberg/standing-for-voice-and-free-expressio
On Free Speech	16-10-2019	https://www.facebook.com/zuck/posts/10109797261606001
Dinner with conservatives	14-10-2019	https://www.facebook.com/zuck/posts/10109760864675741
Chat Joe DeRisi and Steve Quake	10-10-2019	https://www.facebook.com/zuck/videos/vb.4/10109692205948381
Facebook employees	03-10-2019	https://www.facebook.com/zuck/posts/10109570963534151
eaked internal Facebook Q&A session	01-10-2019	https://www.theverge.com/2019/10/1/20892354/mark-zuckerberg-full-transcript-leaked-fa
MZ post about leaked internal conversation	01-10-2019	https://www.facebook.com/zuck/posts/10109531568162791
Oversight Board	17-09-2019	https://www.facebook.com/zuck/posts/10109281036175901
Zuckerberg and Cass Sustain dialaoge	26-06-2019	https://www.facebook.com/zuck/videos/vb.4/10107806730007281
MZ and Macron meeting	10-05-2019	https://www.facebook.com/photo.php?fbid=10107320597760201
F8	30-04-2019	https://www.facebook.com/photo.php?fbid=10107278677733271&set=a.529237706231&
Conversation with Noah Harari about MZ 2019 personal challenge	26-04-2019	https://newsroom.fb.com/news/2019/04/marks-challenge-yuval-noah-harari/
New Board of Directos	12-04-2019	https://www.facebook.com/zuck/posts/10107134779456571
Trip to Dublin	02-04-2019	https://www.facebook.com/photo.php?fbid=10107039318795501&set=a.612287952871& pyiIZEtXuTiMdt5VL1pr5fwgQ5QA3UPvGgIij62PJ7FU_wmbzG7JmwjdveHr6Xgvf3b2 mWtho78XCgtNCcwa9bhrBuITiudGdxkLOpLw&__tn__=-R
MZ published editorial on online platform regulation	30-03-2019	https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules

Privacy on Facebook	06-03-2019	https://www.facebook.com/notes/mark-zuckerberg/a-privacy-focused-vision-for-social-ne
MZ Challeges	08-01-2019	https://www.facebook.com/zuck/posts/10106021347128881
TITLE	**2020**	**LINK**
Jack Dorsey Testified before the senate	17-11-2020	https://www.c-span.org/video/?478048-1/facebook-twitter-ceos-testify-regulating-social-r
On Content Moderation	18-10-2020	https://www.c-span.org/video/?476686-1/social-media-content-moderation
About Voting	26-10-2020	https://www.facebook.com/zuck/posts/10112507395046861
On voting information centre	26-10-2020	https://www.facebook.com/zuck/posts/10112506564655971
Banning Post on Holocoust denial	12-12-2020	https://www.facebook.com/zuck/posts/10112455086578451
Facebook Connect Event	15-09-2019	https://www.facebook.com/zuck/posts/10112340462945131
Voting and missingformation	03-10-2020	https://www.facebook.com/zuck/posts/10112270823363411
Facebook Reality labs	25-08-2020	https://www.facebook.com/zuck/posts/1011222095397706
Antri Trust Law	29-07-2020	https://www.c-span.org/video/?474236-1/heads-facebook-amazon-apple-google-testify-an
No deal with Trump	20-07-2020	https://www.axios.com/mark-zuckerberg-trump-facebook-interview-ae28771f-71b9-4df8-
Death of John Lewis	18-07-2020	https://www.facebook.com/zuck/posts/10112115058961521
On racial Justice	26-06-2020	https://www.facebook.com/zuck/posts/10112048823443031
New Content polices . Elections	26-06-2020	https://www.facebook.com/zuck/posts/10112048980882521
Lift Black voices features. Products	19-06-2020	https://www.facebook.com/zuck/posts/10112027900891991
DACA . Migration	18-06-2020	https://www.facebook.com/zuck/posts/10112025073867371
Black Owned bussines	18-06-2020	https://www.facebook.com/zuck/posts/10112025405303171

Voting info campaing	17-06-2020	https://www.facebook.com/zuck/posts/10112021381970961
LGBT on the Workplace	15-06-2020	https://www.facebook.com/zuck/posts/10112016153399061
Content Policies (beyond binary)	11-06-2020	https://www.facebook.com/zuck/posts/10111985969467901
TRumps controversial posts	03-06-2020	https://www.vox.com/recode/2020/6/3/21279434/mark-zuckerberg-meeting-facebook-em
on George Floyd death	31-05-2020	https://www.facebook.com/zuck/posts/10111969612272851
About Trump	29-05-2020	https://www.facebook.com/zuck/posts/10111961824369871
FB censoring voices?	28-05-2020	https://www.youtube.com/watch?v=Mu7SgOZ2nEs
Interview with Andrew Ross	28-05-2020	https://www.cnbc.com/video/2020/05/28/watch-cnbcs-full-interview-with-facebook-ceo-m
New platform for small bussiens	19-05-2020	https://www.youtube.com/watch?v=p7QjDCah28M
On Facebook Shops	19-05-2020	https://www.facebook.com/zuck/posts/10111930156208121
New product updates	19-05-2020	https://www.facebook.com/zuck/videos/10111929914173161/
Product updates	18-05-2020	https://www.facebook.com/zuck/posts/10111927602176421
Oversight Board	06-05-2020	https://www.facebook.com/zuck/posts/10111886600788531
New products Connectivity	23-04-2020	no link available
New Governonr Board membre	26-03-2020	https://www.facebook.com/zuck/posts/10111716796258171
On Facebook instrastructure	24-03-2020	https://www.facebook.com/zuck/posts/10111707373631191
New baord Memembers	09-03-2020	https://www.facebook.com/zuck/posts/10111636392268411
MZ on Communities on FB	12-02-2020	https://www.facebook.com/zuck/timeline?lst=100028603934067%3A4%3A1583118453

MZ 2020 Personal Challenge	06-01-2020	https://www.facebook.com/zuck/posts/10111311886191191

Appendix II Principles Version 1 and Version 2.

(version 1)

7/17/2017 Facebook

Search Facebook Paloma Home 1

Facebook Principles

We are building Facebook to make the world more open and transparent, which we believe will create greater understanding and connection. Facebook promotes openness and transparency by giving individuals greater power to share and connect, and certain principles guide Facebook in pursuing these goals. Achieving these principles should be constrained only by limitations of law, technology, and evolving social norms. We therefore establish these Principles as the foundation of the rights and responsibilities of those within the Facebook Service.

1. Freedom to Share and Connect

People should have the freedom to share whatever information they want, in any medium and any format, and have the right to connect online with anyone - any person, organization or service - as long as they both consent to the connection.

2. Ownership and Control of Information

People should own their information. They should have the freedom to share it with anyone they want and take it with them anywhere they want, including removing it from the Facebook Service. People should have the freedom to decide with whom they will share their information, and to set privacy controls to protect those choices. Those controls, however, are not capable of limiting how those who have received information may use it, particularly outside the Facebook Service.

3. Free Flow of Information

People should have the freedom to access all of the information made available to them by others. People should also have practical tools that make it easy, quick, and efficient to share and access this information.

4. Fundamental Equality

Every Person - whether individual, advertiser, developer, organization, or other entity - should have representation and access to distribution and information within the Facebook Service, regardless of the Person's primary activity. There should be a single set of principles, rights, and responsibilities that should apply to all People using the Facebook Service.

5. Social Value

People should have the freedom to build trust and reputation through their identity and connections, and should not have their presence on the Facebook Service removed for reasons other than those described in Facebook's Statement of Rights and Responsibilities.

6. Open Platforms and Standards

People should have programmatic interfaces for sharing and accessing the information available to them. The specifications for these interfaces should be published and made available and accessible to everyone.

7. Fundamental Service

People should be able to use Facebook for free to establish a presence, connect with others, and share information with them. Every Person should be able to use the Facebook Service regardless of his or her level of participation or contribution.

8. Common Welfare

The rights and responsibilities of Facebook and the People that use it should be described in a Statement of Rights and Responsibilities, which should not be inconsistent with these Principles.

9. Transparent Process

Facebook should publicly make available information about its purpose, plans, policies, and operations. Facebook should have a process of notice and comment to provide transparency and encourage input on amendments to these Principles or to the Rights and Responsibilities.

10. One World

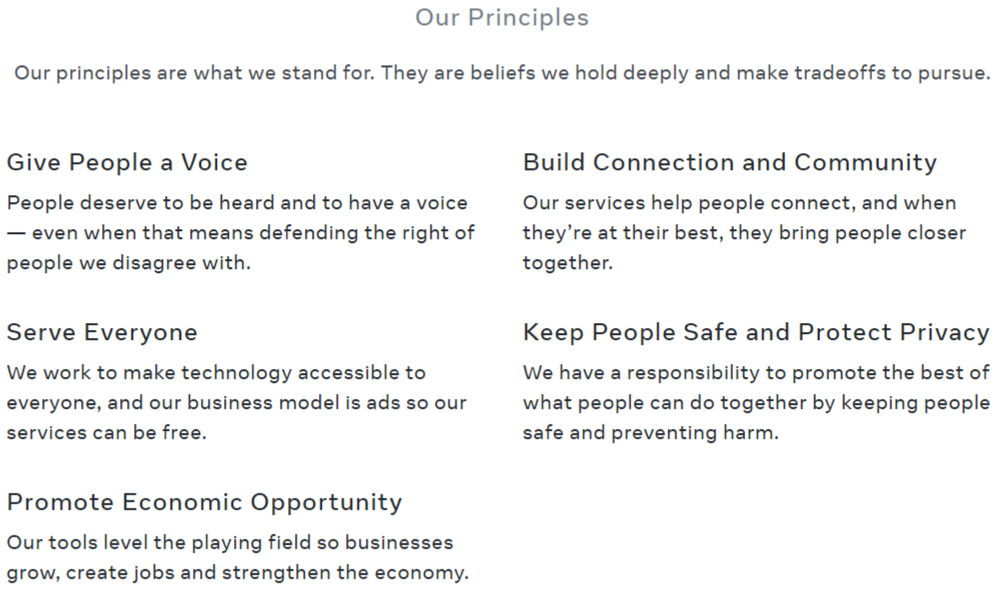
The Facebook Service should transcend geographic and national boundaries and be available to everyone in the world.

Developers Careers Privacy Cookies Terms Help AdChoices

Facebook © 2017
English (UK) English (US) Español Polski Lietuvių Português (Brasil) Русский Română Français (France) Deutsch Italiano

https://www.facebook.com/principles.php 1/1

(version 2)



Our Principles

Our principles are what we stand for. They are beliefs we hold deeply and make tradeoffs to pursue.

Give People a Voice

People deserve to be heard and to have a voice — even when that means defending the right of people we disagree with.

Build Connection and Community

Our services help people connect, and when they're at their best, they bring people closer together.

Serve Everyone

We work to make technology accessible to everyone, and our business model is ads so our services can be free.

Keep People Safe and Protect Privacy

We have a responsibility to promote the best of what people can do together by keeping people safe and preventing harm.

Promote Economic Opportunity

Our tools level the playing field so businesses grow, create jobs and strengthen the economy.

Apendix III UN Sessions Drafting Committee 1947

Meeting **Reference Link Intervetnion**

1st Meeting, Monday, 9 June 1947 :10/06/1947 E/CN.4/AC.1/SR.1 https://undocs.org/E/CN.4/AC.1/SR.2

2nd Meeting, Wednesday, 11 June 1947 :13/06/1947 E/CN.4/AC.1/SR.2 https://undocs.org/E/CN.4/AC.1/SR.2 Page 6

3rd Meeting, Wednesday, 11 June 1947 :13/06/1947 E/CN.4/AC.1/SR.3 https://undocs.org/E/CN.4/AC.1/SR.2 Page 3 and 7

4th Meeting, Thursday, 12 June 1947 :13/06/1947 E/CN.4/AC.1/SR.4 https://undocs.org/E/CN.4/AC.1/SR.2 Evidence of the detail preparation of their speech. Right to comment later time.

5th Meeting, Thursday, 12 June 1947 :17/06/1947 E/CN.4/AC.1/SR.5 https://undocs.org/E/CN.4/AC.1/SR.5 On Fighting Fascism and the destruction of discrimination and inequality. Pag 5 On the role fo the State page 9

6th Meeting, Thursday, 16 June 1947 :16/06/1947 E/CN.4/AC.1/SR.6 https://undocs.org/E/CN.4/AC.1/SR.6 Not enough done on discrimination. Pag 3. Reference to the masses of people instead of individual, page 4

7th Meeting, Tuesday, 17 June 1947 :19/06/1947 E/CN.4/AC.1/SR.7 https://undocs.org/E/CN.4/AC.1/SR.7 On the creation of sub-commitees

8th Meeting, Tuesday, 17 June 1947 :20/06/1947 E/CN.4/AC.1/SR.8 https://undocs.org/E/CN.4/AC.1/SR.7 Aticle 7 Idea on Everyone Page 4 On resisting oppression Art 25. Not included in thesis. Prof Cassin. France

9th Meeting, Wednesday, 18 June 1947 :03/07/1947 E/CN.4/AC.1/SR.9 https://undocs.org/E/CN.4/AC.1/SR.7 On aslylum page 8. discarded for thesis

10th Meeting, Wednesday, 18 June 1947 :20/06/1947

E/CN.4/AC.1/SR.10 https://undocs.org/E/CN.4/AC.1/SR.10

On the need for the rest of commissioners for full preparation

before consolidating any draft. 11th Meeting, Thursday, 19

June 1947 :03/07/1947 E/CN.4/AC.1/SR.11

https://undocs.org/E/CN.4/AC.1/SR.11

12th Meeting, Friday, 20 June 1947 :03/07/1947 E/CN.4/AC.1/SR.12 https://undocs.org/E/CN.4/AC.1/SR.11

13th Meeting, Friday, 20 June 1947 :08/07/1947 E/CN.4/AC.1/SR.13 https://undocs.org/E/CN.4/AC.1/SR.13 ...members seemed to have accepted the expression "all men "make it clear that all human beings were included. On the understanding that all persons were 'included. On historical reflection on the mastery of men over women, - and that the phrase should be modified. page 6. On no discrimination pag 13. On marriage and discrimination against women page 14

14th Meeting, Monday, 23 June 1947:03/07/1947 E/CN.4/AC.1/SR.14 https://undocs.org/E/CN.4/AC.1/SR.14 on discrimination and access to work

15th Meeting, Monday, 23 June 1947 :03/07/1947 E/CN.4/AC.1/SR.15 https://undocs.org/E/CN.4/AC.1/SR.15 On workers and vacation

16th Meeting, Held At Lake Success, 24 June 1947 :03/07/1947 E/CN.4/AC.1/SR.16 https://undocs.org/E/CN.4/AC.1/SR.16

17th Meeting, Held At Lake Success, 24 June 1947 :03/07/1947 E/CN.4/AC.1/SR.17 https://undocs.org/E/CN.4/AC.1/SR.17

18th Meeting, Held At Lake Success, 25 June 1947 :03/07/1947 E/CN.4/AC.1/SR.18 https://undocs.org/E/CN.4/AC.1/SR.18 On implementing the Bill of Rights

Summary Record of the 19th Meeting, Held At Lake Success, 25 June 1947 :03/07/1947 E/CN.4/AC.1/SR.19 https://undocs.org/E/CN.4/AC.1/SR.19

Reports and Actions E/CN.4/21 https://undocs.org/E/CN.4/21