Colonial Platform Gaslighting: Meta's Automated Silencing of Palestinian Narratives

By

Student Number 1983301

A Dissertation submitted in partial fulfilment of the requirements for the degree of
MSc in Big Data and Digital Futures

Supervised by Kavin Narasimhan

University of Warwick
Centre for Interdisciplinary Methodologies
August 2024

Abstract

The ongoing conflict between Israel and Hamas, which escalated on October 7, 2023, has led to a severe humanitarian crisis in Gaza, drawing significant international attention (Cruz, 2023). Amid this conflict, social media platforms, particularly Meta’s Facebook and Instagram, have become crucial spaces for documenting events, expressing opinions, and mobilising for ceasefires and hostage releases (Hamla, 2024). However, these platforms have faced widespread criticism for allegedly censoring pro-Palestinian content, highlighted by a Human Rights Watch report titled "Meta's Broken Promises" (Younes, 2023),

This dissertation investigates how Meta’s content moderation algorithms act as colonial tools, perpetuating cultural gaslighting in the Israeli-Palestinian context. By integrating decolonial theory, postcolonial theory, political epistemology, and digital studies, and drawing on the HRW report, the research introduces the framework of Colonial Platform Gaslighting. A detailed examination of Meta’s policies and practices reveals how these algorithms systematically suppress marginalised narratives to sustain colonial power structures.

The findings demonstrate that Meta’s algorithms disproportionately target Palestinian content, reflecting historical patterns of censorship and silencing. The research uncovers how the company's reliance on a universal human rights framework often disregards the socio-political contexts of oppressed groups, particularly within the MENA region. The paper critiques Meta’s DOI policy for maintaining a simplistic state versus non-state binary, obscuring the complexities of modern geopolitical realities, and highlights the orientalist context within which Meta’s technologies operate.

The dissertation argues that algorithmic moderation represents epistemic violence, exacerbating inequalities, and causing self-censorship and fragmentation within advocacy movements. The paper concludes with policy recommendations for reforming Meta’s content moderation practices to ensure a more equitable and context-sensitive approach. Despite limitations, including restricted access to critical datasets, the study lays the groundwork for further exploration of Colonial Platform

Gaslighting and calls for the decolonisation of digital platforms to better protect marginalised voices.

Table of Contents

Abstract ii

List of Figures vi

Introduction 7

Chapter 1: Theoretical Foundations 10

  Digital Studies 11

  Postcolonial Theory and Decoloniality 11

  Political Epistemology 14

Chapter 2: Historical Evolution of Resistance and Censorship 16

  Historical Context 16

  Early Censorship 18

  The Shift to Institutionalised Censorship 18

  Digital Media and the New Age of Censorship 19

  The Rise of Self-Censorship 20

  The Internet: Battleground for Free Speech and Control 21

  The Age of Social Media and Automation 21

Chapter 3: Meta’s Technologies, Policies and Broken Promises 23

  Overview of Content Moderation Technologies and Mechanisms 23

    Meta’s AI-Driven Approach to Content Moderation 23

    Meta’s Mechanisms of Content Detection 23

  Overview of Meta’s Policies 24

    Corporate Human Rights Policy 24

    The Dangerous Organisations and Individuals Policy 25

  “Meta’s Broken Promises” 27

    Six Patterns of Censorship 28

  Recommendations from Civil Society and Oversight Bodies ... 30
    Reform the DOI Policy ... 30
    Improve Transparency and Reduce Automation ... 30
Chapter 4: Introducing Colonial Platform Gaslighting ... 32
  Cultural Gaslighting ... 32
  Colonial Platform Gaslighting ... 33
  Settler Steals Innocence: All Eyes on Rafah ... 36
Chapter 5: Future Directions and Recommendations ... 38
  Remove Technologies from their Orientalist Context ... 38
  Look Beyond a Universal Standard ... 40
  Reform the Dangerous Definition ... 41
  Future Pathways ... 44
Conclusion ... 45
Appendix 1- Code Used ... 47
    For bar of pie chart: ... 47
    For the map: ... 48
Appendix 2- Use of GAITs ... 51
    Elicit AI ... 51
    Copilot ... 53
    ChatGPT ... 54
Appendix 3- Copy of Academic Integrity Declaration – CIM Dissertation ... 56
Abbreviations ... 58
References ... 59

List of Figures

Figure 1- Theoretical Foundations Visualised 10
Figure 2- Timeline of Epistemic Violence In Palestine... 17
Figure 3- Screenshot of Meta's Corporate Human Rights Policy 25
Figure 4- Screenshot of DOI Tiers.. 26
Figure 5- Breakdown of Reported Censorship Incidents by Company...................... 28
Figure 6- Bridging the Gap 35
Figure 7- Mapping Meta’s Office Locations 40
Figure 8- Meta’s definition of Terrorist Organisations and Individuals....................... 42

Introduction

On October 7, 2023, Hamas launched an attack on Israel, resulting in the deaths of approximately 1,200 Israelis and the abduction of over 200 hostages (Alfonseca, 2023; Cruz, 2023). Israel declared war in response with the intent to dismantle Hamas, leading to a large-scale military campaign in Gaza (Narea, 2023). The ongoing conflict has resulted in a severe humanitarian crisis, with thousands of Palestinian casualties and widespread destruction across Gaza (Hamla, 2024). The humanitarian situation, already dire due to longstanding blockades and restrictions, has been further exacerbated by this conflict, drawing intense international attention and concern.

Amid the escalating violence, social media platforms have emerged as critical arenas for documenting the conflict, expressing opinions, and mobilising demands for a ceasefire and the release of hostages (Hamla, 2024). However, these platforms, especially Meta’s Facebook and Instagram, have been widely criticised for their alleged censorship of pro-Palestinian content. A pivotal report by Human Rights Watch (HRW), titled "Meta's Broken Promises," has exposed systematic suppression of such content on Meta's platforms, highlighting broader concerns about the role of social media companies in moderating content related to conflicts (Younes, 2023). These allegations of biased censorship have sparked widespread condemnation and have brought to light the significant influence of algorithmic content moderation on global discourse.

This research aims to theorise how content moderation algorithms perpetuate historical patterns of oppression in contemporary colonial contexts, specifically through the lens of Meta's practices. Content moderation, in this context, refers to the process by which social media platforms monitor and regulate user-generated content to ensure it aligns with legal requirements, community guidelines, and culturally specific norms of acceptability (Register et al., 2024). The study focuses on how these algorithms serve as tools of colonial control, perpetuating cultural gaslighting within the Palestinian context. This case study is selected as Israel’s occupation of Palestinian territory is described as “indistinguishable from a settler colonial situation” by the UN Special Rapporteur on the situation of human rights in the Palestinian Territory occupied since 1967 (Albanese, 2022). By extending the concept of cultural gaslighting- where the oppressor invalidates the knowledge and experiences of the colonised to maintain epistemic dominance (Ruíz, 2020)- into the digital realm, the paper illustrates how content moderation algorithms systematically suppress certain narratives, reinforcing colonial power dynamics and act as a form of epistemic violence.

To achieve the above, I answer several key research questions. The primary question asks how Meta’s content moderation algorithms and policies act as colonial tools that perpetuate cultural gaslighting in the Israeli-Palestinian context. I also explore how algorithmic censorship reflects ongoing settler colonial and apartheid dynamics, what historical and contemporary socio-political contexts influence these algorithmic biases, and what policy recommendations can be made to address these biases to ensure more equitable treatment of Palestinian content on digital platforms.

The HRW report "Meta’s Broken Promises" serves as the primary case study for this research. This paper confirms, challenges, or expands on the report’s findings and conclusions by examining the report’s references and tracing how these works have been cited in other studies. An interdisciplinary approach, integrating decolonial theory, political epistemology, and digital studies, critically examines the case study to explain the causes and impact of content moderation on Palestinian voices.

Beyond the report, the research conducts an in-depth analysis of Meta’s content moderation policies, terms of service, and corporate human rights policy, along with less obvious Meta resources such as a list of their global offices, to understand the geographical and cultural contexts in which their moderation practices are developed and implemented. This analysis is supported by academic research using databases like Google Scholar and JSTOR, focusing on search terms such as "algorithmic bias," "content moderation," and "digital colonialism in Palestine." Additionally, the AI tool Elicit (2024) is used to identify gaps in the literature and find papers exploring concepts I seek to explore further after the initial literature search. Primary research is deemed unfeasible due to time constraints, limited resources, and the risk of sampling bias, as those likely to participate—such as younger individuals with strong pro-Palestine views—would result in a skewed sample without a balanced representation or a comparative control group.

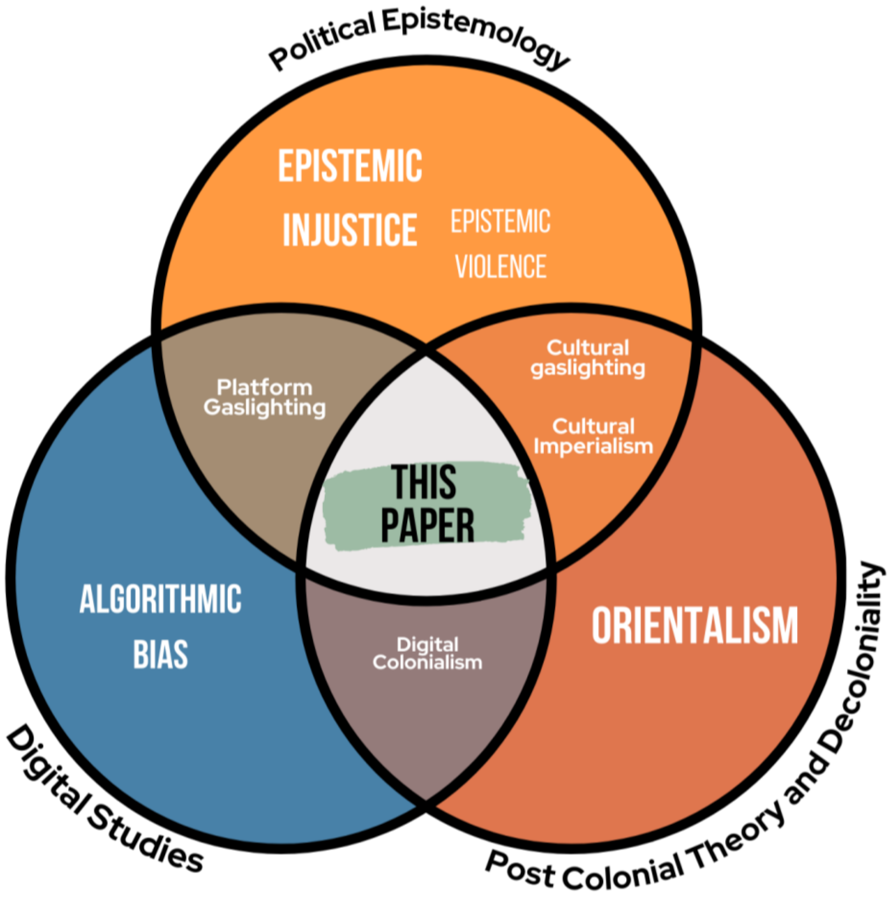
This study contributes to the fields of digital studies, postcolonial and decolonial theory, and political epistemology by offering a comprehensive view of how digital tools can enforce colonial agendas. By introducing the theory of *Colonial Platform Gaslighting*, the research provides contemporary insights into the suppression of Palestinian content and proposes recommendations for enhancing transparency, accountability, and equitable treatment in digital spaces.

This dissertation is structured as follows: The first chapter discusses the key theories underpinning the research within the fields of digital studies, postcolonial and decolonial theory, and political epistemology. The second chapter explores the evolution of censorship in Palestine by Israeli authorities and other actors. The third chapter provides an overview of Meta's content moderation practices and policies, focusing on their impact on Palestinian content as revealed by the HRW report and further research. The fourth chapter extends the concept of cultural gaslighting into the digital context by developing the theory of Colonial Platform Gaslighting. The fifth chapter synthesises the findings to inform policy recommendations, discuss implications for theory and practice, and suggest avenues for future research. The conclusion summarises the key findings and contributions of this paper and discusses its limitations.

Chapter 1: Theoretical Foundations

This chapter establishes the theoretical groundwork for analysing the implications of algorithmic content moderation on colonised peoples. By integrating perspectives from digital studies, postcolonial theory, and political epistemology, it offers a foundation for understanding the systemic inequalities perpetuated by digital tools and their impact on the interconnected online and offline experiences of marginalised groups. Figure 1 visualises the theoretical foundations of this paper, positioning it within the literature.

Figure 1- Theoretical Foundations Visualised



Digital Studies

Digital studies explore the impact of digital technologies on society, culture, and politics. Within this field, this paper focuses on algorithmic bias, which refers to systematic errors in automated systems that often result in unjust outcomes, privileging certain groups and reflecting existing social prejudices (Gorwa, Binns & Katzenbach, 2020; Ricaurte, 2022; Siapera, 2022). My use of the word “privileging” is intentional, as past research shows that these biases frequently benefit already privileged groups, such as white, heterosexual, Western males, while disadvantaging marginalised communities (Beer, 2017; Shahid & Vashistha, 2023). The following chapters will focus on bias in algorithmic content moderation.

By grounding the discussion in the framework of algorithmic bias, this paper challenges the perception of algorithms as neutral and rational tools (Beer, 2017; Ricaurte, 2022). This critical perspective invites a deeper analysis of how algorithms are employed, especially in colonial contexts, where their decisions and outcomes can be critically examined through a decolonial lens.

Postcolonial Theory and Decoloniality

Postcolonial theory critically examines the lasting impacts of colonialism on former colonies and how colonial power dynamics persist today (Bhambra, 2014). This paper draws on postcolonial theory to explain how historical and ongoing colonial practices are perpetuated through modern digital technologies, particularly content moderation algorithms. The concept of digital colonialism, which describes the control and exploitation of digital infrastructures, data, and knowledge, brings postcolonial theory into the digital age. Shahid and Vashistha, (2023) describe digital colonialism as a contemporary manifestation of colonial power dynamics, where the Global North’s control over digital infrastructure perpetuates historical inequities.

Edward Said's (1977) critique of Western scholarship, particularly through his concept of Orientalism, is crucial for analysing digital colonialism in the Palestinian context. Orientalism challenges how Western scholarship has portrayed "the Orient" as exotic, different and backward, justifying colonial dominance through othering (Said, 1977;

Bhambra, 2014). His work shows how claims of universality in European thought—presented as neutral and objective, *much like algorithms*—are actually rooted in specific historical and cultural contexts and are upheld by colonial power (Bhambra, 2014). Throughout this paper, I draw parallels between Orientalism and the use of biased content moderation algorithms, arguing that digital platforms, under the guise of neutrality, continue to privilege certain viewpoints while suppressing others, thereby perpetuating imperial dominance in the digital age. Alimardani and Elswah (2021) have similarly argued that Western social media companies use policies that reflect this colonial framework of othering, disadvantaging users in marginalised regions through media, governance, and digital policies.

Also critical to this paper's analysis is Said's notion of cultural imperialism. Cultural imperialism refers to the way dominant cultures impose their values and norms on other societies, often under the guise of globalisation or modernisation (Said, 1994). Within the context of digital colonialism and AI, particularly pertinent is the role of the United States and other Global North entities, where a few multinational corporations centralise power and resources (Bhattacharyya, 2022; Ricaurte, 2022). This centralisation extends economic influence and perpetuates a digital hegemony that marginalises the 'othered'.

While postcolonial theory helps us understand how colonial legacies prevail in digital practices, decoloniality offers a means of dismantling these structures by advocating for direct action and transformative change with the goal of the liberation and empowerment of colonised peoples (Mignolo, 2011; Bhambra, 2014; Shahid & Vashistha, 2023). A decolonial approach would likely advocate for incorporating diverse perspectives from the Global South into mainstream technology development, addressing Western-centric biases in algorithms.

I use the word 'likely' as, despite the richness of these theoretical frameworks, there has been relatively little work integrating the various trajectories of these fields (Bhambra, 2014), especially within the digital colonialism literature. For example, studies on algorithmic bias adopt a predominantly postcolonial perspective, focusing on how colonial racial hierarchies influence current technologies (Véliz, 2023).

Therefore a more comprehensive synthesis that includes both postcolonial critiques and decolonial methodologies is needed to provide a fuller understanding of digital technologies' roles in perpetuating colonial legacies. This thesis aims to do so by integrating decolonial political epistemology into the ideas mentioned thus far.

Decoloniality is especially relevant here due to its ability to address settler colonialism, which I argue is inadequately addressed by state-centric theories (to be expanded upon in later chapters). Traditional state-centric theories, such as realism and liberalism, emphasise the state as the primary actor in international relations (IR), often overlooking the pervasive impacts of colonialism that extend beyond recognised state boundaries and persist in digital practices (Waltz, 1979; Keohane, 2005, Anon, 2022). For instance, realism’s focus on state self-interest overlooks the colonial histories shaping state behaviours (Waltz, 1979), while liberalism's emphasis on cooperation ignores the historic inequalities affecting state dynamics (Keohane, 2005).

In the context of algorithmic content moderation, particularly as implemented by Meta (formerly Facebook), I find that state-centric theories fall short. They often fail to recognise how digital platforms perpetuate colonial structures and ideologies. In contrast, postcolonial and decolonial perspectives emphasise the importance of addressing the specific historical and cultural contexts of colonialism, which state-centric theories often overlook.

Meta's content moderation algorithms have been criticised for disproportionately targeting Palestinian content, reflecting a broader pattern of digital colonialism where control over digital infrastructures reinforces historical power imbalances (Younes, 2023). In forthcoming chapters I outline how the language of digital rights often mirrors liberal rights frameworks, focusing on individual freedoms and state control. However, these frameworks inadequately address the collective rights and historical injustices faced by marginalised groups, especially in contexts where the state itself is an instrument of colonial oppression. I argue that liberal frameworks, with their emphasis on universal rights, can be co-opted (Peck, 2011) to justify ‘neutral’ content moderation policies that perpetuate colonial biases. Meta’s suppression of Palestinian narratives emphasises how supposedly neutral algorithms are influenced by geopolitical biases and power structures, leaving marginalised groups vulnerable to systemic censorship.

Political Epistemology

Political epistemology explores the intersection of knowledge, power, and politics. Epistemology is the study of knowledge—its nature, origin, and, importantly, its limits (Edenberg & Hannon, 2021). Knowledge can be described as limited as it can be seen as a social and epistemic construct, and the same applies to data (Kitchin, 2014). Though data might seem inherently present, it is constructed by abstracting the world into categories and representations, which then create the building blocks of information and knowledge (Kitchin, 2014). Algorithms use this constructed data to produce outcomes which shape broader notions of reality, thereby operationalising power by establishing and maintaining specific truths (Beer, 2017).

Epistemic injustice, a concept developed by Fricker (2007), refers to the harm done to individuals in their capacity as knowers. This includes practices that undermine someone’s credibility or the legitimacy of their knowledge based on social prejudices. Fricker (2007) identifies two primary forms of epistemic injustice: testimonial injustice, where prejudice leads to an unfair deflation of credibility, and hermeneutical injustice, which occurs when a gap in collective interpretive resources disadvantages someone trying to make sense of their social experiences. Dotson (2011) expands this concept by introducing epistemic violence, which involves systematically excluding or devaluing marginalised groups' knowledge and experiences. Miñoso and Espinosa (2022) take this further by defining epistemic violence as "making the *other* invisible”.

This paper focuses on a decolonial form of epistemic injustice known as cultural gaslighting, introduced by Ruíz (2020). Unlike interpersonal gaslighting, cultural gaslighting operates structurally within settler colonial societies, systematically invalidating the knowledge and experiences of marginalised, particularly colonised, communities (Ruíz, 2020). This tactic preserves the epistemic advantage of dominant groups by obscuring and distorting structural oppressions, thereby perpetuating colonial domination. While cultural gaslighting has been explored in areas like healthcare, its application to the digital realm remains underexamined. The closest concept in the digital context is the introduction of platform gaslighting by Blunt *et al.* (2020), where platforms manipulate users' perceptions through their moderation practices, however, it lacks the decolonial understanding of bias cultural gaslighting possesses. This dissertation bridges this gap by examining cultural gaslighting as a form of digital colonialism through algorithmic content moderation, particularly in the context of Palestinian content. Both theories are explored in depth in Chapter 4, examining how supposedly neutral content moderation algorithms perpetuate colonial power dynamics in the digital age.

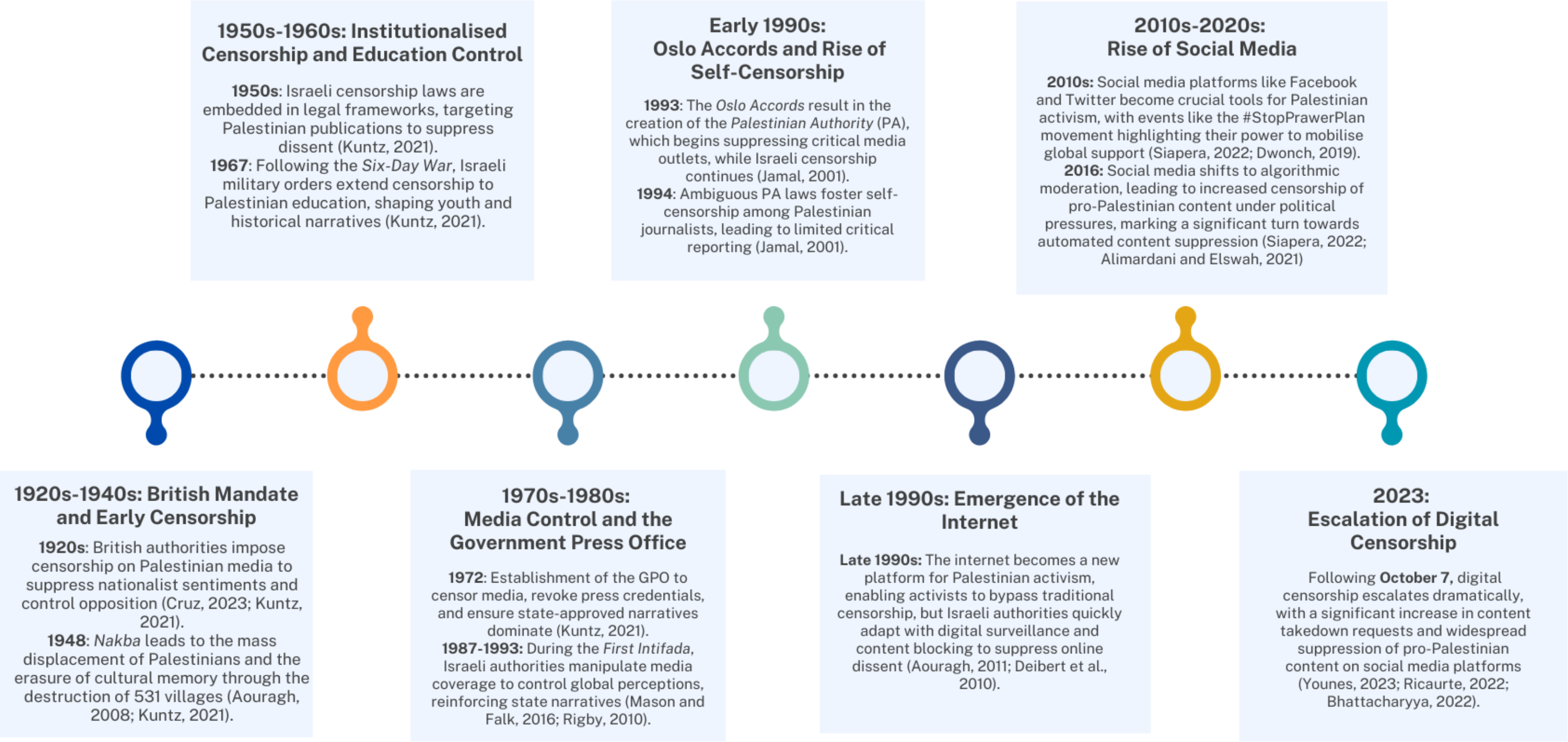
Chapter 2: Historical Evolution of Resistance and Censorship

Historical Context

Following the events of October 7 and the subsequent airstrikes in Gaza, numerous media outlets—including newspapers, magazines, networks, and social media platforms—rushed to document the violence and contextualise it with timelines (Alfonseca, 2023; Cruz, 2023; Narea, 2023; Westfall et al., 2023).. These ranged from simple visualisations to interactive webpages, infographics, and short video essays. Their starting points varied: some traced the origins to the 1800s with the emergence of Zionism (Narea, 2023), others to around 1916 following the partition of Ottoman Empire territories after World War I (Cruz, 2023; Westfall et al., 2023), and many began with the pivotal year of 1948, marked by the establishment of the State of Israel and referred to by Palestinians as the Nakba or ‘catastrophe’ (Alfonseca, 2023). Despite variations in presentation style, all efforts aimed to explain the evolution of violence over the past century.

Said (2000) famously wrote that the greatest battle Palestinians have waged as a colonised people has been over their right to be remembered and reclaim their historical reality. Memory can be manipulated for colonial purposes. Much like physical violence, epistemic violence in the form of censorship and silencing is used to control narratives and rewrite histories, presenting an obstacle to self-determination by questioning such movements' legitimacy. This struggle for control over memory allows groups to assert their right to narrate their history and make demands accordingly. I argue that in colonial contexts, the past legitimises future decisions and actions. In the digital realm, these histories shape how data is formed and understood, determining what we know. In this chapter, I contextualise a specific form of violence, epistemic violence, by outlining the historical evolution of censorship in Palestine. This lays the groundwork for understanding modern digital censorship, like algorithmic content moderation, discussed in subsequent chapters. Figure 2 depicts the major milestones in the censorship of Palestinians.

Figure 2- Timeline of Epistemic Violence In Palestine



1920s-1940s: British Mandate and Early Censorship

1920s: British authorities impose censorship on Palestinian media to suppress nationalist sentiments and control opposition (Cruz, 2023; Kuntz, 2021).
1948: *Nakba* leads to the mass displacement of Palestinians and the erasure of cultural memory through the destruction of 531 villages (Aouragh, 2008; Kuntz, 2021).

1950s-1960s: Institutionalised Censorship and Education Control

1950s: Israeli censorship laws are embedded in legal frameworks, targeting Palestinian publications to suppress dissent (Kuntz, 2021).
1967: Following the *Six-Day War*, Israeli military orders extend censorship to Palestinian education, shaping youth and historical narratives (Kuntz, 2021).

1970s-1980s: Media Control and the Government Press Office

1972: Establishment of the GPO to censor media, revoke press credentials, and ensure state-approved narratives dominate (Kuntz, 2021).
1987-1993: During the *First Intifada*, Israeli authorities manipulate media coverage to control global perceptions, reinforcing state narratives (Mason and Falk, 2016; Rigby, 2010).

Early 1990s: Oslo Accords and Rise of Self-Censorship

1993: The *Oslo Accords* result in the creation of the *Palestinian Authority* (PA), which begins suppressing critical media outlets, while Israeli censorship continues (Jamal, 2001).
1994: Ambiguous PA laws foster self-censorship among Palestinian journalists, leading to limited critical reporting (Jamal, 2001).

Late 1990s: Emergence of the Internet

Late 1990s: The internet becomes a new platform for Palestinian activism, enabling activists to bypass traditional censorship, but Israeli authorities quickly adapt with digital surveillance and content blocking to suppress online dissent (Aouragh, 2011; Deibert et al., 2010).

2010s-2020s: Rise of Social Media

2010s: Social media platforms like Facebook and Twitter become crucial tools for Palestinian activism, with events like the #StopPrawerPlan movement highlighting their power to mobilise global support (Siapera, 2022; Dwonch, 2019).
2016: Social media shifts to algorithmic moderation, leading to increased censorship of pro-Palestinian content under political pressures, marking a significant turn towards automated content suppression (Siapera, 2022; Alimardani and Elswah, 2021)

2023: Escalation of Digital Censorship

Following **October 7,** digital censorship escalates dramatically, with a significant increase in content takedown requests and widespread suppression of pro-Palestinian content on social media platforms (Younes, 2023; Ricaurte, 2022; Bhattacharyya, 2022).

(Jamal, 2001; Aouragh, 2008, 2011; Deibert et al., 2010; Rigby, 2010; Mason and Falk, 2016; Dwonch, 2019; Alimardani and Elswah, 2021; Kuntz, 2021; Bhattacharyya, 2022; Ricaurte, 2022; Siapera, 2022; Cruz, 2023; Younes, 2023)

Early Censorship

During the British Mandate in the 1920s, censorship was a critical tool for maintaining colonial control. British authorities imposed rigorous regulations on Palestinian newspapers, books, and pamphlets to suppress nationalist sentiments and opposition to British rule (Kuntz, 2021; Cruz, 2023).This form of censorship aimed to erase Palestinian memory, identity, and credibility. Colonial powers often manipulated historical narratives to reinforce their dominance, as evident in the British strategy that marginalised Palestinian perspectives (Said, 2000).

An escalation in physical violence was matched by an escalation in epistemic violence. The Nakba in 1948 marked a significant increase in efforts to erase Palestinian identity, as both authority and colonial censorship duties were passed on to the newly established state of Israel (Alfonseca, 2023). The mass displacement of over 800,000 Palestinians and the destruction of 531 villages were deliberate attempts to obliterate their cultural and historical memory (Aouragh, 2011; Kuntz, 2021). Libraries and educational institutions were targeted, and cultural artefacts were destroyed or confiscated (Kuntz, 2021). This erasure of memory aimed to create a narrative void that the new Israeli state could fill with its version of history, a turning point for the use of epistemic violence where erasing cultural memory served as a tool of domination and control (Whitelam, 2018). By controlling the narrative and erasing the history of the Indigenous population, colonial and occupying powers could forge a new national identity that served their interests and legitimised other forms of violence. The destruction of Palestinian villages and the systematic removal of cultural artefacts were not just acts of physical violence but also undermined the legitimacy of Palestinian calls for self-determination.

The Shift to Institutionalised Censorship

Following the establishment of the Israeli state in 1948, censorship laws targeting Palestinian publications were institutionalised. These laws embedded censorship within the state's legal framework to suppress Palestinian dissent or resistance

(Kuntz, 2021). Israeli military orders in the 1960s extended this control to education, restricting content to control the intellectual development of Palestinian youth (Kuntz, 2021). This aimed to indoctrinate Palestinians with a narrative that served Israeli interests, stifling the development of a critical and resistant Palestinian intelligentsia.

This strategy was not unique; similar efforts were seen in France and its colonies post-Napoleon, where education was used to *create* ideal citizens (Said, 2000). Control over education was highly effective due to its insidious nature, a trait later mirrored in digital censorship. By controlling the curriculum, Israeli authorities aimed to shape the historical consciousness of Palestinian youth (Said, 2000). This effort continued the colonial project, using epistemic violence to suppress Palestinian identity, memory, and credibility.

Digital Media and the New Age of Censorship

As media evolved and the world became more interconnected, so did forms of epistemic violence. The establishment of the Government Press Office (GPO) in the 1970s marked a significant shift in Israel's approach to controlling the narrative. Tasked with censoring news and revoking press credentials, the GPO ensured that only state-approved narratives were disseminated, effectively silencing Palestinian voices both within Israel and globally (Kuntz, 2021). This institutionalised media control was a strategic response to the increasing importance of information in shaping public opinion and international perception, especially during the Intifadas.

The First Intifada’s coverage highlighted the power imbalance and humanised the Palestinian struggle, altering global perceptions and challenging the dominant narrative of Israel as the ‘David’ against the Arab ‘Goliath’ (Rigby, 2010). This shift in public opinion was crucial in securing international support for the Palestinian cause, highlighting the importance of narrative control in the battle for self-determination (Mason & Falk, 2016). In response, Israeli authorities recognised the media’s influence on public perception and, through the GPO, intensified efforts to shape media narratives during the First and Second Intifadas (Mason & Falk, 2016). They also increased censorship, strictly controlling media coverage, while targeting Palestinian intellectual resources (Kuntz, 2021). By manipulating media narratives, Israeli authorities aimed to delegitimise Palestinian resistance and justify their actions, reinforcing the Israeli narrative globally (Mason & Falk, 2016).

The Rise of Self-Censorship

The initial success in garnering international sympathy for Palestinians waned due to the GPO’s efforts and ongoing Israeli censorship, which significantly diminished the media’s effectiveness in advocating for Palestinian self-determination (Kuntz, 2021). However, these were not the only factors. Decolonial scholars, primarily focusing on the role of the oppressor, often ignore internal Palestinian censorship. The signing of the Oslo Accords in 1993 and the subsequent establishment of the Palestinian Authority (PA) marked a significant shift, where the PA was given a share of the authority and censorship capabilities (Jamal, 2001).

The PA's new control over media led to the closure of many revolutionary newspapers, due to leadership’s hope for peace through diplomatic agreements (Jamal, 2001). In 1994, PA leader Arafat's security forces arrested over 30 journalists and editors, and the shutdown of low-powered TV stations in the West Bank gave the newly established Palestinian Broadcasting Corporation (PBC) a monopoly (Frankel-Shlosberg, 1996). The continued suppression of critical voices, such as the closure of Al-Nahar for its criticisms of the PA, led other publications to tone down their critiques (Frankel-Shlosberg, 1996). This created a new sociopolitical reality where revolutionary newspaper titles gave way to more conciliatory ones, reflecting journalists’ new role of explaining PA decisions rather than rallying people for self-determination (Jamal, 2001). Following in the PBC’s footsteps, remaining TV and broadcasting media outlets transitioned from a revolutionary stance to a mediator position, further altering the Palestinian media landscape (Jamal, 2001).

Self-censorship became a pressing issue, influenced by Israeli policies and the PA’s unpredictable censorship. Ambiguous laws prohibiting publications against ‘Palestinian Unity,’ a term that was not clearly defined, fostered fear and uncertainty among journalists, leading to widespread self-censorship to avoid repercussions (Jamal, 2001). The self-censorship caused by vague rules and internal divisions, often overlooked by decolonial readings that ignore the PA’s role, is critical to this paper’s discussion of cultural gaslighting in digital spaces, where algorithmic moderation leads to self-censorship for similar reasons, affecting the Palestinian struggle for self-determination.

The Internet: Battleground for Free Speech and Control

The rise of the internet in the late 1990s offered a new avenue for Palestinian activism and expression, circumventing traditional forms of censorship. Activists and journalists used online platforms to mobilise support, disseminate information, and counteract mainstream media biases, making the internet a critical space for resisting dominant portrayals by Israeli and Western media (Aouragh, 2011).

However, Israeli authorities adapted to this new epistemic source by implementing digital surveillance strategies. They monitored Palestinian activists and journalists online, blocking access to Palestinian websites and forums (Aouragh, 2008, 2011). This extended Israeli control into the virtual realm, using technology to enhance traditional censorship methods (Aouragh, 2011). State ownership of internet infrastructure, along with internal surveillance in Palestinian-owned internet cafes, allowed Israeli authorities to control the narrative and suppress dissent (Aouragh, 2011; McMahon, 2014). Thus, the internet did not diminish state control over information but provided new tools for surveillance and censorship (Anon, 2010). In Palestine, these digital tools have been used to monitor, block, and suppress online activism, ensuring Israeli authorities maintain control over the narrative and stifle the potential of the internet as a tool for liberation (Aouragh, 2011).

The Age of Social Media and Automation

Meta’s founder and CEO Mark Zuckerberg (2017) has repeatedly emphasised social media's potential to provide a “platform for all ideas" and give "voice to all people". This vision was initially embraced globally, as platforms like Facebook allowed individuals from heavily censored regions, including Palestine, to broadcast their voices. In 2013, young Palestinians used hashtags such as #StopPrawerPlan and #AngerStrike on Twitter and Instagram to mobilise international support, contributing to the Prawer Plan's withdrawal (Dwonch, 2019)

However, 2016 marked a significant shift towards increased digital repression of pro-Palestine content (Alimardani & Elswah, 2021). In September, Facebook faced threats from Israel to block its platform unless it complied with demands to delete specific content, highlighting growing pressures on social media companies to censor political narratives within the region (Alimardani & Elswah, 2021). This trend continued in 2021 when Israeli officials met with Facebook and TikTok executives to urge the removal of "anti-Israel" content (Alimardani & Elswah, 2021; Hamla, 2023).

2016 also saw a significant change in content moderation practices, with a shift toward algorithmic systems driven by the global political climate and technological advancements. Facebook, for instance, significantly increased its investment in AI technologies to automate content moderation. According to Siapera (2022), this shift to AI-driven moderation disproportionately affects Palestinian users, whose posts and accounts are suspended under vague and misapplied terms like ‘hate speech’ or ‘incitement’. Meta’s algorithms, while intended to be neutral, often embed existing biases and political pressures, perpetuating digital repression (Siapera, 2022).

Social media's dual role as both a powerful platform for grassroots activism and an instrument for digital repression has become even more pronounced since October 7, 2023. In addition to offline silencing and intimidation within Palestine, artists, cultural workers, and academics have faced increased digital censorship by social media platforms across the globe (Younes, 2023). Since October 7, Israel’s Cyber Unit has sent Meta over 5,700 content takedown requests, repeating previous trends (Younes, 2023). It is unclear whether Meta has complied with Israeli requests, but it is evident that Meta now possesses censorship capabilities similar to those of the British Mandate, Israel, and the PA. As the next chapters will explore, the evolution of social media from a liberating force to a controlled ecosystem mirrors broader patterns of epistemic violence, where the control of narratives significantly impacts cultural and social visibility (Bhattacharyya, 2022; Ricaurte, 2022).

Chapter 3: Meta’s Technologies, Policies and Broken Promises

Overview of Content Moderation Technologies and Mechanisms

Meta’s AI-Driven Approach to Content Moderation

Meta has developed a wide range of AI technologies to improve its content moderation capabilities. One key tool is the Linformer model, which analyses content within specific cultural and regional contexts for accurate and sensitive assessments (Meta, 2023c). This is a step in the right direction in dealing with the nuances of moderating its diverse user base. Another significant technology is the Reinforced Integrity Optimiser, which enhances the detection of hate speech and harmful content by learning from online signals and adapting to new behaviour patterns (Meta, 2023c). SimSearchNet identifies subtle variations in content, such as different versions of the same meme where one might be harmful and the other harmless, helping to mitigate the spread of viral misinformation (Meta, 2023c).

To address the multilingual nature of its platforms, Meta uses the XLM-R model, they claim improves the detection of hate speech across various languages, improving the moderation of non-English content (Meta, 2023c). The LLaMA (Large Language Model Meta AI) series, including the latest version LLaMA 3.1, further advances language processing and content moderation capabilities by incorporating additional features for better contextual understanding and accurate detection of harmful content (Meta, 2024a).

Meta’s Mechanisms of Content Detection

Meta employs various mechanisms to enhance its harmful content detection efforts. The Hasher-Matcher-Actioner (HMA) tool automates the hashing process for images and videos, creating unique digital fingerprints for each piece of content and matching them against a database of known violations (Clegg, 2022). This scalable solution allows for continuous updates allowing Meta to respond to new threats (Clegg, 2022).

In addition to hashing, Meta utilises convolutional neural networks (CNNs) for image recognition tasks. CNNs process image data through multiple layers of convolutional filters, learning spatial hierarchies and features such as edges, textures, and patterns, making them highly effective for detecting inappropriate images and identifying specific objects (Goodfellow, Bengio & Courville, 2016). For text and language processing, Meta employs recurrent neural networks (RNNs). Designed for sequential data, RNNs maintain an internal state and process data sequences, making them particularly effective for understanding text context, detecting hate speech, and identifying harmful language (Goodfellow, Bengio & Courville, 2016). These models and mechanisms are part of Meta's broader strategy to integrate advanced AI technologies with ethical standards to ensure fair content moderation.

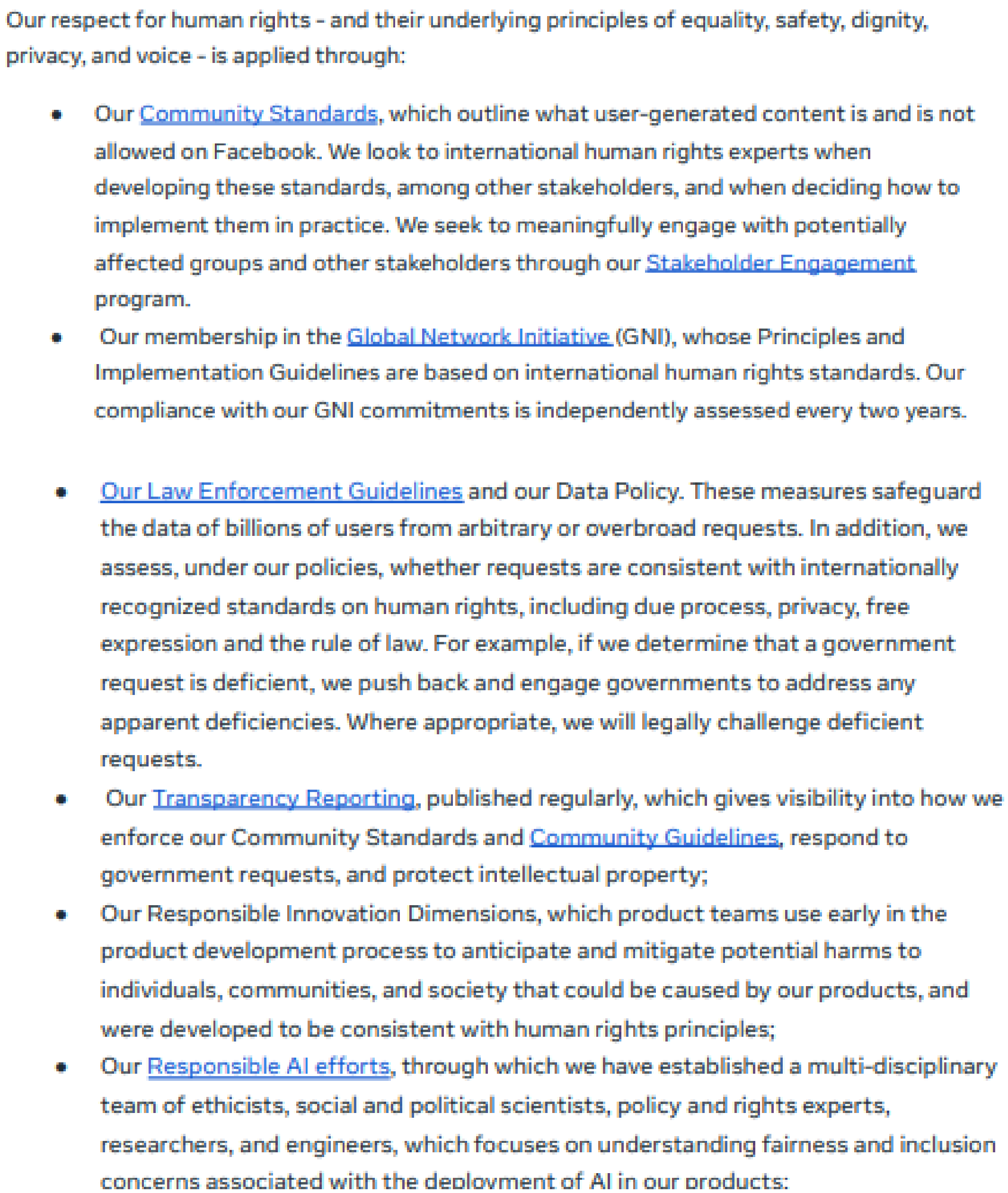
Overview of Meta’s Policies

Meta's approach to content moderation is deeply influenced by its policies, which guide the creation and deployment of its content moderation technologies (Meta, 2023b; Saltman & Hunt, 2023). According to (Meta, 2023b) their development of AI technologies for content moderation is grounded in ethical frameworks that prioritise human rights and societal values. Meta (2023b) claims this ethical framework helps prevent biases and ensures that the technologies do not infringe on user rights.

Corporate Human Rights Policy

Meta’s approach to content moderation is influenced by its *Corporate Human Rights Policy*, which aligns with international human rights standards such as the International Bill of Human Rights and the UN Guiding Principles on Business and Human Rights (Meta, 2021). This policy emphasises key principles including the protection of freedom of expression, ensuring privacy, preventing discrimination, and maintaining user safety (Meta, 2021). These principles guide Meta's efforts to balance the need to manage harmful content with the imperative to protect user rights, ensuring that their content moderation practices respect and uphold the aforementioned fundamental human rights. The policy is maintained through mechanisms such as Human Rights Impact Assessments, stakeholder engagement, and transparency in reporting, which collectively help Meta mitigate risks and enhance positive impacts on human rights (Meta, 2021). Figure 3 is a screenshot of the policy.

Figure 3- Screenshot of Meta's Corporate Human Rights Policy



Our respect for human rights - and their underlying principles of equality, safety, dignity, privacy, and voice - is applied through:

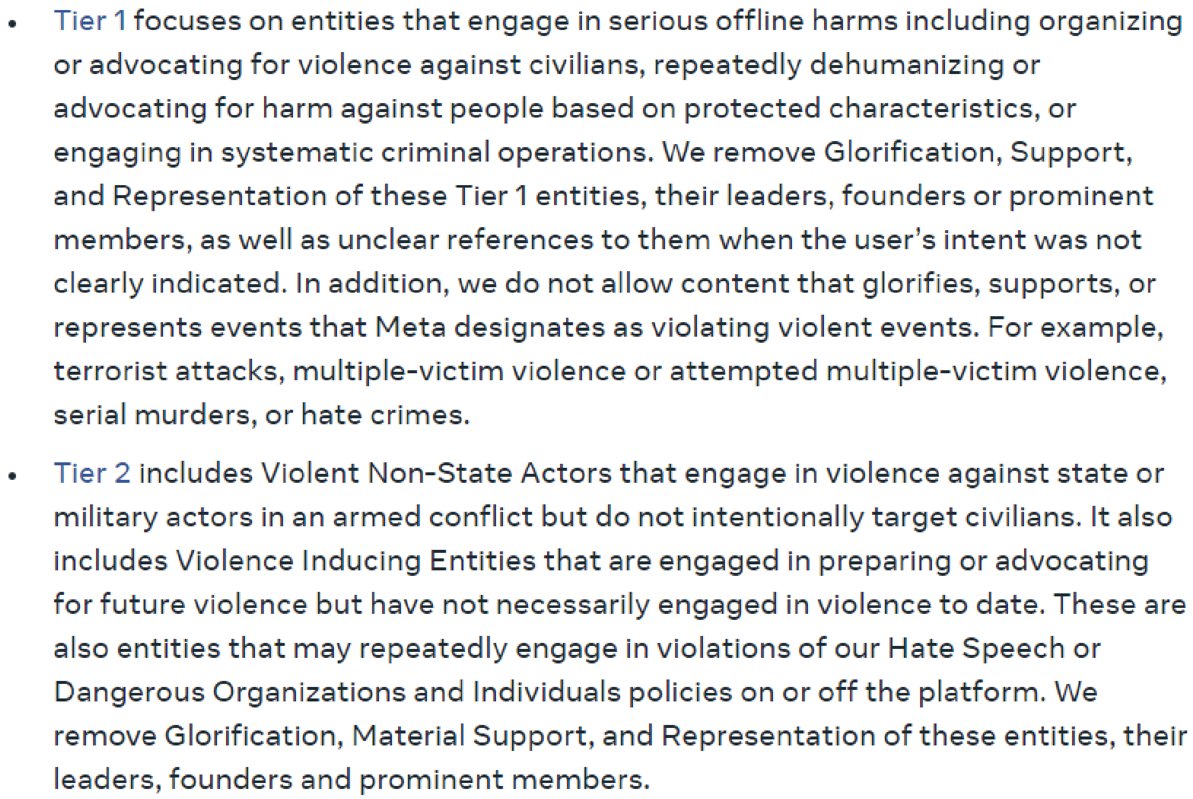
- Our Community Standards, which outline what user-generated content is and is not allowed on Facebook. We look to international human rights experts when developing these standards, among other stakeholders, and when deciding how to implement them in practice. We seek to meaningfully engage with potentially affected groups and other stakeholders through our Stakeholder Engagement program.
- Our membership in the Global Network Initiative (GNI), whose Principles and Implementation Guidelines are based on international human rights standards. Our compliance with our GNI commitments is independently assessed every two years.
- Our Law Enforcement Guidelines and our Data Policy. These measures safeguard the data of billions of users from arbitrary or overbroad requests. In addition, we assess, under our policies, whether requests are consistent with internationally recognized standards on human rights, including due process, privacy, free expression and the rule of law. For example, if we determine that a government request is deficient, we push back and engage governments to address any apparent deficiencies. Where appropriate, we will legally challenge deficient requests.
- Our Transparency Reporting, published regularly, which gives visibility into how we enforce our Community Standards and Community Guidelines, respond to government requests, and protect intellectual property;
- Our Responsible Innovation Dimensions, which product teams use early in the product development process to anticipate and mitigate potential harms to individuals, communities, and society that could be caused by our products, and were developed to be consistent with human rights principles;
- Our Responsible AI efforts, through which we have established a multi-disciplinary team of ethicists, social and political scientists, policy and rights experts, researchers, and engineers, which focuses on understanding fairness and inclusion concerns associated with the deployment of AI in our products;

Source: (Meta, 2021)

The Dangerous Organisations and Individuals Policy

Meta’s Dangerous Organisations and Individuals (DOI) policy outlines the principles and processes for identifying and managing content related to entities that pose a risk to user safety (Meta, 2023a). The DOI policy is a crucial component of Meta's broader content moderation strategy, ensuring that the platform remains safe for all users. The DOI policy includes criteria for defining what constitutes a dangerous individual or organisation, which typically involves assessments of their activities and behaviours both online and offline (Meta, 2023a). In employing the automated detection systems mentioned above and human review to enforce this policy, Meta (2023) claims to value accuracy and fairness in the identification process. Moreover, the policy is regularly reviewed and updated to adapt to new threats and changing societal contexts (Meta, 2023a). Figure 4 provides a screenshot of the different tiers within the DOI Policy.

Figure 4- Screenshot of DOI Tiers



- Tier 1 focuses on entities that engage in serious offline harms including organizing or advocating for violence against civilians, repeatedly dehumanizing or advocating for harm against people based on protected characteristics, or engaging in systematic criminal operations. We remove Glorification, Support, and Representation of these Tier 1 entities, their leaders, founders or prominent members, as well as unclear references to them when the user's intent was not clearly indicated. In addition, we do not allow content that glorifies, supports, or represents events that Meta designates as violating violent events. For example, terrorist attacks, multiple-victim violence or attempted multiple-victim violence, serial murders, or hate crimes.
- Tier 2 includes Violent Non-State Actors that engage in violence against state or military actors in an armed conflict but do not intentionally target civilians. It also includes Violence Inducing Entities that are engaged in preparing or advocating for future violence but have not necessarily engaged in violence to date. These are also entities that may repeatedly engage in violations of our Hate Speech or Dangerous Organizations and Individuals policies on or off the platform. We remove Glorification, Material Support, and Representation of these entities, their leaders, founders and prominent members.

Source: Meta, (2023a)

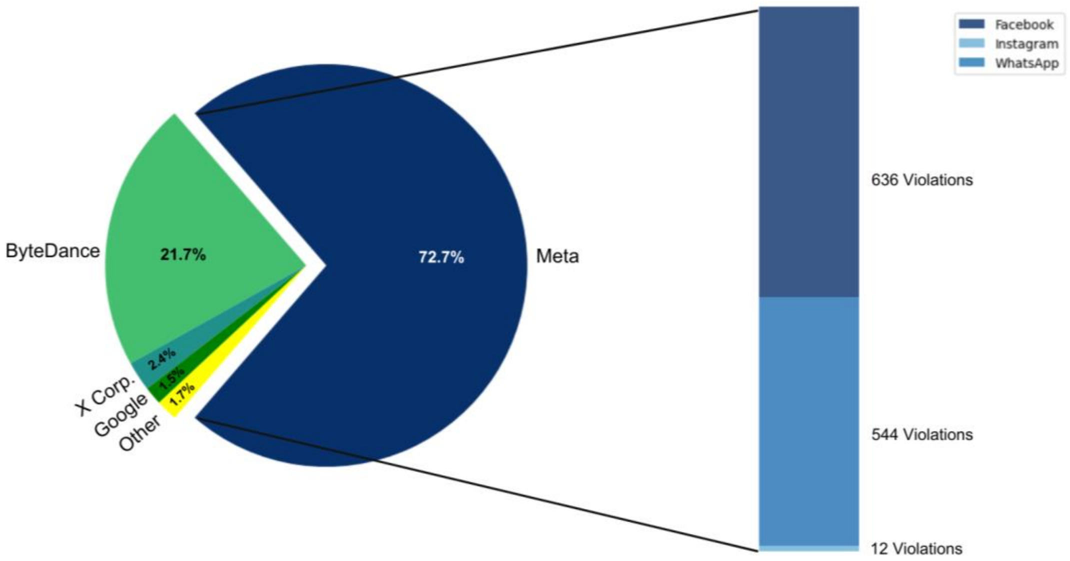
“Meta’s Broken Promises”

Despite these efforts, Meta has been accused of violating human rights by excessively censoring harmless content supportive of Palestine and Palestinians (BSR, 2022; Hamla, 2023; Younes, 2023). These actions infringe on individuals' rights to freedom of expression, especially given the crucial role social media plays in social movements and the digital public sphere. Meta’s restrictions can also be seen to undermine the rights to freedom of assembly, association, and participation in public affairs (Younes, 2023).

HRW’s analysis of content removal on Instagram and Facebook reveals that Meta’s automated moderation tools often struggle to accurately differentiate between peaceful and violent comments (Younes, 2023). These findings were unsurprising to many as such systems are generally limited in their ability to interpret the contextual nuances necessary to determine whether a post supports or glorifies terrorism, leading to overly broad speech restrictions and the improper labelling of content as violent or abusive (Alimardani & Elswah, 2021; Siapera, 2022; Hamla, 2023). Meta has not disclosed data on the error rates of these automated systems or the extent to which automation processes complaints and appeals, raising concerns about algorithmic bias.

HRW determined this through gathering evidence of online censorship from social media users worldwide, confirming 1,049 cases from over 60 countries where harmless speech was removed or suppressed on Meta’s platforms (Younes, 2023). This dataset likely underrepresents the actual volume of censorship, as reports continued to be submitted after the initial cutoff date (Younes, 2023). These findings were not unique to HRW, Hamla's (2024) cross-platform study of censorship in the Palestinian context found 1,191 violations on Meta’s part, far more than any other organisation. Figure 5 shows the distribution of confirmed wrongful censorship of Palestine-related content across different digital platforms.

Figure 5- Breakdown of Reported Censorship Incidents by Company



Source: Hamla (2024)

Six Patterns of Censorship

HRW identified six distinct patterns in the censorship of pro-Palestinian content on Meta’s platforms.:

1. **Post and Comment Removal**: Content removal includes posts, stories, and comments, often without notification. For instance, a Facebook post questioning the justification of killing civilians was removed under Community Standards on “bullying and harassment” (Younes, 2023). Slogans like “From the river to the sea, Palestine will be free,” and comments such as “Free Palestine,” “Ceasefire Now,” and “Stop the Genocide” were frequently removed under “spam” Community Guidelines without considering their context (Hamla, 2023; Younes, 2023). Meta has also hidden the Palestinian flag from comment sections, claiming it harasses or targets others (Younes, 2023).

2. **Account Suspension and Disabling**: There has been a disproportionate number of account suspensions or permanent disabling of pro-Palestinian users, particularly since 2016 (Alimardani & Elswah, 2021). Meta cited

Community Guidelines or Standards violations, often relating to the DOI policy, “adult nudity and sexual activity,” “violent and graphic content,” and “spam” (Alimardani & Elswah, 2021). For example, HRW found that content depicting clothed Palestinians over ruins in Gaza was inaccurately flagged under the “adult nudity and sexual activity” policy (Younes, 2023). The “spam” policy was the most recurrently cited reason behind suspensions, frequently applied erroneously to Palestine-related content (Younes, 2023).

3. **Engagement Restrictions**: Restrictions on engagement, such as limitations on liking, commenting, sharing, and reposting stories, have been imposed for periods ranging from 24 hours to three months (Younes, 2023). These restrictions significantly reduce user interaction with content that supports Palestinian causes, impacting the visibility and reach of such content.

4. **Feature Usage Limitations**: Users discussing Palestinian issues have faced restrictions on features like Instagram/Facebook Live and monetisation. This limits their ability to broadcast live content, generate revenue, and reach a broader audience (Younes, 2023). This is especially problematic given that a growing number of Palestinian users are relying on monetisation features to raise money for crossing the Rafah border to safety (Hamla, 2024) .

5. **Shadow Banning**: Shadow banning involves a significant decrease in the visibility of posts, stories, or accounts without any formal communication from the platform about the restriction (Cotter, 2023). Despite Meta's claim of fixing a “bug” that reduced reach on re-shared Stories, users continued to report these cases (Hamla, 2023, 2024; Younes, 2023). HRW argue this insidious form of censorship makes it difficult for users to understand why their content's visibility has diminished, further stifling free speech (Younes, 2023). Affected users often see a drastic drop in engagement metrics like likes, shares, and comments (Cotter, 2023). As a result, shadow banning can lead to self-censorship, where users refrain from posting certain content to avoid such visibility issues (Younes, 2023).

6. **Search and Discovery Interference**: This involves disabling searches for accounts and reducing content distribution or reach. Users reported difficulty finding accounts that post pro-Palestinian content, further reducing the visibility of such narratives (Younes, 2023).

Recommendations from Civil Society and Oversight Bodies

Reform the DOI Policy

Human Rights Watch and other organisations have consistently called for an overhaul of Meta’s Dangerous Organisations and Individuals (DOI) policy to better align with international human rights standards (BSR, 2022; Younes, 2023; Hamla, 2024). The DOI policy, which bans "praise" and "substantive support" of designated groups (Meta, 2023a), has been criticised for its broad and vague definitions that can cover protected speech (Younes, 2023). The policy’s current definitions of “praise” and “substantive support” are seen as overly restrictive, often leading to the removal of content that neutrally discusses or reports on groups labelled as terrorist organisations (Younes, 2023). The UN Special Rapporteur has expressed concerns that the policy does not comply with international human rights law, particularly regarding its reliance on U.S. terrorist lists, which do not align with global standards (Ní Aoláin, 2021).

Improve Transparency and Reduce Automation

Recommendations have also focused on increasing transparency. Civil society groups and Meta’s Oversight Board have urged Meta to publish the list of designated dangerous organisations and provide more details on how government requests for content removal are handled (Ní Aoláin, 2021; BSR, 2022; Hamla, 2023; Younes, 2023). There have also been calls for increased transparency regarding automated content moderation processes and technologies used in censoring content related to Palestine, including the disclosure of classifiers and their error rates (Younes, 2023; Hamla, 2024). To address concerns around excessive automation, implementing a human-in-the-loop principle—where humans oversee more AI decisions—has been recommended to ensure meaningful oversight, in line with the UN Guiding Principles on Business and Human Rights (UNGPs) (Younes, 2023). In Chapter 5, I present further recommendations based on the theoretical framework developed in Chapter 4.

Chapter 4: Introducing Colonial Platform Gaslighting

Cultural Gaslighting

Cultural gaslighting, as defined by Ruíz (2020), is a strategic tool employed in settler colonial societies to legitimise and perpetuate colonial domination. Unlike interpersonal gaslighting, which manipulates an individual's perception of reality (Johnson et al., 2021), cultural gaslighting functions on a structural level. It creates environments—referred to as "abusive ambients"—designed to systematically undermine and erase the knowledge and experiences of marginalised colonised communities (Ruíz, 2020). These environments encompass languages, narratives, and practices that entrench colonial authority while silencing Indigenous voices.

Cultural gaslighting plays a crucial role in maintaining the epistemic advantage of dominant groups. By curating knowledge that obscures and distorts structural oppressions, it deflects scrutiny from the colonial foundations of power (Ruíz, 2020). Ruíz (2020), argues that this is not an unintended consequence of colonialism that can be attributed to individual bias but a deliberate strategy to suppress resistance, ensuring the ongoing cultural genocide and dispossession of Indigenous peoples. Through control over the production and circulation of meaning (as discussed in Chapter 2), colonisers sustain narratives that justify their continued occupation and exploitation of Indigenous lands, systematically silencing resistance and reinforcing their dominance.

Ruíz (2020) explores these colonising scripts—the above narratives and practices embedded in settler colonial societies that perpetuate structural violence—particularly in medical contexts. Here, they manifest as "medical gaslighting," where the health concerns and practices of Indigenous women and women of colour are systematically downplayed or ignored, as they are not the blueprint (Ruíz, 2020). This reflects the broader structural biases that are crucial to maintaining colonial power, ensuring that marginalised voices are silenced, and their struggles remain unacknowledged.

These women often bear the burden of explaining their struggles to medical professionals and lobbying for change, instead of receiving the healthcare they sought (Ruíz, 2020; Liddell & Lilly, 2022). This dynamic reflects the concept of “settler moves to innocence” (Tuck & Yang, 2012) where settlers and their descendants attempt to absolve themselves of responsibility for the ongoing impacts of colonisation. It often manifests in demands that Indigenous peoples educate colonisers about their oppression, reinforcing colonial power dynamics (Ruíz, 2020). These demands shift the burden of explanation onto the oppressed, allowing colonisers to control the narrative and maintain moral authority (Ruíz, 2020). Cultural gaslighting is thus a form of epistemic violence that not only silences and devalues the knowledge of the ‘other,’ but also shifts the burden of explaining these injustices onto the marginalised, enabling colonisers to evade moral responsibility.

Colonial Platform Gaslighting

In extending Ruíz's (2020) concept of cultural gaslighting to the digital realm, I introduce the idea of *Colonial Platform Gaslighting.* Cultural Gaslighting provides a critical framework for understanding how structural denial and epistemic violence operate within settler colonial societies. I argue that the colonising scripts Ruíz (2020) identifies in medical contexts are also deeply embedded in digital spaces. This framework can also effectively illustrate how algorithmic moderation on social media platforms, such as shadow banning, suppresses content and shapes meaning.

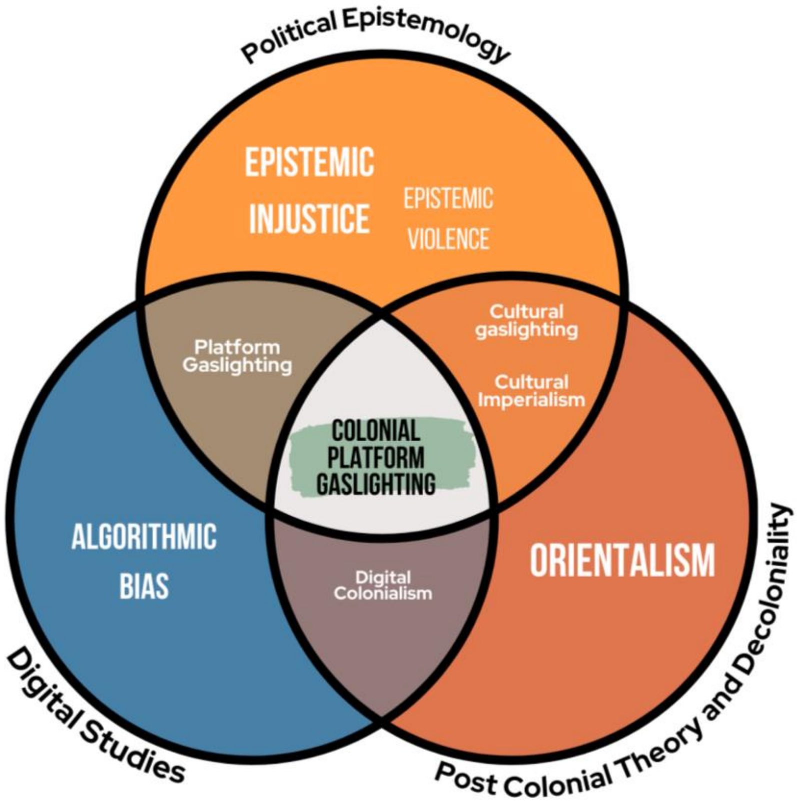
Colonialism has deeply influenced global power dynamics, shaping the distribution of wealth, technological infrastructure, and the design of algorithms. Developed predominantly in the Global North, these algorithms often reinforce existing power hierarchies (Shahid & Vashistha, 2023). The underlying codes in these systems act as *colonising codes*, digital equivalents of the *colonising scripts* discussed earlier. Like colonising scripts, colonizing codes perpetuate structural violence by systematically marginalising and silently silencing oppressed communities, thereby maintaining the epistemic dominance of colonial powers. Unlike the more transparent yet effective scripts, these codes are insidious and often opaque (Younes, 2023).

Colonial platform gaslighting also draws on Blunt *et al.'s* (2020) introduction to platform gaslighting, which refers to the systematic denial by social media companies of practices like shadow banning, where users' content is made invisible without their knowledge. These companies often attribute user concerns to technical glitches or misunderstandings, effectively dismissing the realities of censorship (Blunt et al., 2020). This denial strategy mirrors historical colonial tactics, where oppressive actions were frequently denied or reframed as necessary measures to maintain order or security (Said, 2000; Ruíz, 2020)

While platform gaslighting highlights the denial of censorship practices like shadow banning, it often overlooks the deeper colonising scripts or codes that are embedded within these digital systems. In contexts such as Palestine, or the ongoing suppression of Indigenous and Black online activists (Register et al., 2024), these colonising codes are evident. Content that challenges dominant narratives is not merely suppressed due to technical glitches or vague policies but through the design of algorithms that reinforce existing power structures. In chapter 5, I argue that the categorisation of content as "high-risk" or "low-value" frequently aligns with its potential to disrupt these structures, particularly when it advocates for marginalised communities or contests settler-colonial narratives.

Colonial Platform Gaslighting bridges the gap between decolonial cultural gaslighting and platform gaslighting by revealing how colonial structures are reproduced and enforced through digital platforms, where colonising codes embedded in algorithms perpetuate the same epistemic violence found in offline contexts. In Chapter 5 this framework goes beyond merely identifying platform censorship, instead recognising and challenging the colonial roots within digital technologies. Figure 6 situates Colonial Platform Gaslighting within the existing literature.

Figure 6- Bridging the Gap



We see Colonial Platform Gaslighting in action in the Palestinian context (see Chapter 3). Meta (formerly Facebook) claimed to have resolved a "bug" that affected content related to Palestine, yet reports of shadow banning persisted (Younes, 2023). This scenario is a digital manifestation of historical colonial practices, where the suppression of voices advocating for justice is reframed as necessary for maintaining stability. The same patterns of epistemic violence that Ruíz identifies in medical contexts are present in these digital interactions, but on a global scale and with wider reach and impact. I argue that the reach and impact are wider because this example of colonial platform gaslighting occurred across 60 countries to silence a single narrative (Younes, 2023). This suggests that in the digital age, cultural gaslighting has evolved, with algorithmic moderation now operating on a global scale, affecting millions of users and transcending borders. The automation and speed of these systems intensify gaslighting’s effects, creating new dimensions of harm outlined in the next section.

Settler Steals Innocence: All Eyes on Rafah

The impact of Colonial Platform Gaslighting extends far beyond mere censorship. While Ruíz's (2020) framework highlights the mistrust and silencing that occur between Indigenous women and healthcare workers, and Blunt *et al.* (2020) discuss similar dynamics between users and platforms, the consequences of algorithmic moderation extend even further. The influence of platform gaslighting is particularly insidious, as it shapes both the public narrative and the internal dynamics of decolonial movements.

One outcome of Colonial Platform Gaslighting is self-censorship. Similar to the self-censorship discussed in Chapter 2, users on platforms like Instagram and Facebook, particularly those involved in contentious or politically sensitive issues, often alter their behaviour to avoid having their content removed or restricted (Register et al., 2024). Younes (2023) documents how, after repeated experiences of automated content removal, users become increasingly hesitant to engage in discussions about Israel and Palestine, leading to a significant reduction in their participation and expression. This self-censorship not only stifles individual voices but also diminishes the overall visibility of critical issues, contributing to the broader silencing effect of these platforms.

I argue that as the "settler moves to innocence" within this digital context, there is a deeper layer to the impact: the settler does not just absolve themself of responsibility for ongoing injustices but also deprives the colonised of their innocence or perceived innocence. Activists are forced to make unprecedented and ethically challenging decisions, such as downplaying the severity of issues in their content to avoid censorship and maintain engagement. This dynamic extends beyond traditional media institutions and journalists subject to Israeli or Palestinian censorship (see Chapter 2); it now affects a much broader global community, including supporters of Palestine, who must navigate the risk of being silenced by algorithms designed to filter out "high-risk" content.

This issue is not unique to the Palestinian cause. Indigenous activists, as noted by Register *et al.* (2024) often face similar challenges in balancing the need to educate others with the risk of censorship. Avoiding specific words might help evade algorithmic suppression, but it can dilute the message, leading to internal mistrust and divisions within movements, as activists adopt different strategies to cope with these constraints. The case of "All Eyes on Rafah" illustrates this dynamic.

In May 2024, an AI-generated image depicting tent camps for displaced Palestinians, accompanied by the slogan “All Eyes on Rafah,” went viral on Instagram and was shared over 47 million times (Burga, 2024; Davies, 2024). The image's widespread sharing was partly due to its ability to circumvent Meta’s restrictive content algorithms, as it lacked graphic content and avoided recognisable keywords that might trigger removal (Davies, 2024). However, some activists expressed concern that the image sanitised the reality on the ground, criticising it as performative activism that downplayed the severity of the situation (Davies, 2024; Haughton, 2024). The creator of the image issued a public apology, acknowledging the unintended consequences of their work and navigating the balance between visibility and accountability (Davies, 2024; Haughton, 2024). This incident illustrates a key consequence of Colonial Platform Gaslighting: it forces activists to make difficult decisions that can lead to internal conflict and division.

The broader implications of this are significant. I do not intend to unpack the intentions or validity of individuals' activism, as that is not the focus of my paper. However, I will say the fact that an AI-generated image became the most shared post about the conflict—largely because it could navigate colonial algorithms—illustrates the silencing effect of Colonial Platform Gaslighting. This dynamic not only affects the visibility of certain narratives but also shapes the strategies and cohesion of advocacy movements, creating a landscape where the quest for engagement often comes at the cost of deeper, more honest communication and solidarity.

Chapter 5: Future Directions and Recommendations

In the previous chapters, I extended Ruíz's theory into the digital realm to introduce the concept of colonial platform gaslighting, using Meta's ongoing censorship of Palestine as a case study. Now, I apply this framework to refine the policy recommendations from Chapter 3 and outline a practical path forward. While various avenues exist for scholars, policymakers, and platform developers, my approach remains firmly grounded in verifiable facts. I avoid speculation and do not assume deliberate wrongful moderation or collaboration with Israel, as some have suggested (Younes, 2023; Hamla, 2024). Instead, I base my recommendations on Meta’s publicly shared policies and technologies (see Chapter 3) and the historical trends analysed earlier (see Chapter 2). This allows me to propose reforms that are both practical and achievable.

Remove Technologies from their Orientalist Context

The HRW report discussed earlier primarily focused on wrongful content moderation of English content. However, Younes (2023) suggests that the situation for Arabic content was even worse, a concern echoed by BSR (2022) and Hamla (2024) who found that Meta’s AI struggled more with Arabic content compared to its understanding of content in English and Hebrew. These reports suggest that the issue lies not only in technological shortcomings but also in the orientalist context within which these technologies operate. Rather than merely investing in more technologies (Chapter 3), I argue that Meta must work to extract its technologies from this orientalist framework.

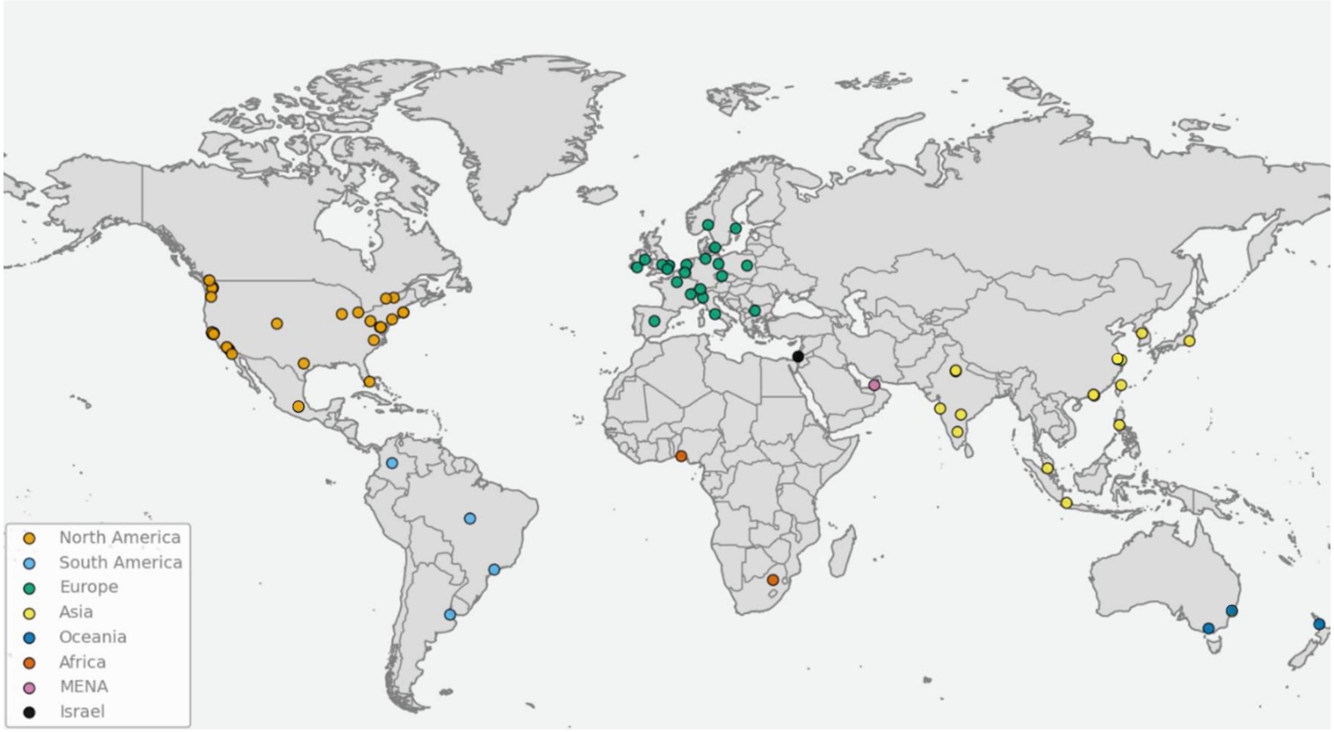
For instance, posts using the hashtag "Al-Aqsa" were censored because the platform conflated content related to the Al-Aqsa Mosque with the Al-Aqsa Brigades, a militant group (Hamla, 2023) Similarly, the term "shahid" or "shaheed," which can refer to a martyr, a person’s name, or even a popular Arabic streaming service, is frequently censored by Meta’s automated systems (Singh, 2024). In October 2023, Instagram apologised for wrongly labelling Palestinian users as "terrorists" for using the Arabic phrase "alhamdulillah" alongside the Palestinian flag emoji (Younes, 2023).

This leads me to argue that the issue is not purely technological but deeply contextual. Meta operates two offices in the MENA region: one in Tel-Aviv which covers Israel and one in Dubai for the rest of the region (Meta, 2024b). While the Tel Aviv office may adequately handle Hebrew content—leading to fewer wrongful takedowns of Israeli posts (BSR, 2022)—the Dubai office alone is insufficient for managing the 28+ Arabic dialects and over 60 languages spoken across the MENA region including Palestine and Arabic-speaking areas of Israel (Eberhard, Simons & Fennig, 2024).

This lack of regional representation can be understood through Orientalism, where the *other* is treated as a monolith, disregarding the region's linguistic and cultural diversity. This leads to cultural imperialism, where Western-centric perspectives and understandings dominate the digital landscape, distorting how content from non-Western regions is moderated. I argue this is one of the primary causes behind the colonial platform gaslighting outlined in Chapter 4, where minimal regional representation perpetuates an unequal power dynamic reminiscent of colonialism.

Figure 7 illustrates this disparity, showing a concentration of Meta offices and personnel in the Global North, particularly in the USA and Europe. This imbalance highlights the need to decolonise these platforms and AI technologies by involving more people from the region in their development and decision-making processes. As Meta’s user base and epistemic relevance in the Global South continue to grow, so too should the representation of these users in key decision-making roles.

Figure 7- Mapping Meta's Office Locations



Source: (Meta, 2024b)

Look Beyond a Universal Standard

As discussed in Chapter 3, Meta's corporate human rights policy heavily relies on universal human rights frameworks. While this might seem commendable, this reliance is problematic, as the language of rights is often co-opted by those with power. After 9/11, Western leaders, for example, used feminist rhetoric about the oppression of Afghan women to justify military interventions, illustrating how human rights discourse can be manipulated to serve the interests of powerful states (Peck, 2011). States like the U.S., despite advocating for universal human rights, often evade accountability for their own violations, benefiting from perceived immunity. (Peck, 2011). This lack of accountability is deeply rooted in the ways these states amassed the power to shape rights—through colonialism, slavery, and dispossession (Peck, 2011). Meta's policies, therefore, do not exist in a vacuum; they reflect broader power dynamics that shape their implementation and impact.

This idea is echoed by scholars like Siapera (2022) who criticise the reliance on Western ethics frameworks in addressing issues of race and coloniality in technology. Content moderation practices disproportionately affect marginalized groups, such as Black users, trans users, activists, and LGBTQ communities; for example, anti-white supremacist rhetoric is often flagged as hate speech, while white supremacy itself is not (Siapera, 2022; Register et al., 2024). In the context of Israel-Palestine, criticism of the Israeli government, no matter how carefully worded, is often removed from platforms like Instagram (Younes, 2023).

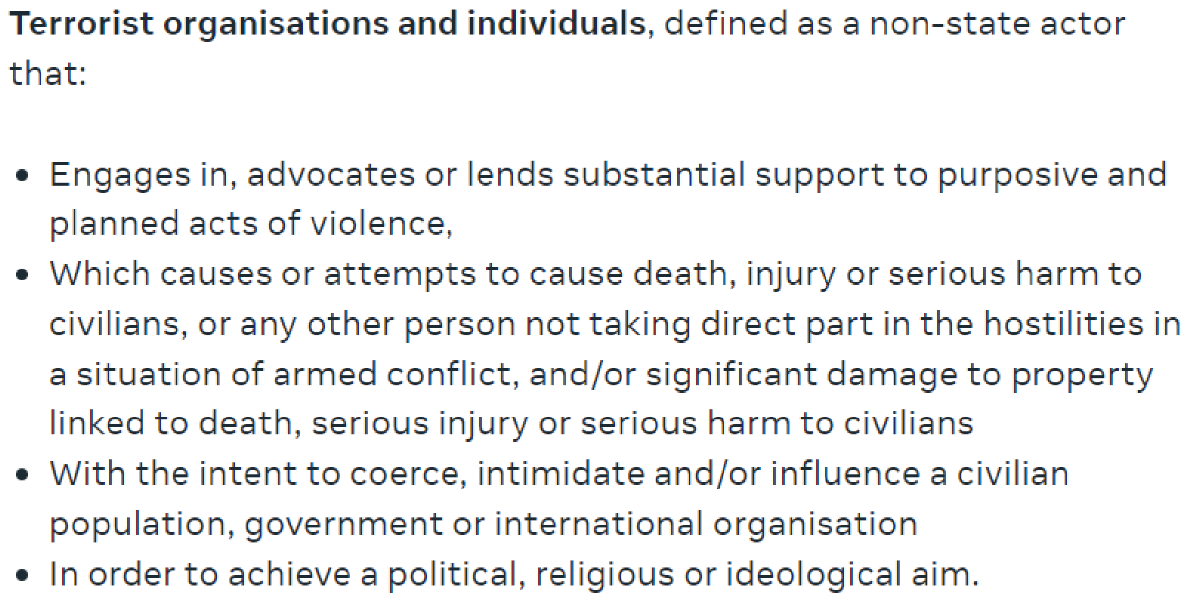
This reveals a major flaw in Meta’s policies: the failure to distinguish between oppressors and the oppressed. Although Meta acknowledges the challenges in moderating hate speech, particularly when it is used self-referentially or to raise awareness (Meta, 2023b), its policies often dismiss the perspectives of the most affected communities, perpetuating colonial dynamics (Siapera, 2022). A recent example of this was Meta's (2024c) formal announcement that it would remove all anti-Zionist speech in July.

To address these issues, I suggest that Meta must move beyond a universal approach to upholding rights, which overlooks historical and ongoing contexts, and instead meaningfully engage with the specific needs of oppressed groups. This requires not only greater representation within the company but also deeper collaboration with both Western and non-Western organisations, such as Hamla, to ensure content moderation practices are more transparent, inclusive, and sensitive to the diverse realities of global users.

Reform the Dangerous Definition

I argue that Meta's DOI policy is flawed not only due to its vagueness (see Chapter 3) but also because of the binary framework it uses to categorise terrorist organisations and individuals exclusively as non-state actors. This binary framework is not just a technicality; it reflects a wider worldview rooted in IR theories like liberalism and realism (see Chapter 1). However, this binary approach falls short in fairly assessing contexts of settler colonialism and apartheid. Figure 8 is a snapshot of Meta’s definition of terrorist organisations and individuals.

Figure 8- Meta’s definition of Terrorist Organisations and Individuals



Terrorist organisations and individuals, defined as a non-state actor that:

- Engages in, advocates or lends substantial support to purposive and planned acts of violence,
- Which causes or attempts to cause death, injury or serious harm to civilians, or any other person not taking direct part in the hostilities in a situation of armed conflict, and/or significant damage to property linked to death, serious injury or serious harm to civilians
- With the intent to coerce, intimidate and/or influence a civilian population, government or international organisation
- In order to achieve a political, religious or ideological aim.

Source: (Meta, 2023a)

In colonial contexts, I argue the state versus non-state binary inherently favours the settler state, recognised as the legitimate authority, over Indigenous or occupied populations, who are typically labelled as non-state actors. This dichotomy ignores the complexities of power dynamics, historical injustices, and governance legitimacy. The settler state is often viewed as the "legitimate" actor, while Indigenous resistance or support for occupied populations is easily framed as illegitimate or in some cases, terrorist (Ak, 2024). This framing reinforces the settler state's dominance and marginalises the voices and rights of indigenous or occupied people.

For example, Meta’s DOI policy often equates expressions of Palestinian support with Hamas the de facto governing body of the Gaza Strip (not the entirety of Palestine) and a recent addition to the US terror list (Younes, 2023; Hamla, 2024). I argue this is largely because the State of Palestine lacks universal recognition, including by the United States, where Meta is based, and where U.S. terrorist designations heavily influence the platform's policies (See Chapter 3). As a result, harmless content that expresses support for Palestine is frequently flagged as violent or abusive (Younes, 2023) leading to wrongful censorship. Numerous instances have been documented where such content was mistakenly removed under the DOI policy (see Chapter 3), highlighting the biases inherent in this binary framework.

The policy's exclusion of state actors from being classified as terrorists creates a significant blind spot. State terrorism—violent acts committed by a government against its own citizens or others for political purposes—is widely recognised in academic and legal contexts (Primoratz, 2005). Meta's policy restricts discussions that condemn state-perpetrated violence, effectively silencing critiques of Israel's actions against Palestinians, which some scholars and activists describe as state terrorism (Pieterse, 1984; Ak, 2024). This rigid state/non-state binary has serious real-world implications, as it stifles essential discourse on such violence.

As we have seen before, even among other platforms, Meta seems to be the most aggressive in its censorship efforts (see Figure 5). Outside of Meta, discussions about state-based violence occur in a more open and nuanced epistemic environment, allowing for critical examination. However, on Meta, the binary classification imposed by the DOI policy stifles such discourse, creating a contrast between the two epistemic environments and perpetuating its own form of epistemic violence. By maintaining this binary, Meta’s DOI policy reinforces the status quo, upholding the settler state's dominance and silencing alternative narratives, effectively engaging in Colonial Platform Gaslighting.

To address these issues, I suggest that Meta reforms its DOI policy to move beyond the simplistic state versus non-state binary. The policy should reflect the complexities of modern geopolitical realities, acknowledging that state actors can engage in actions warranting scrutiny and that non-state actors are not inherently illegitimate. Such reform would enable more equitable content moderation and create a digital space where diverse perspectives, including from marginalised groups, can be heard and understood in their full context, rather than being dismissed by a rigid binary framework.

Future Pathways

As stated above, Meta should reform its content moderation policies by implementing revised definitions that move away from state-centric biases and take their technologies out of their orientalist context. In the meantime, they could engage in beta testing these changes to evaluate how they impact the rate of wrongful censorship. While the primary responsibility for these reforms lies with Meta, academics and data scientists also have a crucial role in this process. By collaborating to refine and implement these revised definitions into the open-source code provided by Meta, the research community can help improve the accuracy and fairness of content moderation. In my upcoming research, I will focus on testing the practical application of these changes by conducting small-scale experiments using the modified open-source code from Meta and other platforms. This approach will allow me to assess the effectiveness of these revisions in real-world scenarios and develop my framework accordingly.

Conclusion

This paper maps the historical evolution of censorship in the Palestinian context, beginning with traditional methods used by British and Israeli colonial powers. Over time, censorship becomes institutionalised, with both Israeli and Palestinian authorities engaging in narrative control. In the digital age, platforms like Meta inherit this legacy, using content moderation algorithms as modern instruments of censorship.

Meta's algorithms disproportionately suppress Palestinian content, mirroring historical tactics of silencing Palestinian voices. For example, shadow banning resembles the subtle yet pervasive educational controls of the past, where suppression lacks tangible proof but is deeply felt (Said, 2000). Meta's vague DOI policy fosters self-censorship (Younes, 2023), akin to how the Palestinian Authority's ambiguous rules lead to self-censorship among Palestinians (Jamal, 2001).

This paper identifies digital censorship on platforms like Meta as a form of epistemic violence. I introduce the framework of Colonial Platform Gaslighting, which extends the concept of cultural gaslighting into the digital realm. Bridging digital studies with decolonial and postcolonial theory, as well as political epistemology, it illustrates how digital platforms, under the guise of neutrality and universal rights, systematically invalidate and suppress marginalised narratives, perpetuating colonial epistemic dominance. Instances such as Meta's denial of shadow banning by attributing it to a bug (Younes, 2023), and the banning of anti-Zionist speech without considering context (Meta, 2024c), demonstrate colonial platform gaslighting in action.

The fear of algorithmic suppression leads users, especially those advocating for Palestinian rights, to engage in self-censorship or dilute their messages (Younes, 2023). This not only reduces the visibility of crucial issues but also has broader implications for global advocacy and the ability of marginalised groups to mobilise. The internal divisions caused by the "All Eyes on Rafah" campaign (Burga, 2024) demonstrate how biased moderation and policies can fracture liberation movements, adding a new dimension to Ruiz’s (2020) concept of cultural gaslighting—where individuals begin to question and invalidate each other rather than just themselves.

While the engagement in colonial platform gaslighting is not assumed to be deliberate, it is attributed to historical standards and the orientalist context in which Meta’s technologies operate. To address these issues, I recommend three key actions: First, Meta should extract its technologies from this orientalist framework by increasing regional representation in the MENA region. Second, Meta must move beyond a universal human rights approach that often overlooks the historical and ongoing contexts of oppression. Third, Meta should reform its DOI policy to transcend the simplistic state versus non-state binary, acknowledging the complexities of modern geopolitical realities.

The study has its limitations, particularly the inability to conduct a quantitative content analysis using relevant datasets from Apodaca and Uzcátegui-Liggett's (2024) investigation on Instagram’s automated censorship of Palestinian content and the Palestinian Observatory of Digital Rights Violations database (7or, 2024). Access to these datasets would have enabled a more detailed examination of algorithmic suppression patterns, providing concrete examples to further develop the framework of Colonial Platform Gaslighting. Additionally, the opacity of Meta’s algorithms restricts the depth of their critique and analysis.

Future research aims to further mitigate these limitations by testing the practical application of proposed changes through small-scale experiments using modified open-source code from Meta and other platforms. Continued efforts will be made to gain access to relevant datasets to further refine the framework. This paper contributes to ongoing discussions about digital rights and censorship, advocating for a re-evaluation of digital platform practices and the decolonisation of digital spaces to ensure fair and equitable treatment of all voices.

Appendix 1- Code Used

For bar of pie chart:

```
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.patches import ConnectionPatch
from matplotlib import cm
overall_labels = ["Meta", "ByteDance", "X Corp.", "Google", "Others"]
overall_sizes = [1192, 356, 39, 24, 26 + 2]
meta_color = "#08306b"  # Dark blue for Meta
bytedance_color = cm.viridis(0.7)
xcorp_color = cm.viridis(0.5)
google_color = "green"
others_color = "yellow"
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(14, 8))
fig.subplots_adjust(wspace=0)
explode = [0.1] + [0] * (len(overall_labels) - 1)
angle = -180 * overall_ratios[0]
wedges, texts, autotexts = ax1.pie(overall_sizes, autopct='%1.1f%%', startangle=angle,
                          labels=overall_labels, explode=explode,
                          colors=[meta_color, bytedance_color, xcorp_color, google_color, others_color])
for text, autotext in zip(texts, autotexts):
    text.set_rotation(45)
    text.set_fontsize(12)
    autotext.set_rotation(-45)
    autotext.set_fontsize(12)
    text.set_horizontalalignment('right')
bottom = 1
width = 0.2
for j, (height, label, color) in enumerate(reversed(list(zip(meta_sizes, meta_labels, [meta_color] *
len(meta_labels))))):
    bottom -= height / sum(meta_sizes)
    bc = ax2.bar(0, height / sum(meta_sizes), width, bottom=bottom, color=color, label=label,
            alpha=0.8)
    ax2.bar_label(bc, labels=[f"{height}"], label_type='center', fontsize=12, weight='bold')
ax2.set_title('Meta Platforms Breakdown')
ax2.legend()
ax2.axis('off')
ax2.set_xlim(-2.5 * width, 2.5 * width)
theta1, theta2 = wedges[0].theta1, wedges[0].theta2
center, r = wedges[0].center, wedges[0].r
bar_height = sum(meta_ratios)
x = r * np.cos(np.pi / 180 * theta2) + center[0]
y = r * np.sin(np.pi / 180 * theta2) + center[1]
con = ConnectionPatch(xyA=(-width / 2, bar_height), coordsA=ax2.transData,
                xyB=(x, y), coordsB=ax1.transData)
con.set_color([0, 0, 0])
con.set_linewidth(2)
```

```
ax2.add_artist(con)
x = r * np.cos(np.pi / 180 * theta1) + center[0]
y = r * np.sin(np.pi / 180 * theta1) + center[1]
con = ConnectionPatch(xyA=(-width / 2, 0), coordsA=ax2.transData,
                      xyB=(x, y), coordsB=ax1.transData)
con.set_color([0, 0, 0])
con.set_linewidth(2)
ax2.add_artist(con)

plt.show()
```

I followed a tutorial by The Matplotlib Development Team (2024) to help achieve the above.

For the map:

The following code was generated with the help of OpenAI ChatGPT (2024) as I knew how to generate choropleth maps but needed a method to create a map with markers. I first pasted the locations to create the data frame then asked what libraries to load/download and how to go about it. I then went over the code and fixed errors. I also followed a tutorial by Project Pythia Community (2024) for further guidance.

```
import matplotlib.pyplot as plt
import cartopy.crs as ccrs
import cartopy.feature as cfeature
import pandas as pd
from geopy.geocoders import Nominatim

color_palette = {
    'North America': '#E69F00',
    'South America': '#56B4E9',
    'Europe': '#009E73',
    'Asia': '#F0E442',
    'Oceania': '#0072B2',
    'Africa': '#D55E00',
    'MENA': '#CC79A7',
    'Israel': '#000000'
}
```

I asked ChatGPT (2024) to turn the data I had pasted into this line of code/define it in JupyterLab and group it for me

```
data = {
    'City': [
        'Austin, TX', 'Bellevue, WA', 'Boston, MA', 'Burlingame, CA', 'Cambridge, MA', 'Chicago, IL',
        'Denver, CO', 'Detroit, MI', 'Durham, NC', 'Foster City, CA', 'Fremont, CA', 'Irvine, CA',
        'Los Angeles, CA', 'Menlo Park, CA', 'Miami, Florida', 'Montreal, Canada', 'Mountain View, CA',
        'New York, NY', 'Newark, CA', 'Northridge, CA', 'Ottawa, Canada', 'Pittsburgh, PA',
        'Redmond, WA', 'Reston, VA', 'San Diego, CA', 'San Francisco, CA', 'Sausalito, CA',
        'Seattle, WA', 'Sunnyvale, CA', 'Vancouver, Canada', 'Vancouver, WA', 'Washington, DC',
        'Bogotá, Colombia', 'Brasilia, Brazil', 'Buenos Aires, Argentina', 'Mexico City, Mexico',
        'São Paulo, Brazil', 'Amsterdam, Netherlands', 'Berlin, Germany', 'Brussels, Belgium',
        'Cambridge, UK', 'Copenhagen, Denmark', 'Cork, Ireland', 'Dubai, United Arab Emirates',
        'Dublin, Ireland', 'Geneva, Switzerland', 'Hamburg, Germany', 'Johannesburg, South Africa',
        'Lagos, Nigeria', 'Leamington Spa, UK', 'London, UK', 'Madrid, Spain', 'Milan, Italy',
        'Oslo, Norway', 'Paris, France', 'Prague, Czech Republic', 'Rome, Italy', 'Sofia, Bulgaria',
        'Stockholm, Sweden', 'Tel Aviv, Israel', 'Warsaw, Poland', 'Zurich, Switzerland',
        'Auckland, New Zealand', 'Bangalore, India', 'Gurgaon, India', 'Hong Kong', 'Hyderabad, India',
        'Jakarta, Indonesia', 'Kuala Lumpur, Malaysia', 'Manila, Philippines', 'Melbourne, Australia',
        'Mumbai, India', 'New Delhi, India', 'Seoul, South Korea', 'Shanghai, China', 'Shenzhen, China',
        'Suzhou, China', 'Sydney, Australia', 'Taipei, Taiwan', 'Tokyo, Japan'
    ],
    'Group': [
        'North America', 'North America', 'North America', 'North America', 'North America', 'North America',
        'North America', 'North America', 'North America', 'North America', 'North America', 'North America',
        'North America', 'North America', 'North America', 'North America', 'North America',
        'North America', 'North America', 'North America', 'North America', 'North America',
        'North America', 'North America', 'North America', 'North America', 'North America',
        'North America', 'North America', 'North America', 'North America', 'North America',
        'South America', 'South America', 'South America', 'South America',
        'South America', 'Europe', 'Europe', 'Europe',
        'Europe', 'Europe', 'Europe', 'MENA',
        'Europe', 'Europe', 'Europe', 'Africa',
        'Africa', 'Europe', 'Europe', 'Europe',
        'Europe', 'Europe', 'Europe', 'Europe', 'Europe',
        'Europe', 'Europe', 'Israel', 'Europe', 'Europe',
        'Oceania', 'Asia', 'Asia', 'Asia', 'Asia',
        'Asia', 'Asia', 'Asia', 'Oceania',
        'Asia', 'Asia', 'Asia', 'Asia',
        'Asia', 'Oceania', 'Asia', 'Asia'
    ]
}
df_locations = pd.DataFrame(data)
geolocator = Nominatim(user_agent="geoapiExercises")
def get_lat_lon(city):
    location = geolocator.geocode(city)
    if location:
        return location.latitude, location.longitude
    else:
        return None, None
df_locations['Latitude'], df_locations['Longitude'] = zip(*df_locations['City'].apply(get_lat_lon))
df_locations = df_locations.dropna()

import cartopy.io.shapereader as shpreader
from cartopy.feature import ShapelyFeature
```

```
# Create a map with Cartopy  (ChatGPT, 2024)
fig, ax = plt.subplots(figsize=(20, 12), subplot_kw={'projection': ccrs.Miller()})
ax.set_global()
ax.set_facecolor('#F2F4F4')
land_50m = cfeature.NaturalEarthFeature('physical', 'land', '50m', facecolor='#DCDCDC',
edgecolor='grey', linewidth=1.2)
ax.add_feature(land_50m)
ax.add_feature(cfeature.OCEAN, facecolor='#F2F4F4')
ax.set_extent([-180, 180, -60, 90], crs=ccrs.PlateCarree())

# Add coastlines and borders with slightly thicker outlines (ChatGPT, 2024)
ax.add_feature(cfeature.COASTLINE, edgecolor='grey', linewidth=1.2)
ax.add_feature(cfeature.BORDERS, edgecolor='grey', linewidth=1.2)

for group, color in color_palette.items():
    group_df = df_locations[df_locations['Group'] == group]
    ax.scatter(
        group_df['Longitude'],
        group_df['Latitude'],
        color=color,
        s=80,
        alpha=0.9,
        marker='o',
        edgecolor='k',
        transform=ccrs.PlateCarree(),
        label=group,
        zorder=5
    )

# Add the legend in the original position
legend = ax.legend(loc='lower left', fontsize=14, frameon=True, facecolor='white',
edgecolor='grey', labelcolor='grey')

# Show the plot
plt.show()
```

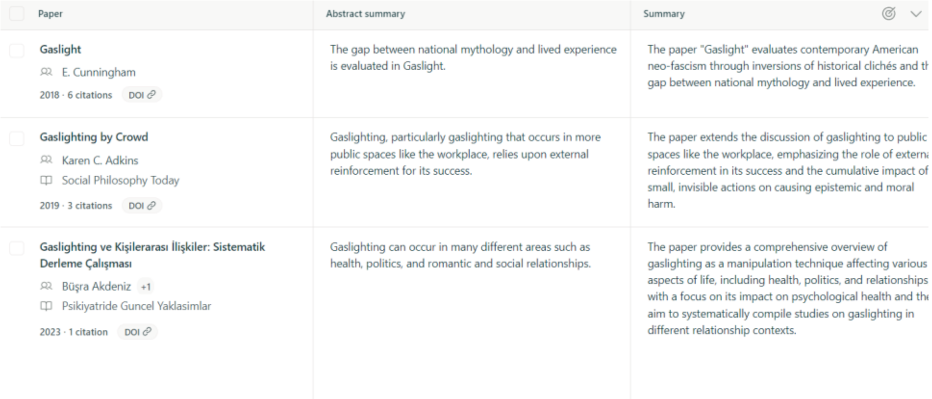
I used Canva and PowerPoint to manually create the rest of the visualisations.

Appendix 2- Use of GAITs

Elicit AI

Me: find papers on cultural gaslighting (believe it was initially explored by elena ruiz and its kind of a decolonial outlook on epistemic harms i suppose if you are unfamiliar). anything related to algorithms, digital cultural gaslighting etc would be a bonus but any literature on cultural gaslighting is the priority. take your time. also inform me if anyone has looked at cultural gaslighting through algorithms before or if this is indeed a gap in the literature

Elicit: Cultural gaslighting, a concept initially explored by Elena Ruiz, is a form of manipulation that causes the victim to doubt their own reality and is often reinforced by external factors (Adkins, 2019). It can occur in various contexts, including health, politics, and relationships (Akdeniz, 2023). However, there is a lack of research on this topic, particularly in non-Western contexts (Akdeniz, 2023). While there is a growing body of literature on the harmful and discriminatory effects of algorithms (Maalsen, 2023), there is a gap in the literature on the intersection of cultural gaslighting and algorithms. This presents an opportunity for future research to explore the potential for algorithms to perpetuate or mitigate cultural gaslighting.

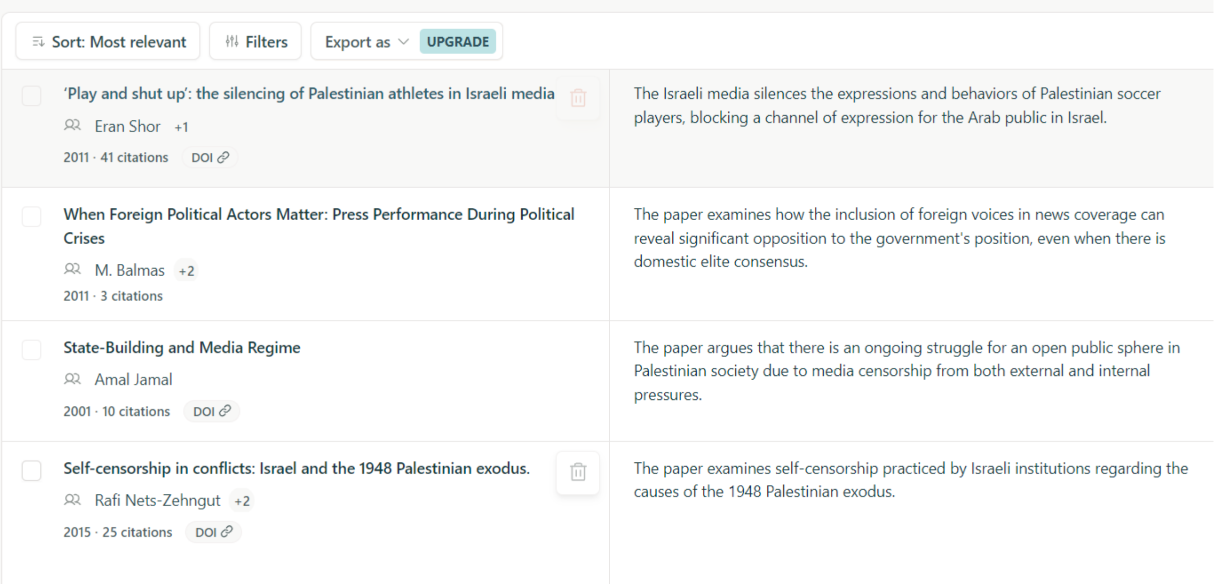


☐ Paper	Abstract summary	Summary
☐ **Gaslight** E. Cunningham 2018 · 6 citations DOI	The gap between national mythology and lived experience is evaluated in Gaslight.	The paper "Gaslight" evaluates contemporary American neo-fascism through inversions of historical clichés and th gap between national mythology and lived experience.
☐ **Gaslighting by Crowd** Karen C. Adkins Social Philosophy Today 2019 · 3 citations DOI	Gaslighting, particularly gaslighting that occurs in more public spaces like the workplace, relies upon external reinforcement for its success.	The paper extends the discussion of gaslighting to public spaces like the workplace, emphasizing the role of externa reinforcement in its success and the cumulative impact of small, invisible actions on causing epistemic and moral harm.
☐ **Gaslighting ve Kişilerarası İlişkiler: Sistematik Derleme Çalışması** Büşra Akdeniz +1 Psikiyatride Guncel Yaklasimlar 2023 · 1 citation DOI	Gaslighting can occur in many different areas such as health, politics, and romantic and social relationships.	The paper provides a comprehensive overview of gaslighting as a manipulation technique affecting various aspects of life, including health, politics, and relationships with a focus on its impact on psychological health and th aim to systematically compile studies on gaslighting in different relationship contexts.

Note: This above literature table was not helpful, I did not end up using any of the sources. I did, however, find the reassurance that I had identified a gap in the literature helpful.

Me: i want papers that refute/argue against the fact that there has been israeli censorship of palestinian voices

Elicit: These papers challenge the notion of Israeli censorship of Palestinian voices, highlighting complex media dynamics in the region. While Palestinian artists and institutions face multiple layers of censorship, including self-censorship and external pressures (Persekian, 2020), the issue is not solely attributable to Israeli actions. The Palestinian Authority (PA) itself has been criticized for limiting press freedom and using censorship during its self-governance period (Nossek & Rinnawi, 2003). Israeli media policies have been designed to treat Palestinian-Israelis as a linguistic minority, limiting their participation in national culture-building (Schejter, 2008). However, during political crises, Israeli national newspapers have shown the ability to produce counterviews challenging government positions and increasingly include Palestinian voices over time (Balmas et al., 2015). These findings suggest that media censorship in the region is a multifaceted issue involving various actors and political dynamics, rather than a simple case of Israeli suppression of Palestinian voices.



☐ 'Play and shut up': the silencing of Palestinian athletes in Israeli media Eran Shor +1 2011 · 41 citations DOI	The Israeli media silences the expressions and behaviors of Palestinian soccer players, blocking a channel of expression for the Arab public in Israel.
☐ When Foreign Political Actors Matter: Press Performance During Political Crises M. Balmas +2 2011 · 3 citations	The paper examines how the inclusion of foreign voices in news coverage can reveal significant opposition to the government's position, even when there is domestic elite consensus.
☐ State-Building and Media Regime Amal Jamal 2001 · 10 citations DOI	The paper argues that there is an ongoing struggle for an open public sphere in Palestinian society due to media censorship from both external and internal pressures.
☐ Self-censorship in conflicts: Israel and the 1948 Palestinian exodus. Rafi Nets-Zehngut +2 2015 · 25 citations DOI	The paper examines self-censorship practiced by Israeli institutions regarding the causes of the 1948 Palestinian exodus.

Note: I did end up using Amal Jamal's paper in my dissertation so I made sure to include my use of (Elicit, 2024) as part of my literature search methodology. The remainder of the questions I asked generated results that were not helpful to me (I did not have access to most papers and many were in Spanish) and I eventually ran out of free tokens so stopped using it.

Copilot

While proofreading, I used Copilot (2024) to go over paragraphs where I had used many commas to check for comma splices. Below is an example of a time it noticed one due to issues with spacing.

You

Sent by you:

check this paragraph for comma splices: Cultural gaslighting, as defined by Ruíz (2020), is a strategic tool employed in settler colonial societies to legitimise and perpetuate colonial domination. Unlike interpersonal gaslighting, which manipulates an individual's perception of reality, cultural gaslighting functions on a structural level. It creates environments—referred to as "abusive ambients"—designed to systematically undermine and erase the knowledge and experiences of marginalised communities(Ruíz, 2020). These environments encompass languages, narratives, and practices that entrench colonial authority while silencing Indigenous voices.

Copilot

Sent by Copilot:

Your paragraph is well-written, but there is one comma splice. Here’s the corrected version:

Original: Cultural gaslighting, as defined by Ruíz (2020), is a strategic tool employed in settler colonial societies to legitimise and perpetuate colonial domination. Unlike interpersonal gaslighting, which manipulates an individual’s perception of reality, cultural gaslighting functions on a structural level. It creates environments—referred to as “abusive ambients”—designed to systematically undermine and erase the knowledge and experiences of marginalised communities(Ruíz, 2020). These environments encompass languages, narratives, and practices that entrench colonial authority while silencing Indigenous voices.

Corrected: Cultural gaslighting, as defined by Ruíz (2020), is a strategic tool employed in settler colonial societies to legitimise and perpetuate colonial domination. Unlike interpersonal gaslighting, which manipulates an individual’s perception of reality, cultural gaslighting functions on a structural level. It creates environments—referred to as “abusive ambients”—designed to systematically undermine and erase the knowledge and experiences of marginalised communities (Ruíz, 2020). These environments encompass languages, narratives, and practices that entrench colonial authority while silencing Indigenous voices.

The comma splice was between “communities” and “(Ruíz, 2020)”. Adding a space before the parenthesis resolves this issue.

Would you like any further assistance with this paragraph or anything else?

ChatGPT
<u>Note:</u> I initially tried to use (OpenAI ChatGPT, 2024b) for this task but it was not as accurate so I switched to Copilot. Here is the same example when I had tried using ChatGPT.

Please check this paragraph for comma splices: Cultural gaslighting, as defined by Ruíz (2020), is a strategic tool employed in settler colonial societies to legitimise and perpetuate colonial domination. Unlike interpersonal gaslighting, which manipulates an individual's perception of reality, cultural gaslighting functions on a structural level. It creates environments—referred to as "abusive ambients"—designed to systematically undermine and erase the knowledge and experiences of marginalised communities(Ruíz, 2020). These environments encompass languages, narratives, and practices that entrench colonial authority while silencing Indigenous voices.

ChatGPT said:

ChatGPT

The paragraph you provided does not contain any comma splices. The sentences are well-structured, and the commas are used correctly. Here's the paragraph again for reference:

Cultural gaslighting, as defined by Ruíz (2020), is a strategic tool employed in settler colonial societies to legitimize and perpetuate colonial domination. Unlike interpersonal gaslighting, which manipulates an individual's perception of reality, cultural gaslighting functions on a structural level. It creates environments—referred to as "abusive ambients"—designed to systematically undermine and erase the knowledge and experiences of marginalized communities (Ruíz, 2020). These environments encompass languages, narratives, and practices that entrench colonial authority while silencing Indigenous voices.

All the sentences are properly punctuated, and there are no comma splices.

<u>Note:</u> Uses of AI in fixing code were outlined in Appendix 1.

Appendix 3- Copy of Academic Integrity Declaration – CIM Dissertation

We're part of an academic community at Warwick. Whether studying, teaching, or researching, we're all taking part in an expert conversation which must meet standards of academic integrity. When we all meet these standards, we can take pride in our own academic achievements, as individuals and as an academic community.

Academic integrity means committing to honesty in academic work, giving credit where we've used others' ideas and being proud of our own achievements.

In submitting my work I confirm that:

1. *I have read the guidance on academic integrity provided in the Student Handbook and understand the University regulations in relation to Academic Integrity. I am aware of the potential consequences of Academic Misconduct.*
2. *I declare that the work is all my own, except where I have stated otherwise.*
3. *No substantial part(s) of the work submitted here has also been submitted by me in other credit bearing assessments courses of study (other than in certain cases of a resubmission of a piece of work), and I acknowledge that if this has been done this may lead to an appropriate sanction.*
4. *Where a generative Artificial Intelligence such as ChatGPT has been used I confirm I have abided by both the University guidance and specific requirements as set out in the Student Handbook and the Assessment brief. I have clearly acknowledged the use of any generative Artificial Intelligence in my submission, my reasoning for using it and which generative AI (or AIs) I have used. Except where indicated the work is otherwise entirely my own.*
5. *I understand that should this piece of work raise concerns requiring investigation in relation to any of points above, it is possible that other work I have submitted for assessment will be checked, even if marks (provisional or confirmed) have been published.*
6. *Where a proof-reader, paid or unpaid was used, I confirm that the proof-reader was made aware of and has complied with the University's proofreading policy.*
7. *I consent that my work may be submitted to Turnitin or other analytical technology. I understand the use of this service (or similar), along with other methods of maintaining the integrity of the academic process, will help the University uphold academic standards and assessment fairness.*

Privacy statement

The data on this form relates to your submission of coursework. The date and time of your submission, your identity, and the work you have submitted will be stored. We will only use this data to administer and record your coursework submission.

Word Count: 10,034

Abbreviations

AI: Artificial Intelligence

BSR: Business for Social Responsibility

CNN: Convolutional Neural Network

DOI: Dangerous Organisations and Individuals

GPO: Government Press Office

HMA: Hybrid Moderation Algorithm

HRW: Human Rights Watch

IR: International Relations

LLM: Large Language Model

MENA: Middle East and North Africa

PA: Palestinian Authority

PBC: Palestinian Broadcasting Corporation

RNN: Recurrent Neural Network

UN: United Nations

UNGP: United Nations Guiding Principles

UNSRCT: United Nations Special Rapporteur on Counter Terrorism

XLM-R: Cross-Lingual Language Model - RoBERTa

References

7or (2024) *Home - 7or*. 2024. The Palestinian Observatory of Digital Rights Violations (7or). https://7or.7amleh.org [Accessed: 20 August 2024].

Ak, G. (2024) “Crime of the Century”: Israel’s State Terrorism and International Law Violations in Gaza Strip. *Journal of Humanity, Peace and Justice*. 1 (1), 65–95. doi:10.26650/hupej.2024.10.1.1468793.

Albanese, F. (2022) Situation of Human Rights in the Palestinian Territories Occupied since 1967. *The Palestine Yearbook of International Law Online*. 18 (1), 291–297. doi:10.1163/22116141_018010012.

Alfonseca, K. (2023) *A look into the long history of the Israeli-Palestinian conflict*. 2023. ABC News. https://abcnews.go.com/International/timeline-long-history-israeli-palestinian-conflict/story?id=103875134 [Accessed: 1 July 2024].

Alimardani, M. & Elswah, M. (2021) *Digital Orientalism: #SaveSheikhJarrah and Arabic Content Moderation*. https://papers.ssrn.com/abstract=3900520.

Anon (2024) *Does Meta censor pro-Palestinian content? | Al Jazeera World Documentary*. 12 March. https://www.youtube.com/watch?v=YgLVMwOf834.

Aouragh, M. (2008) Everyday resistance on the internet: the Palestinian context. *Journal of Arab and Muslim Media Research*. 1 (2), 109–130. doi:10.1386/jammr.1.2.109/1.

Aouragh, M. (2011) Google-Books-ID: aGGJDwAAQBAJ. *Palestine Online: Transnationalism, the Internet and the Construction of Identity*. Bloomsbury Publishing.

Apodaca, T. & Uzcátegui-Liggett, N. (2024) *How We Investigated Shadowbanning on Instagram – The Markup*. 25 February 2024. https://themarkup.org/automated-censorship/2024/02/25/how-we-investigated-shadowbanning-on-instagram [Accessed: 20 August 2024].

J. Baylis, S. Smith, P. Owens, J. Baylis, S. Smith, & P. Owens (eds.) (2022) *The Globalization of World Politics: An Introduction to International Relations*. Ninth Edition, Ninth Edition. Oxford, New York, Oxford University Press.

Beer, D. (2017) The social power of algorithms. *Information, Communication & Society*. 20 (1), 1–13. doi:10.1080/1369118X.2016.1216147.

Bhambra, G.K. (2014) Postcolonial and decolonial dialogues. *Postcolonial Studies*. 17 (2), 115–121. doi:10.1080/13688790.2014.966414.

Bhattacharyya, S. (2022) *Global Debates in the Digital Humanities*. Minneapolis, UNITED STATES, University of Minnesota Press. http://ebookcentral.proquest.com/lib/warw/detail.action?docID=6921430.

Blunt, D., Wolf, A., Coombes, E. & Mullin, S. (2020) *Posting Into the Void: Studying the Impact of Shadowbanning on Sex Workers and Activists*. 2020. Hacking//Hustling. https://hackinghustling.org/posting-into-the-void-content-moderation/ [Accessed: 9 August 2024].

BSR (2022) *Human Rights Due Diligence of Meta's Impacts in Israel and Palestine* | *Reports* | *Sustainable Business Network and Consultancy* | *BSR*. https://www.bsr.org/en/reports/meta-human-rights-israel-palestine.

Burga, S. (2024) *Why the 'All Eyes on Rafah' AI Post Is Going Viral on Social Media*. 29 May 2024. TIME. https://time.com/6983344/all-eyes-on-rafah-ai-post-viral/ [Accessed: 13 August 2024].

Clegg, N. (2022) *Meta Launches New Content Moderation Tool as It Takes Chair of Counter-Terrorism NGO* | *Meta*. 2022. https://about.fb.com/news/2022/12/meta-launches-new-content-moderation-tool/ [Accessed: 28 July 2024].

Copilot (2024) *Microsoft Copilot: Your everyday AI companion*. 2024. Microsoft Copilot: Your everyday AI companion. https://ceto.westus2.binguxlivesite.net/ [Accessed: 20 August 2024].

Cotter, K. (2023) "Shadowbanning is not a thing": black box gaslighting and the power to independently know and credibly critique algorithms. *Information, Communication & Society*. 26 (6), 1226–1243. doi:10.1080/1369118X.2021.1994624.

Cruz, M. (2023) The Question of Palestine- Timeline of Events. *Question of Palestine*. https://www.un.org/unispal/timeline/.

Davies, A. (2024) *All Eyes on Rafah: The post shared by 47m people*. 30 May 2024. BBC News. https://www.bbc.com/news/articles/cjkkj5jejleo [Accessed: 13 August 2024].

R. Deibert, J. Palfrey, R. Rohozinski, & J. Zittrain (eds.) (2010) Accepted: 2019-01-17 23:55. *Access Controlled: The Shaping of Power, Rights, and Rule in Cyberspace*. The MIT Press. https://library.oapen.org/handle/20.500.12657/26076.

Dotson, K. (2011) Tracking Epistemic Violence, Tracking Practices of Silencing. *Hypatia*. 26 (2), 236–257. doi:10.1111/j.1527-2001.2011.01177.x.

Dwonch, A.S. (2019) *Palestinian youth activism in the internet age: Online and offline social networks after the Arab Spring.* London, Bloomsbury Publishing.

Eberhard, Simons, G.F. & Fennig, C.D. (2024) Arabic | Ethnologue Free. *Ethnologue: Languages of the World*. https://www.ethnologue.com/language/ara/.

Edenberg, E. & Hannon, M. (2021) Google-Books-ID: fcUqEAAAQBAJ. *Political Epistemology*. Oxford University Press.

Elicit (2024) *Evolution of Cultural Gaslighting and Censorship of Palestinians | Elicit*. 2024. https://elicit.com/notebook/9ce727e7-ebf2-4c54-8d02-c56d8be02a66 [Accessed: 20 August 2024].

Frankel-Shlosberg, L. (1996) The Palestinian news game: Walking on eggs in the Arafat era. *Columbia Journalism Review*.35 (1) p.16.

Fricker, M. (2007) Google-Books-ID: lncSDAAAQBAJ. *Epistemic Injustice: Power and the Ethics of Knowing*. Clarendon Press.

Goodfellow, I., Bengio, Y. & Courville, A. (2016) *Deep Learning*. Cambridge, MA, USA, MIT press. https://www.deeplearningbook.org/.

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

Hamla (2023) *Briefing on The Palestinian Digital Rights Situation Since October 7th, 2023*. 2023. https://7amleh.org/2023/11/01/briefing-on-the-palestinian-digital-rights-situation-since-october-7th-2023 [Accessed: 12 May 2024].

Hamla (2024) *Hashtag Palestine 2023: Palestinian Digital Rights During War*. 2024. https://7amleh.org/2024/01/17/hashtag-palestine-2023-palestinian-digital-rights-during-war [Accessed: 1 August 2024].

Haughton, C.-S. (2024) *All Eyes on Rafah: Why is the trend being criticised?* 29 May 2024. ITV News. https://www.itv.com/news/2024-05-29/all-eyes-on-rafa-whys-this-trend-being-criticised [Accessed: 13 August 2024].

Jamal, A. (2001) State-Building and Media Regime: Censoring the Emerging Public Sphere in Palestine. *Gazette (Leiden, Netherlands)*. 63 (2–3), 263–282. doi:10.1177/0016549201063002010.

Johnson, V., Nadal, K., Sissoko, G. & King, R. (2021) “It’s Not in Your Head”: Gaslighting, ‘Splaining, Victim Blaming, and Other Harmful Reactions to Microaggressions. *Perspectives on Psychological Science*. 16, 1024–1036. doi:10.1177/17456916211011963.

Keohane, R.O. (2005) Google-Books-ID: HnvpdocqT9EC. *After Hegemony: Cooperation and Discord in the World Political Economy*. Princeton University Press.

Kitchin, R. (2014) Big Data, New Epistemologies and Paradigm Shifts. *Big Data and Society*. 1 (1). doi:10.1177/2053951714528481.

Kuntz, B. (2021) Stolen Memories: Israeli State Repression and Appropriation of Palestinian Cultural Resources. *Journal of Radical Librarianship*. 7, 17–52.

Liddell, J.L. & Lilly, J.M. (2022) "There's so much they don't cover:" Limitations of healthcare coverage for Indigenous women in a non-federally recognized tribe. *SSM - Qualitative Research in Health*. 2, 100134. doi:10.1016/j.ssmqr.2022.100134.

Mason, V. & Falk, R. (2016) Assessing Nonviolence in the Palestinian Rights Struggle. *State Crime Journal*. 5 (1), 163–186. doi:10.13169/statecrime.5.1.0163.

McMahon, S.F. (2014) *The Boycott, Divestment, Sanctions campaign: contradictions and challenges*. 5 (4), 65–81.

Meta (2021) *Corporate Human Rights Policy*. 2021. https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf.

Meta (2023a) *Dangerous organisations and individuals | Transparency Centre*. 2023. https://transparency.meta.com/en-gb/policies/community-standards/dangerous-individuals-organizations/ [Accessed: 5 June 2024].

Meta (2023b) *Detecting violations | Transparency Centre*. 2023. https://transparency.meta.com/en-gb/enforcement/detecting-violations/ [Accessed: 5 June 2024].

Meta (2024a) *Expanding our open source large language models responsibly*. 2024. Meta AI. https://ai.meta.com/blog/meta-llama-3-1-ai-responsibility/ [Accessed: 28 July 2024].

Meta (2023c) *How technology detects violations | Transparency Centre*. 2023. https://transparency.meta.com/en-gb/enforcement/detecting-violations/technology-detects-violations/ [Accessed: 5 June 2024].

Meta (2024b) *Meta Office Locations*. 2024. Meta Careers. https://www.metacareers.com/locations [Accessed: 14 August 2024].

Meta (2024c) *Update from the Policy Forum on our approach to 'Zionist' as a proxy for hate speech | Transparency Centre*. https://transparency.meta.com/en-gb/hate-speech-update-july2024/.

Mignolo, W.D. (2011) *The Darker Side of Western Modernity: Global Futures, Decolonial Options*. Duke University Press. doi:10.2307/j.ctv125jqbw.

Miñoso, Y.E. & Espinosa, L.C. (2022) Ethnocentrism and Coloniality in Latin American Feminisms: The Complicity and Consolidation of Hegemonic Feminists in Transnational Spaces. *Hypatia*. 37 (3), 498–509. doi:10.1017/hyp.2022.32.

Narea, N. (2023) *A timeline of Israel and Palestine's complicated history*. 19 October 2023. Vox. https://www.vox.com/world-politics/23921529/israel-palestine-timeline-gaza-hamas-war-conflict [Accessed: 28 June 2024].

Ní Aoláin, F. (2021) *Input of the United Nations Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism to the Facebook Oversight Board Concerning its 'Community Guidelines' and 'Community Standard on Dangerous Individuals and Organizations'*. 2021, https://www.ohchr.org/sites/default/files/Documents/Issues/Terrorism/UNSRCT_Facebook_Oversight_Board_Input2021.docx.

OpenAI ChatGPT (2024a) *ChatGPT*. 2024. Help with Map. https://chatgpt.com/c/56e18f06-f98b-4721-8018-f55f2f7cf682 [Accessed: 21 August 2024].

OpenAI ChatGPT (2024b) *ChatGPT*. 2024. Request to Proofread. https://chatgpt.com/ [Accessed: 20 August 2024].

Peck, J. (2011) *Ideal illusions: How the US government co-opted human rights*. 1st edition. Henry Holt and Company.

Pieterse, J. (1984) State Terrorism on a Global Scale: The Role of Israel. *Crime and Social Justice*. (21/22), 58–80.

Primoratz, I. (2005) State terrorism and counterterrorism. In: *Ethics of Terrorism & Counter-Terrorism*. pp. 69–82.

Project Pythia Community (2024) *Introduction to Cartopy — Pythia Foundations*. 2024. https://foundations.projectpythia.org/core/cartopy/cartopy.html [Accessed: 20 August 2024].

Register, Y., Grasso, I., Weingarten, L.N., Fury, L., Chinea, C.E., Malloy, T.J. & Spiro, E.S. (2024) Beyond Initial Removal: Lasting Impacts of Discriminatory Content Moderation to Marginalized Creators on Instagram. *Proceedings of the ACM on Human-Computer Interaction*. 8 (CSCW1), 1–28. doi:10.1145/3637300.

Ricaurte, P. (2022) Ethics for the majority world: AI and the question of violence at scale. *Media, Culture & Society*. 44, 016344372210996. doi:10.1177/01634437221099612.

Rigby, A. (2010) Google-Books-ID: vnFhywAACAAJ. *Palestinian Resistance and NONVIOLENCE.* Passia.

Ruíz, E. (2020) Cultural Gaslighting. *Hypatia*. 35 (4), 687–713. doi:10.1017/hyp.2020.33.

Said, E.W. (1994) *Culture and imperialism*. 1st Vintage Books ed. New York, Vintage Books.

Said, E.W. (2000) Invention, Memory, and Place. *Critical Inquiry*. 26 (2), 175–192.

Said, E.W. (1977) Orientalism. *The Georgia Review*. 31 (1), 162–206.

Saltman, E. & Hunt, M. (2023) *Borderline Content: Understanding the Gray Zone*.

Shahid, F. & Vashistha, A. (2023) Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony? *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18. doi:10.1145/3544548.3581538.

Siapera, E. (2022) AI Content Moderation, Racism and (de)Coloniality. *International Journal of Bullying Prevention*. 4 (1), 55–65. doi:10.1007/s42380-021-00105-7.

Singh, J. (2024) Meta to end ban on the word ‘shaheed’ on oversight board’s recommendations. *Reuters*. 2 July. https://www.reuters.com/technology/meta-end-ban-word-shaheed-oversight-boards-recommendations-2024-07-02/.

The Matplotlib Development Team (2024) *Bar of pie — Matplotlib 3.9.1 documentation*. 2024. https://matplotlib.org/stable/gallery/pie_and_polar_charts/bar_of_pie.html#sphx-glr-gallery-pie-and-polar-charts-bar-of-pie-py [Accessed: 1 August 2024].

Tuck, E. & Yang, K.W. (2012) *Decolonization is not a metaphor*. 1 (1), 1–40.

Véliz, C. (2023) *Oxford Handbook of Digital Ethics*. Oxford Handbooks. OUP Oxford. https://books.google.co.uk/books?id=g4jlEAAAQBAJ.

Waltz, K. (1979) *Theory of International Politics*. Boston, Mass., McGraw-Hill Education.

Westfall, S., Murphy, B., Taylor, A. & Pietsch, B. (2023) *The Israeli-Palestinian conflict: A chronology*. 9 October 2023. Washington Post. https://www.washingtonpost.com/world/2023/israel-palestine-conflict-timeline-history-explained/ [Accessed: 1 July 2024].

Whitelam, K.W. (2018) The poetics of the history of Israel: Shaping Palestinian history*. In: *Revealing the History of Ancient Palestine*. Routledge. p.

Younes, R. (2023) Meta’s Broken Promises. *Human Rights Watch*. https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and.

Zuckerberg, M. (2017) *Facebook Post*. 27 September 2017. Facebook. https://www.facebook.com/zuck/posts/i-want-to-respond-to-president-trumps-tweet-this-morning-claiming-facebook-has-a/10104067130714241/ [Accessed: 18 August 2024].