Abstract
Social media platforms, while influential tools for human rights activism, free speech, and mobilization, also bear the influence of corporate ownership and commercial interests. This dual character can lead to clashing interests in the operations of these platforms. This study centers on the May 2021 Sheikh Jarrah events in East Jerusalem, a focal point in the Israeli-Palestinian conflict that garnered global attention. During this period, Palestinian activists and their allies observed and encountered a notable increase in automated content moderation actions, like shadow banning and content removal. We surveyed 201 users who faced content moderation and conducted 12 interviews with political influencers to assess the impact of these practices on activism. Our analysis centers on automated content moderation and transparency, investigating how users and activists perceive the content moderation systems employed by social media platforms, and their opacity. Findings reveal perceived censorship by pro-Palestinian activists due to opaque and obfuscated technological mechanisms of content demotion, complicating harm substantiation and lack of redress mechanisms. We view this difficulty as part of algorithmic harms, in the realm of automated content moderation. This dynamic has far-reaching implications for activism’s future and it raises questions about power centralization in digital spaces.
Opaque algorithms, transparent biases: Automated content moderation during the Sheikh Jarrah Crisis

“I think people are absolutely not happy. They don’t trust platforms. Almost everyone has a complaint, either because they’ve been censored, their account has been blocked, or they have been attacked, or subjected to hate speech or disinformation. If users found another platform with the same scale and spread as Twitter or Facebook, they would leave these platforms immediately. It’s only because there are no others. This is where everyone is, right? And that’s why you’re forced to stay there.”
— Participant 1, a Palestinian political activist and influencer.

Social media platforms like Instagram, Facebook, and X serve as crucial spaces for political discourse and activism. Their widespread use, global reach, and user-friendly interfaces not only make them not only platforms for news consumption (Kwak, et al., 2010; Flintham, et al., 2018; Wang and Mark, 2017) but also powerful tools for community organizing and collective action. This is particularly true in regions where freedom of expression is constrained by state power. A prominent example is their organizing role in the Arab Spring of 2011, where the massive mobilization brought down long-standing regimes and fundamentally changed the political landscape of the Middle East (Simpson, 2018; Saeed, et al., 2011; O’Callaghan, et al., 2014; Lotan, et al., 2011; Kavanaugh, et al., 2016; Al-Ani, et al., 2012; Sen, et al., 2010).

The design and affordances of social media platforms present themselves as spaces for users to express themselves and their creativity and reach a broad audience (Theocharis, et al., 2023; Ahuja, et al., 2018). These affordances encourage the viral spread and creative crafting of political content and facilitate the formation of “affective publics” — networked communities mobilized through expressions of sentiment that manifest as online solidarity or offline action (Papacharissi, 2016). However with regards to movements challenging the status quo, like Black Lives Matter and #MeToo, and in global conflicts such as the Russian invasion of Ukraine, these platforms are a two sided coin. On the one hand, they are powerful tools for both those seeking to enhance freedom and those seeking to curtail it, such as governments. On the other hand, their private ownership and opacity raise crucial questions about user influence, design choices, policies, and specifically, content moderation (Myers West, 2018).

Our study focuses on the automated content moderation policies enacted during the May 2021 events in Sheikh Jarrah (SJ), a historically disputed neighborhood in East Jerusalem, occupied by Israel. The tensions escalated when Palestinian families were expelled from their houses and detained, and Israel changed the agreed upon status quo in Al Aqsa mosque during Ramadan, leading to a deadly conflict in the Gaza Strip. In response, activists from around the world, including prominent figures (Denman, 2022; Yee and El-Naggar, 2021), turned to social media to shed light on life in East Jerusalem under Israeli occupation. During these events, activists used hashtags like #SaveSheikhJarrah to document and denounce various forms of violence and injustice. However, their posts faced various forms of restrictions in the name of content moderation, including shadow banning, sensitivity filters, and even account restrictions.

Drawing inspiration from Myers West (2018), our approach centers on the lived experiences of activists and their allies, examining content moderation decisions within the unique context of SJ. Specifically, we explore three aspects of users’ experiences with content moderation: their perceptions of how it works, the impact on the visibility and reach of pro-Palestine voices, and the actions they take to circumvent it. Our primary objective is to address the gap between user perceptions and the actual enforcement of content moderation policies. We aim to identify users’ needs for ‘meaningful transparency’ (Suzor, et al., 2019), focusing on policies that would provide users and researchers with more detailed and meaningful explanations of the moderation process, and the ability to appeal decisions, thereby addressing the platform’s shortcomings.

Given the critical role that social media platforms play in conflict zones and the opaque nature of social media policies and actions, which often leads to perceptions of algorithmic bias and potentially unjust
Opaque algorithms, transparent biases: Automated content moderation during the Sheikh Jarrah Crisis

censorship, our study is timely. It is particularly crucial for Palestinian activists and their allies, who not only heavily rely on these platforms for awareness and mobilization but also face limited alternative spaces to voice their concerns and gather support. Our study employs a mixed-methods approach, combining survey data from 201 social media users who experienced content moderation and in-depth interviews with 12 activist influencers. This allows us to examine both the breadth and depth of lived experiences with content moderation, how users interpret the responsibilities of social media platforms in upholding free expression, and whether they see it as intentional policy decisions or design flaws.

**Motivation**

Our motivation is rooted in addressing the notable under-representation of the Arab World within Western academic discourse. In parallel, we champion the development and implementation of less biased content moderation practices, while genuinely considering the diverse perspectives of various stakeholders. The primary objective of our research is to contribute meaningfully to the ongoing discourse around fostering equitable and more transparent digital spaces for the benefit of activists and users globally. Furthermore, our research seeks to confront the perils and harms that accompany the expansion of algorithmic moderation, especially as it pertains to a protected group, such as the Palestinians. We ask for a critical re-evaluation of the prevailing discourse surrounding “the use of automated decision making in moderation, and not let firms hide behind the veil of black-boxed complexity” as platforms currently seek to disengage from content policy discussions, and the more automated tools become ubiquitous (Gorwa, *et al.*, 2020)

**Contribution**

1. We extend the discourse on platform governance, particularly focusing on the power imbalances in the development and deployment of AI systems for content moderation that could potentially stifle activist voices.
2. We provide a nuanced understanding of user experiences faced by activists in Israel-Palestine during the SJ crisis, thereby filling a crucial knowledge gap that can inform the development of more equitable automation standards for content moderation, especially with regards to conflict areas.

---

### 2. Contextual background: The Sheikh Jarrah and Silwan Neighborhoods Incident

The land disputes in Sheikh Jarrah (SJ) are intricate and steeped in a historical context that dates back to the nineteenth century. While a comprehensive review is beyond the purview of this study, we provide a summary of contextually relevant information. In 1967, Israel captured SJ from Jordan during the Six-Day War and subsequently annexed it — a move in violation of international law. SJ is part of East Jerusalem, a conglomeration of neighborhoods currently housing nearly 600,000 inhabitants, of whom over 60 percent are Palestinian. SJ is located in close proximity to the Old City of Jerusalem, which houses the Al-Aqsa Mosque, one of Islam’s holiest sites, along with monuments key to Christianity and Judaism.

Palestinian families, many of whom have been residing in SJ since the nineteenth century, find themselves embroiled in ongoing property disputes. These disputes often involve far-right Israeli groups claiming ownership of certain houses, some of which are now occupied by Jewish-Israeli settlers. The neighborhood has thus become an epicenter of political activism, tension, and conflict. Palestinian inhabitants predominantly hold the status of ‘permanent’ residents, a classification that is itself a subject of contention. While some aspire to full Israeli citizenship, others perceive Israel as an illegitimate occupying force.

Socioeconomically, SJ is one of Jerusalem’s most impoverished areas, characterized by high youth unemployment rates and neglected public infrastructure. This volatile mix of political, social, and economic factors perpetuates unrest among both Israeli and Palestinian communities residing in the area.
In May 2021, a resurgence of tensions over property ownership and the status quo regarding Al Aqsa mosque catalyzed significant protests (Bateman, 2021). Young activists utilized social media platforms, such as Facebook, Twitter, YouTube, Instagram, and TikTok, to disseminate images and narratives reflecting the situation in Sheikh Jarrah. Hashtags like #SaveSheikh-Jarrah and #FreePalestine became viral rallying cries. Their efforts garnered unprecedented global attention (Yee and El-Naggar, 2021), galvanizing the Palestinian diaspora and rejuvenating the protest movement internationally. In the United States, this led to mounting pressure on government officials.

Subsequently, however, platforms like Facebook, Instagram, and X placed restrictions on content — ranging from content notices to shadow banning — particularly on hashtags such as #AlAqsaMosque and the word “Palestine” in both English and Arabic (source: Human Rights Watch Report). In contrast, a recent study focused on the role of TikTok in the context of SJ by analyzing three major hashtags: #SaveSheikhJarrah, #SavePalestine, and #FreePalestine. These hashtags collectively garnered billions of views and revealed that TikTok’s platform affordances allowed for a nuanced and multilayered presentation of pro-Palestinian narratives. Unlike other platforms, TikTok enables users to blend direct political messages with challenge-based content, making the activism more dynamic and engaging (Abbas, et al., 2022). This differential treatment across platforms raises critical questions about the role of social media corporations in shaping public discourse and influencing political activism. It also invites further inquiry into the algorithms and moderation policies that govern these platforms, and how they may be contributing to, or inhibiting, social movements and political change.

3. Related work

3.1. Social media and political activism in the MENA region

The advent of social media has changed various aspects of life in the MENA region, but its role in political activism is particularly noteworthy. This was especially evident during the Arab Spring of 2011–2012, which marked one of the first major political events in which social media platforms played a pivotal role. This was not only significant for the activists and citizens involved but also for the platforms themselves, as they navigated uncharted territory in governance and content moderation.

Platforms like Facebook, Twitter, and YouTube became instrumental in mobilizing citizens, disseminating grievances, and attracting global attention (Kavanaugh, et al., 2011; Lotan, et al., 2011; Wulf, et al., 2013a, 2013b). These platforms acted as megaphones for activists, often working in tandem with other media like satellite TV to apply pressure on autocratic regimes and deter violent suppression of protests (Al-Ani, et al., 2012; el Nawawy and Khamis, 2013; Kavanaugh, et al., 2011; Starbird and Palen, 2012; Aday, et al., 2012).

The strategies for leveraging social media in political activism varied even between fairly similar contexts within the MENA region (where the Arab Spring occurred), as evidenced by the well documented different approaches in Tunisia, Egypt, and Syria (Wulf, et al., 2013a, 2013b; Rohde, et al., 2016; Aal, et al., 2018). In Palestine, Internet activism has evolved from information disseminated about the humanitarian and political situation to online political organizing and critical discourse (Nabulsi, 2014). This shift has redistributed the power dynamics of how #Palestine is represented, moving from ‘fact-based, hard news’ produced by mainstream media to decentralized networks of news ‘produsers’ — individuals who are both producers and users, and citizen journalists, reporting real-time events as witnesses (Siapera, 2014). For Palestinians and Israelis in an ongoing, high-stakes conflict, these platforms are a vital lifeline for real-time communication. Any disruption to this channel could have serious implications, leaving communities without crucial information and highlighting the critical role of platform governance in such volatile situations.
During the Arab Spring, social media platforms were heralded as instruments of “liberation” and free expression. Globally, social media platforms are often seen as an infrastructure for public discourse, or the current ‘town square,’ where public opinion is moulded and reflected (Plantin and Punathambekar, 2019; Roberts, 2016; Gillespie, 2018). For instance, during Egypt’s six-day Internet shutdown in 2011, platforms like Twitter and Facebook provided tools like Speak To Tweet, earning public and activist praise. However, this level of support for activism hasn’t been consistent and has evolved over time, as highlighted by the contrasting case of SJ, which our study delves into. By focusing on the MENA region, an area often overlooked in content moderation discussions, our research aims to shed light on the complex interplay of voice and silence on these platforms. We explore users’ perceptions of governance choices that dictate how various causes are treated — whether they’re spotlighted, moderated, or censored.

3.2. Platform governance and content moderation

Public debates on the boundaries of acceptable content on social media have intensified in recent years. While there’s consensus on removing illegal content such as child pornography, the lines between free speech and objectionable harmful speech are shifting, and with them, content moderation systems — both algorithmic and human — are changing, and unstable objects (Rieder and Skop, 2021). In this digital landscape, content moderation has evolved to a central commercial necessity and a commodity, often operating behind the scenes shaping user experience (Roberts, 2016; Gillespie, 2018). For our study, we adopt Grimmelmann’s (2015) definition of moderation as “the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse” [1]. Gillespie offers a more nuanced definition relevant to online spaces: a set of practices used by platforms to manage, evaluate, and sometimes remove user-generated content based on platform policies and societal norms. We further frame our research within “online commercial content moderation,” a term introduced by Sarah T. Roberts (Roberts, 2016). She emphasizes the role of human judgment in moderation decisions, even as algorithms and machine learning tools are increasingly used. However, the lack of transparency around these automated tools, due to both technical opacity and commercial secrecy, and the lack of clarity from companies how they figure into or replace human decision-making, raises concerns about their effectiveness and fairness.

In this paper, we also delve into the concept of censorship, a term often avoided in academic discourse in favor of less charged phrases like “content moderation” or “information control.” This avoidance largely stems from the challenges in proving censorship claims, given the opaque policies governing these practices (Gillespie, 2018). While fair application of content moderation aims to balance the interests of various stakeholders and maintain a ‘clean’ digital space, censorship is often understood as a more extreme, unilateral form of control, often linked to suppressing speech for ideological, political, or commercial reasons (Gillespie, 2018; Tufekci, 2017). Identifying the line between responsible curation and suppression, potentially based on opaque or unfair criteria, is challenging (Lavin, 2020; Gillespie, 2018). Platforms, while hosting user content, also cater to advertisers as a major revenue source (Rieder and Sire, 2014). This commercial aspect significantly influences content moderation, as seen in a recent case where advertisers withdrew from X following Elon Musk’s endorsement of an antisemitic post, risking a loss of up to $US75 million in revenue. This highlights how financial outcomes can drive censorship or content demotion, underscoring the conflict between commercial viability and free speech. Such scenarios are vital in understanding the digital expression landscape, especially its effects on marginalized groups, and the integrity of public discourse. The line between moderation and censorship can blur, especially with sensitive political topics. As social media continues to evolve, the need for academic research on content moderation and censorship grows, particularly in how they intersect with democracy, human rights, and social justice.

Moderation practices range from visible actions like account deletions, content warnings or deplatforming, to subtler tactics such as shadow banning, which involves content demotion and is often impossible for users to notice or prove. These less noticeable practices pose significant legal and policy concerns due to their obfuscation and lack of a clear counterfactual, leaving users unable to contest them effectively (Leerssen, 2023; Gillespie, 2022). According to Gillespie these techniques of “reduction of visibility,” are
becoming more prevalent in large platforms’ content moderation regimes. He warns that these methods moderate “not only what any one user is likely to see but also what society is likely to attend to, take seriously, struggle with, and value” [2]. Platforms like Meta assert that their content moderation policies aim to create a safe and credible environment for users (Facebook, n.d.). Some platforms release reports on algorithmic moderation, for example, Meta releases a quarterly ‘community standards enforcement report’. This report details in quite broad terms what percentage of content violations were identified by algorithms. However, Meta’s reports lack transparency about the training and implementation of these algorithms. For instance, a persisting technical issue in computer science research is the poor abilities of NLP technologies with under resourced languages like Arabic and Hebrew [3]. One can assume that Meta might run into these issues with these languages, dominant in the region we are focusing on.

An investigation by ProPublica into Facebook’s content moderation practices reveals inconsistencies and controversial decision-making processes. They expose how Facebook’s internal guidelines can lead to unequal treatment of posts, sometimes favoring elites and governments over grassroots activists and racial minorities. For example, a U.S. congressman’s call for violence against ‘radicalized’ Muslims remained on the platform, while a post by a Black Lives Matter activist was removed. The report suggests that Facebook’s approach to content moderation is not only complex but also fraught with challenges in balancing free speech, public safety, and commercial interests (Angwin and Grassegger, 2017). Studies on YouTube and Twitter also highlight the complexities and challenges of content moderation (Ma and Kou, 2022; Jhaver, et al., 2021). Many organizations and scholars have raised concerns over the topic, fueling public debate (Thach, et al., 2022; Common, 2020; Vaccaro, et al., 2020). Legal experts have criticized social media platforms for making unpredictable and non-transparent decisions, arguing that this leads to a new type of corporate censorship [4].

In this context, activists often produce or share real-time content for wider dissemination, forcing platforms to weigh their commercial benefits against risks. This dynamic raises questions about free speech, as platforms must decide whether to amplify or restrict politically charged content, sometimes even violent content from conflict zones influencing the success or failure of activist movements. In summary, the literature paints a complicated picture of content moderation practices that intersect with issues of free speech, transparency, commercial influence, and power dynamics. As we transition into the methods section, we will outline the research design and tools we employed to delve deeper into these intricacies, particularly focusing on the experiences of activists and the role of automated systems in content moderation.

Methods

Our mixed-methods study combined surveys and semi-structured interviews to investigate the content moderation experiences of social media users and activists advocating for Palestinian rights, particularly focusing on events in May 2021 in Sheikh Jarrah. The survey targeted social media users who posted about the events and encountered content moderation in any form on their posts or interactions with posts from others. It was distributed to a sample of 230 respondents following recruitment calls on Twitter, Facebook, and Instagram in Arabic, English, and German through the team’s social networks. A total of 201 responses were considered for analysis. For a deeper understanding, we also conducted 12 interviews with individual activist influencers — with over 10k followers — who had encountered content moderation. The study was reviewed and approved by the IRB at Indiana University.

4.1. Survey and interview protocols

Adopting a social constructivist approach, our study aimed to capture the subjective experiences of participants, including activists and social media users. This approach was grounded in theories from Vygotskii (1979) as well as Berger and Luckmann (2023). This was complemented by a case study
framework, in line with methodologies outlined by Yin (2009), Stake (1995), and Merriam and Tisdell (2015), to delve into the specific socio-political events and context of Sheikh Jarrah in May 2021.

The survey, conducted in English and Arabic and hosted on Qualtrics, comprised 29 questions, both closed and open-ended. Participants had to be at least 18 years old, familiar with the events in Sheikh Jarrah, and have experienced content moderation related to posts about Palestine during the event or shortly thereafter. Eligible respondents provided descriptive accounts of their experiences, their perceived reasons for moderation, their counter-strategies, and concluded with demographic information.

Additionally, we invited influencers for online interviews, briefing them on protocols and obtaining informed consent for recording. Using semi-structured and ethnographically-influenced interviews, we sought deeper insights into experiences with content moderation. Participants’ languages varied between those who conversed in English, Arabic, or German. During the interviews, participants discussed their backgrounds, posting activities since May 2021, platform preferences, and views on the importance of social media for activists. The interview process was carefully structured to encourage open-ended responses, ensuring participant confidentiality and comfort. Sessions typically lasted around 45 minutes and were conducted with two researchers present: one to conduct the interview and another to take notes. When a note-taker was unavailable, interviewers employed a reflective approach post-interview to document key takeaways. The team exchanged insights and learnings via e-mail and met twice a month via video conference.

4.2. Recruitment and data collection

We aimed for diverse voices by using social media platforms like Facebook, X, and Instagram, known for their high regional use and noted content moderation issues that we observed. Besides these platforms, we engaged online communities, forums, and popular hashtags like #sheikh Jarrah and #Palestine. Acknowledging our sample’s limitations, we recognize that focusing on these platforms and only English and Arabic might have excluded certain users. Our recruitment, from December 2021 to November 2022, included retweets from influencers, journalists interested in the topic, and sharing in WhatsApp groups enabling us to engage with a broader demographic beyond our immediate research network. For interviews, we used snowball sampling to find influencers with large followings affected by content restrictions. Our linguistically diverse team enabled engagement with participants from various geographies, including Israelis supporting the Palestinian cause. However, we accept that the sample does not cover all major platforms, leaning more towards Instagram, Facebook, and X, excluding TikTok and YouTube.
Table 1: Cross tabulation of survey participants’ demographics

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Woman</th>
<th>Man</th>
<th>Not specified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>88</td>
<td>71</td>
<td>42</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-24</td>
<td>52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25-34</td>
<td>63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35-44</td>
<td>26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>45-64</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not specified</td>
<td>57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bachelor’s degree</td>
<td>68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doctorate degree</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Master’s degree</td>
<td>38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some college/university</td>
<td>34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>some secondary/high school</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not specified</td>
<td>37</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.3. Participants

A total of 230 participants were recruited for the survey. After excluding those who did not engage with the events in Palestine ($N=11$) or did not report experiencing content moderation ($N=29$), 201 responses were included in the final dataset. Participants hailed from diverse nations, including Canada, Oman, Kuwait, and the U.S., and responded in either English or Arabic. The detailed demographics are depicted in Table 1.

Twelve influencers (M=7, F=5), who are all pro-Palestinian activists with at least 10,000 followers on one platform, were voluntarily recruited for interviews without incentives. To maintain confidentiality, full demographic details are not provided, but some are available in Table 2. Recordings were pseudonymized and securely stored.

Table 2: Demographic details of the interview participants

<table>
<thead>
<tr>
<th>P</th>
<th>Gender</th>
<th>Education</th>
<th>Profession</th>
<th>Location of residence</th>
<th>Country of birth</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>F</td>
<td>Masters degree</td>
<td>Director of an NGO</td>
<td>Palestine</td>
<td>Palestine</td>
<td>English</td>
</tr>
<tr>
<td>P2</td>
<td>M</td>
<td>Bachelor’s degree</td>
<td>Engineer</td>
<td>Palestine</td>
<td>Palestine</td>
<td>Arabic</td>
</tr>
<tr>
<td>P3</td>
<td>F</td>
<td>Masters degree</td>
<td>Company owner</td>
<td>Netherlands</td>
<td>Palestine</td>
<td>English</td>
</tr>
<tr>
<td>P4</td>
<td>M</td>
<td>Bachelor’s degree</td>
<td>Student</td>
<td>Germany</td>
<td>Germany</td>
<td>German</td>
</tr>
<tr>
<td>P5</td>
<td>M</td>
<td>Masters degree</td>
<td>Engineer</td>
<td>Palestine</td>
<td>Palestine</td>
<td>Arabic</td>
</tr>
<tr>
<td>P6</td>
<td>F</td>
<td>Masters degree</td>
<td>Country manager in an NGO</td>
<td>Sweden</td>
<td>Palestine</td>
<td>English</td>
</tr>
<tr>
<td>P7</td>
<td>F</td>
<td>Masters degree</td>
<td>Journalist</td>
<td>Palestine</td>
<td>Palestine</td>
<td>Arabic</td>
</tr>
<tr>
<td>P8</td>
<td>M</td>
<td>Masters degree</td>
<td>Director of an NGO and Journalist</td>
<td>UK</td>
<td>Palestine</td>
<td>English</td>
</tr>
<tr>
<td>P9</td>
<td>M</td>
<td>Masters degree</td>
<td>Director of an NGO</td>
<td>Israel</td>
<td>Israel</td>
<td>English</td>
</tr>
<tr>
<td>P10</td>
<td>M</td>
<td>Bachelor’s degree</td>
<td>Social media manager at an NGO</td>
<td>Israel</td>
<td>Israel</td>
<td>English</td>
</tr>
<tr>
<td>P11</td>
<td>M</td>
<td>Bachelor’s degree</td>
<td>Journalist</td>
<td>Israel</td>
<td>USA</td>
<td>English</td>
</tr>
<tr>
<td>P12</td>
<td>F</td>
<td>Masters degree</td>
<td>Policy manager at an NGO</td>
<td>Germany</td>
<td>Palestine</td>
<td>English</td>
</tr>
</tbody>
</table>

4.4. Data analysis
Opaque algorithms, transparent biases: Automated content moderation during the Sheikh Jarrah Crisis

Five researchers applied thematic coding to the transcripts, adopting an inductive approach to focus on the actual experiences of activists with content moderation (Braun and Clarke, 2006; Guest, et al., 2011). The team convened multiple times to discuss and refine initial codes. This method entailed identifying recurrent themes through meticulous line-by-line examination before interpretation, leveraging thematic analysis’ suitability for phenomenological research and concentrating on subjective human experiences. The inductive approach allowed us to conceptualize shared experiences with content moderation and highlighted discrepancies between user perceptions and explanations provided by social media platforms. Our data collection yielded qualitative data and descriptive statistics, offering a narrative that encapsulates the lived, direct experiences of pro-Palestinian social media users who have encountered content restriction or deletion during a time of violent escalation in the conflict.

4.5. Positionality

Our team’s diverse expertise span HCI, informatics, journalism, media, and communication, combined with our diverse backgrounds from Israel, Saudi Arabia, Morocco, Germany, and of Palestinian descent, along with various religious beliefs, enriches our analysis and helps mitigate potential biases. This aligns with feminist epistemology, particularly focused on not obfuscating the situatedness of the knowledge that each researcher’s unique perspective brings to the research process (Haraway, 1988). We have observed worrying practices of content moderation related to posts about Sheikh Jarrah, especially those in non-English languages. Driven by these observations, we scrutinized and highlighted these seemingly biased and restrictive online practices. Our multifaceted political and linguistic proficiency enables a nuanced exploration from diverse situated perspectives, offering insights into the disparate moderation regimes encountered by political entities on social media. The team engaged in extensive discussions over several days to reach a consensus on themes, adding rigor and depth to our analytical process. This approach has led to certain challenges in the process. However, we see them as productive challenges that strengthen our conclusions, as we had to examine different perspectives to reach them. We address some of the debates we had throughout the discussion section below. Despite differing interpretations of historical details or the perspectives on the future resolution of the conflict, we all recognize the human rights violations in Palestine-Israel and Israel’s breaches of international law.

5. Findings

Through inductive analysis, we explore the lived experiences of activists and allies in relation to content moderation during the May 2021 events. Our study focuses on users’ initial encounters with moderation, the types of content that triggered it, their perceived reasons for moderation, and the actions they took in response. Our survey data reveals a notable pattern: a surge in content moderation or restriction, predominantly targeting Palestinian content in Arabic during the SJ events. The platforms most implicated in these actions were Instagram (42.86 percent), Facebook (25.27 percent), X (17.58 percent), and YouTube (6.96 percent), aligning with the platforms most frequently used by our participants. Another theme was the clear discrepancy between the perceptions of users and the justifications of moderation provided by the platforms, raising issues of opacity, power dynamics and lack of redress mechanisms. In the following we unpack these themes and discuss them in details.

5.1. The surge: Experiences of content moderation

Both survey and interview data indicate widespread encounters with content moderation. Among survey respondents, 12.93 percent experienced moderation on their own content, 33.17 percent on both their own and others’ content, and 53.73 percent witnessed it on content by others. These moderation actions span a spectrum of visibility, from overt methods like content restriction, often accompanied by notices to users, to covert tactics like shadow banning that users did not receive clear notice of. In our study, the most
frequently reported types of moderation were content restriction (29.6 percent), shadow banning (17.3 percent), content deletion (17.3 percent), and account restriction (15.3 percent).

Interviews with activists and influencers further substantiated these findings. Starting with content restriction and deletion, which were among the most frequently reported issues in the surveys. P2 recounted that two of his stories were deleted, one depicting settlers beating an Arab, and another featured settlers insulting the Prophet Mohammad. He said, “I remember those stories disappeared, and I also received warnings about them.” Similarly, P5 discussed a video he re-shared that depicted the killing of Al-Jazeera journalist Shireen Abu Akleh. He stated, “The video displays how this journalist was killed, and it was subsequently restricted.” Participants also reported more severe, visible actions like account restrictions, temporary posting bans, or complete deplatforming. These actions often followed instances where their content had been flagged by other users as inappropriate. According to the Electronic Frontier Foundation (EFF), such flagging initiates a review process, which can be either manual, involving human moderators, or automated. The user responsible for the content usually receives a notification about this action. Depending on the severity or frequency of the violations, punitive measures like temporary suspension or indefinite account banning may follow. P8, a journalist and NGO director, noted that his NGO’s Twitter page had been restricted. He described: “When someone tries to follow our NGO on Twitter, they’re warned that we have sensitive content. However, if you check our Twitter feed, you’ll see that we only post text-based stories; we don’t share graphic videos or pictures. Despite this, our content seems to have limited reach.” Upon verification, we found P8’s claim to be accurate.

Among our interviewees, shadow banning was a common experience, yet its causes and intentions remain unclear to them. Shadow banning is an elusive practice where users’ posts are hidden from everyone except the poster themselves, or their visibility is limited (Savolainen, 2022). This form of moderation can manifest in various ways, including content demotion, making the affected user’s posts invisible to the wider audience without their knowledge. Users subjected to shadow banning “continue posting to a fictional audience instead of creating a new account” (Savolainen, 2022). For instance, P3 noticed a decline in Instagram story views: “I observed a decline in the views of my Instagram stories compared to my other posts. I’m pretty sure of it, and since then, the views have never returned to their previous levels.” P8, saw reduced reach despite gaining followers, indicating that their content was reaching fewer people. He reported, “Even though we’ve gained more followers during the event, our posts don’t seem to reach everyone. It’s ironic because we started in 2015 with only a couple of thousand followers. Back then, whenever we posted, it would reach thousands of people. Now, for example, on Facebook, we have 75,000 followers. But when we post our powerful stories, only a few hundred people seem to see them.”

The opaqueness in content moderation, particularly of shadow banning, highlights the complexities and challenges in understanding these practices. As Savolainen warned, “shadow banning” may misleadingly imply deliberate, unified censorship, but it often stems from varied, platform-specific algorithms and policies [5]. It is difficult to discern the true nature of such practices, which might be incorrectly perceived as intentional suppression.

Our findings illustrate a complex landscape of content moderation, notably impacting posts about Palestine and often differing by language and political perspective. For instance, P11 found it ‘ironic’ that Arabic posts about Israeli police actions faced less suppression, contrasting with other diminished-view posts. In contrast, our Israeli participants reported no moderation on Hebrew-language posts but observed shadow bans on content posted by Palestinians. For example, P9 shared an instance of a post that was reported but not removed: “there was one instance one image with the dead teenager, child, that I posted with Yair Lapid column where he said that ‘we are ashamed to the bottom of our soul, we kill children. That’s what we become‘ ... I got some notification that was reported ... I don’t think it was taken down.” Disparities in moderation across different content types, including images (P6, P9), videos (P5), stories (P2) live-streamed videos (P2) and plain text (P9), further highlight these issues.

This reflects a broader theme of intransparency and induced uncertainty among users, complicated by content policies and platform designs. Leerssen’s observation that shadow banning does not have an
Opaque algorithms, transparent biases: Automated content moderation during the Sheikh Jarrah Crisis

This lack of clarity and evidence of precise causes, as highlighted in our study, is indicative of the veiled operation of power. The platforms’ intentional opaqueness denies users — and researchers alike — access to understanding the true nature of their experiences. This intransparency becomes particularly harmful when platforms suppress speech about the lives of Palestinians, as it prevents a clear understanding of whether these experiences are due to intentional censorship or other factors. The question of shadow banning vividly shows the uncertainty and obscured power dynamics in digital expression, which affect marginalized voices and the integrity of public discourse.

Organizations like Human Rights Watch and Business for Social Responsibility (BSR) have criticized these opaque, possibly biased moderation policies, particularly by Facebook and Instagram. BSR’s report, ordered by Meta itself, highlighted Meta’s unequal enforcement during the May 2021 crisis: “Meta’s actions appear to have had an adverse human rights impact on the rights of Palestinian users to freedom of expression, freedom of assembly, political participation, and non-discrimination, and therefore on the ability of Palestinians to share information and insights about their experiences as they occurred” [6]. One example from the report was the erroneous blocking of the hashtag #AlAqsa due to what they claim was a mix-up caused by a Meta contract worker who pulled it from an updated list of terms published by the U.S. Treasury Department that contained the ‘Al Aqsa Brigade’, resulting in #AlAqsa being hidden from search results (Business for Social Responsibility [BSR], 2022), effectively shadow banning a term crucial to the Palestinian narrative. Similarly, the HRW comprehensive report accuses Meta of wrongfully suppressing and removing content related to human rights issues in Israel and Palestine. The report highlights flawed content moderation mechanisms, lack of transparency, and potential governmental influence [7] as key concerns. The Arab Center for the Advancement of Social Media (7amleh) also reported over 500 violations against Palestinian digital rights across various platforms, among them account restrictions and reduced visibility.

Language, location and images: Moderation triggers across platforms. Participants commonly believed that specific keywords act as triggers for content restriction. Terms like ‘Mukawama’, ‘shaheed’ (resistance, martyr in Arabic), and ‘Palestine’ were frequently cited. For instance, P2 observed that a story on Instagram was shadow banned due to the inclusion of the word ‘Palestine.’ P2 elaborated: “I posted a story about a foundation using only the words ‘art’ and ‘Palestinian.’ While other stories had engagement around 26k to 28k views, this particular story only garnered 2,000 views. When I posted another story on a different topic, the engagement increased.” P8 and P1 also noted that posts containing the word ‘Palestine’ faced restrictions, while as P11 mentioned that stories tagged with the location ‘Palestine’ were shadow-banned. P1 added, “Anything with the words Palestine, occupation, or related terms was consistently censored,” and to “trick the algorithm”, they included characters like * and $ to workaround the restriction, which proved to be successful. P1’s tactic of using special characters like * and $ (known as hashing) to avoid restrictions of terms like “Palestine” and “occupation” is a form of algorithmic resistance. This practice, along with other circumvention methods reported by activists and their allies, aims to “trick the algorithm” and bypass its restrictions [8]. This aligns with the concept of “Algorithms of Resistance” discussed in the literature, where individuals operate within the framework of algorithms to adapt and resist their ontological power. Such resistance is a form of “repair politics,” aiming to correct or challenge the algorithm’s inherent biases or limitations (Velkova and Kaun, 2021).

There are also the problematic policies of restricting such words as ‘shaheed’ or resistance. These might be controversial terms, but they are key to the Palestinian cause. Restricting their use can be seen as over policing of legitimate speech as explained by P5: “a group was suspended because their title contained the word ‘shaheed’, which translates to ‘martyr’ in English. The group is associated with the ‘Village of the Hill of Resistance and Martyrs.’ It has been completely deleted.”

Users also assumed that violent or gory images trigger moderation. P9 said he usually avoids sharing images of wounded or killed Palestinians, assuming that would lead to sanctions on the specific content in the short term, and his account in the long term. P10 said “I know that Facebook is very careful about [...] they have an issue of like sensitive content that they will blur it, requiring user action to view it.” P10 refers
to social media’s “Sensitive Content Filter,” a feature that automatically blurs violent or graphic images, requiring user interaction to view them. This mechanism is designed to protect users from potentially upsetting imagery. P5 added that “For me, I see this content to be normal as long as it does not contain violation of people because whatever is posted is actually a representation of reality and the events happening in Palestine. For example, posting a picture of Israeli settlers or Zionists or a group of ten people committing acts of violence against a ninety years old man means that we are only presenting reality as it is.”

People experiencing the violence on the ground see the sensitive content filter as akin to censorship, as a covering of human pain and hence minimizing its alleged transformative power. Part of the reason they share photographic evidence of violence is to ‘present reality as is’ to people outside of the war zone, calling on them to be more active in resisting violence, and witness their suffering. This raises important questions about the balance between protecting users from sensitive content and accurately representing real-world events, as well as the role of algorithmic filtering in the portrayal and treatment of war violence and atrocities on social media.

In summary, our findings reveal that specific language, location tags, and images serve as key triggers for content moderation on social media platforms. The restriction of terms vital to the Palestinian cause, such as ‘Palestine,’ raises serious concerns about the over-policing of legitimate speech. Additionally, inconsistency in moderation policies across platforms adds a layer of confusion and frustration for users.

5.2. Navigating the gray area: User interpretations and platform justification

Our data highlights a disparity between the platforms’ official explanations for content moderation and the users’ and activists’ perceptions of these actions. This growing mistrust has led users to carry out a two-front resistance: offline for their cause and online against platform policies.

Platform justification: In some cases, participants reported receiving notifications about content suspension or warnings of deletion, but many discovered the moderation through alternative means. For instance, P3 realized their content was moderated when they saw others discussing it: “just as everybody else started talking about it ... I realized mine might be restricted.” Similarly, P6 discovered her restriction after attending a 7amleh event, noting that her posts had unusually low engagement during the ‘Sheikh Jarrah’ events. Despite multiple inquiries to Meta’s support team, P12 received denials about the existence of shadow banning, although she remains convinced it was implemented.

Participants who did receive notifications were often told their content was ‘sensitive’ or violated ‘community standards.’ For example, P3 contested the ‘sensitive’ label on their content, stating, “it has nothing to do with sensitivity.” Appeals to the platforms, if ever made, were generally met with silence, leading participants to form their own theories about the moderation. It’s crucial to note that the term ‘sensitive’ is highly contextual, varying across cultures and locations. Moreover, political activists inherently discuss topics that could be considered sensitive. Broad definitions of ‘sensitive content’ by platforms could stifle meaningful debate on controversial issues. Although Facebook’s community guidelines claim to allow graphic content in discussions about human rights abuses or armed conflicts, our respondents’ experiences suggest otherwise.

User interpretations: A significant 82.6 percent of survey respondents attribute the surge in moderation to what they perceive as a platform bias favoring Israel. They attribute this not just to algorithms, keywords, or violent imagery, but to a deeper systemic bias, extending beyond mere poor NLP or insufficient investment in Arabic content moderation. P5, for instance, emphasized that the issue transcends technical constraints: “I was very stressed because I had all these platforms available but couldn’t send messages. If it had been a normal Internet connection issue, we would have bought more data or found a workaround. However, this issue is beyond our means.” P6 and P8 echoed this sentiment, suggesting that social media companies “do not care” about Palestinians. According to P6, “[Platforms] don’t want to be in a situation where they have to justify and be legally accountable for content that could be interpreted as antisemitic or
similar. So, they tend to support pro-Israeli groups.” P8 noted a discrepancy in the deactivation of Palestinian pages compared to those of Israeli settlers promoting violence. P12, alleged that social media platforms are in ‘cooperation’ with Israel, leading to a systemic muting of Palestinian voices: “with every escalation of violence, the trend is censoring Palestinian voices. So, there is definitely a discrimination against them. This might be due to the platform feeling pressured by the Israeli government ... .”

This perception of alleged platform collusion with the Israeli government sparked debate among our multicultural research team, with varied perspectives based on our diverse experiences and situated knowledges. Some viewed Palestinian beliefs as part of a larger set of beliefs about platforms, that align with certain socio-technical imaginaries. Others, pointed to the seemingly unequal stance that a term such as socio-technical imaginaries implies, and asked — whose views remain conspiratorial imaginations and whose imaginaries take shape as visionary, in guiding policy and real technical abilities of these systems.

And while these beliefs about Israeli state influence are often dismissed as conspiratorial for lack of proof, this induced uncertainty is a feature of the harmful suppression of speech about the lives of Palestinians. The inability to validate ones own experiences through chains of causation is tied to the work of algorithms and other opaque mechanisms at play on social media. In fact, the powerlessness and arbitrariness they experience can be seen as a specific form of algorithmic harm. Our respondents felt that social media platforms are failing to maintain a neutral space for political discourse. Instead, they seem to enact policies that users perceive as biased, thereby undermining trust and limiting meaningful discussions on critical issues, such as the Israeli-Palestinian conflict.

Another prominent theme in user perceptions concerned the different treatment Palestinians received during the SJ events compared to Ukrainians facing Russian aggression. P3 emphasized what he saw as inconsistency in content moderation between Ukraine and Palestine, stating that the difference is “100% politically motivated ... In the case of Ukraine, there’s no restriction, even though they’re also occupied. But it’s politically charged; the sentiment is against Russia, which isn’t an ally. It’s about who they want to support, not about the content being violent.” P5 observed that while posts depicting violence against Russians are permitted, key terms like ‘Shaheed’ and ‘Al-Aqsa,’ essential to Palestinian resistance discourse, are restricted or banned. This viewpoint has been corroborated by multiple reports (Shtaya, 2022; Kelley, 2020; Human Rights Watch, 2021; Beydoun, 2022). Beydoun (2022) addressed the glaring inconsistencies in global reactions to crises, particularly contrasting the treatment of Ukrainians and Palestinians. He argued that geopolitical interests, race, and religion play a significant role in shaping these reactions. Similarly, our study’s participants felt that the moderation policies on social media platforms are not merely technical but are influenced by political agendas. The term ‘resistance,’ for instance, is allowed in posts supporting Ukraine but is restricted or banned when used in the context of Palestine. This selective moderation echoes Beydoun’s argument about the “racialization of freedom fighters and terrorists”, where Ukrainians are lauded for their resistance while Palestinians are often labeled as extremists for similar acts. Such disparities in content moderation claimed by our respondents not only reflect but also perpetuate the geopolitical biases that Beydoun discusses, undermining the platforms’ role as neutral spaces for discourse. This sets the stage for our subsequent section, which delves into the disconnect between user perceptions and platform justifications.

5.3. The gray area: Between justification and perception

At this point, our findings reveal a clear mismatch between the justifications social media platforms offer for content moderation and users’ perceptions of why their content is being moderated. Nearly all participants indicated that platform policies and notifications failed to provide compelling reasons for the moderation they encountered. For instance, P3 noted that his posts did not violate any of the community guidelines explicitly stated by the platform.

In an effort to comprehend the algorithms governing moderation, some participants sought to educate themselves on algorithmic bias. P9 attributed the issue to cultural ‘sensitivities,’ arguing that Facebook’s design and moderation policies are more in tune with Western cultural norms and that Israeli cultural
values, being more Westernized, are more compatible with platform policies on content moderation.

“There is a different cultural structure at play when we talk about Israel and a Western company. Aligning with Israel is more in line with Western sensitivities, and therefore those views are amplified on the platform and integrated into the design of the algorithms. This doesn’t mean that Palestinians have to adopt the sensitivities of Israel or the West, but it does contribute to a divergence in perceptions. While Palestinians will always consider those who died fighting for the land as martyrs, this is not a view that Israel and Western sensitivities will accommodate. Arguing won’t change this; in the end, the dominant view has the power to shape these technologies.” — P9.

This aligns with Benjamin’s (2019) view that algorithms mirror dominant social orderings. P12, an activist and digital rights expert, discussed how these policies are often enforced in partnership with the Israeli Ministry of Justice’s Cyber Unit, which reports accounts for takedown without due legal process [9]. The unit’s actions, alongside Meta’s links to the Israeli government, indicate a blend of corporate and state-level attempt at narrative control.

“Many of the policies related to hate speech, incitement to violence, terrorism, and extremism are notably opaque. These definitions are fluid and difficult to pin down, even for humans. In the Palestinian context, it’s evident to us that the platforms have adopted the narrative and terminology of the Israeli government. They have also banned all individuals and organizations on the U.S. designated terrorist list and the Treasury sanctions list. This creates clear lines or links between the platform’s definitions and those set by the Israeli or U.S. government. This is further complicated by how these policies are enforced.” — P12.

Our study reveals widespread skepticism among participants about platform accountability and transparency in content moderation, particularly related to political discourse and activism. In terms of seeking redress, users express doubt in the effectiveness of seeking clarification or appeals from platforms, anticipating generic justifications like ‘sensitive material’ or ‘community standards violations’ for moderation actions. This skepticism is reinforced by the lack of detailed explanations for increased moderation and perceived favoritism, a sentiment echoed by journalists and observers. Moreover, the difficulty in objectively confirming users’ censorship experiences highlights a power imbalance in favor of platform companies.

6. Discussion: The role of automated content moderation in the MENA region

Our study reveals a disconcerting landscape of content moderation affecting users in the MENA region, particularly focusing on the events surrounding Sheikh Jarrah (SJ) in 2021. Participants reported experiencing a spectrum of moderation actions, including sensitive content filters and shadow banning to outright account deletions. This surge in moderation, often perceived as automated and inconsistent across platforms — triggered by varying keywords and influenced by content language, context, or geolocation —
coupled with the lack of transparency in platform justifications and induced uncertainty in user perceptions, paints a bleak picture.

This landscape, reflective of non-transparent and inconsistent practices, signifies an oppressive suppression of speech, especially in relationship to Palestinian issues — amounting to blatant censorship. This underscores a significant challenge in empirically validating user experiences of censorship due to the asymmetrical power wielded by platforms. This disconnect between lived experience and their confirmability centralizes power even further within the hands of platform companies, further victimizing users already impacted by these practices. This phenomenon, indicative of algorithmic harm, raises critical questions about objectivity and fairness in managing digital discourse.

6.1. Digital authoritarianism

Our research suggests that activists and their allies encountered a form of “private censorship” [10] — presented as content moderation, implemented by corporations such as Meta and Twitter, and, as perceived by users, in collaboration with a local government, particularly in the context of pro-Palestinian narratives. This form of censorship echoes tactics traditionally associated with authoritarian regimes. In contrast, while TikTok (owned by the Chinese parent company), has been noted for allowing pro-Palestinian content (Abbas, et al., 2022), it is also known for censoring information related to the Uyghur crisis in China (Ryan, et al., 2020). This selective censorship challenges the traditional view that digital authoritarianism is exclusive to non-democratic systems.

While authoritarian states have long used digital tools for surveillance, repression, and manipulation, our findings indicate an expanding trend of tech companies in democratic societies assuming tactics taken from a digital authoritarianism textbook. This trend underscores the growing power of these corporations in shaping public discourse, a role comparable to that traditionally held by state entities. The apparent alignment of these companies with government agendas and their role in narrative moulding raises serious concerns about their place in democracies. This blurring of lines between public and private sector control over information poses significant challenges to freedom of expression and privacy.

The emergence of authoritarian-like practices under the guise of opaque content moderation policies in democracies necessitates a critical reevaluation and challenge to these forms of digital control and censorship.

6.2. Algorithmic bias, accountability, and transparency

Our findings contribute to the growing body of literature on algorithmic bias, accountability, and transparency. While existing research on algorithmic bias in content moderation has primarily focused on biases related to identity, such as race, and gender (Haimson, et al., 2021), there has been limited focus on the moderation of activist content. Our research addresses this gap by focusing on the lived experiences of activists, particularly those engaged with the SJ events in Palestine. Prior work has established that algorithms can perpetuate existing social and cultural biases (Benjamin, 2019; Burrell and Fourcade, 2021). Our study extends this understanding by demonstrating that such biases have concrete repercussions, especially in politically sensitive areas like the MENA region. The tangible impacts include a heightened sense of vulnerability, targeted harassment, and the effective silencing of activists, all of which are amplified by algorithmic processes.

The power dynamics inherent in the Israel-Palestine conflict are further complicated by the influence of private corporations. These corporations apply a distinct form of power through their content moderation algorithms, adding an additional layer of asymmetry to an already imbalanced situation. Israel, as a globally recognized nation-state, has a geopolitical advantage. Conversely, Palestinians, who are engaged in an ongoing struggle for basic freedoms and recognition of their national movement as a justified political movement and not just as terrorism, are systematically disadvantaged. This disadvantage is not solely geopolitical but is also algorithmically enforced, governing what is visible and audible in the digital public sphere. Such corporate-driven silencing further distorts the power dynamics, making it increasingly difficult
for marginalized voices to gain traction. In an era in which social media can significantly sway geopolitical processes, the imperative for algorithmic impartiality and fairness is heightened to ensure that the multiple perspectives in such conflicts are fairly represented.

6.3. Resisting algorithmic power

Our study sheds light on the lived experiences of activists fighting for the Palestinian cause, and the inventive strategies employed by them and their allies to circumvent and challenge content moderation algorithms. Although we reserve a detailed exploration of these tactics for a separate study, our findings highlight the resilience and agency of individuals, especially when confronting oppressive systems. Activists involved in the SJ events not only engaged in physical forms of resistance but also waged a digital battle against real-time algorithmic silencing. The perceived uselessness of redress and skepticism has fueled the drive to discover alternative methods to understand and adapt to algorithmic triggers creatively as they evolve. This form of platform power resistance involves engaging in “repair politics” to address perceived shortcomings in algorithmic systems (Velkova and Kaun, 2021).

This agency and creativity echo the novel use of social media for activism during the 2011 Arab Spring. Interestingly, the same human agency once used to bypass governmental controls, a decade later is being directed at resisting algorithms. The platforms once celebrated for their liberating role, are now criticized for contributing to the silencing of activists voices. In response, The Electronic Frontier Foundation has issued an open letter urging these platforms to stop silencing voices from the MENA region and to adopt more equitable content moderation policies, including investments in regional expertise with focus on war and conflict zones (Kelley, 2020).

The work of Velkova and Kaun (2021) addressed the potential role of traditional media in shaping algorithmic output and suggested that resistance in the age of algorithmic governance may increasingly depend on collaboration with traditional media. We found this notion to be true, at least in the scale of our study, as we found external validation from journalists and other observers to be invaluable. Such external validation enhances the credibility of user perceptions, thereby enriching the discourse on algorithmic accountability and transparency.

7. Implications

Our study offers two primary implications: one concerning the social ramifications of perceived biases and opaque content moderation, and the other focusing on the accountability, fairness, and transparency of social media platforms.

Social implications: Perceived bias, and lack of transparency created an environment of loss of trust and increased feelings of marginalization among activists advocating for Palestinian causes during the SJ events. This loss of trust is not merely a social inconvenience but a significant factor that alters the dynamics of online discourse and potentially exacerbates regional tensions. The sense of being systematically silenced has led to increased frustration, anxiety, and feelings of helplessness among activists. This emotional toll, in turn, contributes to the formation of echo chambers and polarized discourse, further isolating marginalized voices. These findings align with existing literature on the social implications of algorithmic bias and lack of transparency (Burrell and Fourcade, 2021; Benjamin, 2019).

Platform accountability, fairness, and transparency: The second implication of our study focuses on the platforms themselves. We contribute to the literature on algorithmic accountability by demonstrating that biases in content moderation are not merely abstract concepts but have tangible impacts, particularly in politically sensitive areas like the MENA region and the Israel-Palestine conflict. The inherent power dynamics in content moderation practices add another layer of complexity to an already fraught situation,
Opaque algorithms, transparent biases: Automated content moderation during the Sheikh Jarrah Crisis

Governing content on influential social media platforms is crucial for drawing international attention to social movements, requiring the formulation of fair, transparent, and globally respectful policies. This includes establishing independent oversight and conducting regular evaluations for fairness and impartiality, especially in politically charged or culturally sensitive contexts. A pertinent example is the critique of Meta’s automated moderation during the Israel-Hamas conflict by its Oversight Board analysis that revealed flaws in Meta’s content removal decisions, particularly related to sensitive content from the conflict. The Board highlighted the necessity for increased human intervention in moderation during crisis situations to prevent errors by automated systems [11].

It is crucial to avoid suppressing viewpoints in a conflict, because they are controversial, for commercial reasons and in order to protect profitability. Platforms must make sure to maintain impartiality and fairness in dealing with different political struggles, considering their impact on geopolitical movements and democratic dialogues. While it may sound naively optimistic, we believe that by genuinely investing in the livelihood and well-being of the users they serve, private corporations can restore trust through the enactment of fair and transparent practices.

8. Conclusion and limitations

Our study, focusing on social media content moderation in the MENA region during the 2021 Sheikh Jarrah events, reveals opaque policies and power imbalances. It enriches existing research, exploring complexities of algorithmic content moderation, its inherent biases, and decision-making processes behind it, which are often overlooked in public scrutiny and criticism (Roberts, 2019; Gillespie, 2018). We also demonstrate with qualitative data a specific form of algorithmic harm tied to content moderation, one that involves not being able to articulate and explain the causality of content decisions and their results. While we provide valuable insights, our work is limited by its regional focus, reliance on qualitative interviews, and lack of Hebrew content review [12].

Our findings, primarily focused on the experiences of those affected by content moderation, highlight a significant knowledge gap regarding the actual criteria for censorship. This is suggestive of a form of digital authoritarianism, not from states, but by private firms exerting control over public discourse through non-transparent governance mechanisms. We document the challenges activists face in navigating content moderation, often feeling that the rules are influenced by platforms’ political alignments and commercial interests. These systemic imbalances call for significant changes in both policy and the design and implementation of content moderation systems, such as clearer notifications for problematic content. We also call upon democratic governments to end their complicity in these practices. Academic research can play a pivotal role in developing fairer policies and mechanisms for social media platforms. Our focus on the silencing of Palestinian voices during an ongoing, complex conflict serves as a clear example of the power dynamics shaping public discourse on social media. This study underscores the urgent need for transparent and equitable mechanisms that can nurture and structure digital speech and debate, standing against the tide of digital authoritarianism.

9. Afterword

This paper was completed in January 2024, focusing on data from events in 2021. However, since 7 October 2023, we have observed a notable increase in social media censorship related to the escalating violence. Currently, we are actively collecting additional data on this matter, as this censorship is
Opaque algorithms, transparent biases: Automated content moderation during the Sheikh Jarrah Crisis

manifesting in various forms.

Our research began in the fall of 2021, a year marked by numerous reports of forced displacement of Palestinians in occupied East Jerusalem and that led to a war between Hamas and Israel in Gaza and violent clashes between Jews and Arabs within Israel. The current wave of censorship is more centered on the ongoing war in Gaza, the devastating human toll it has inflicted, and the unprecedented public debate concerning Palestinian rights. This has been followed by widespread demonstrations and protests worldwide.

It is important to note that our study’s scope does not encompass the on-going events occurring post-7 October 2023, which we are witnessing unfold on our screens. We emphasize that our data collection and analysis predate these developments, so our findings are not directly applicable to the current situation. The present cases of social media censorship, still evolving, will be the focus of our future research. Our study aims to provide a comprehensive understanding of these evolving dynamics in digital communication and their broader social implications.

About the authors

Norah Abokhodair is an independent Ph.D. researcher specializing in algorithmic bias and cross-cultural research.
E-mail: norah [dot] abokhodair [at] gmail [dot] com

Yarden Skop is a research associate and Ph.D. candidate at the University of Siegen.
E-mail: yarden [dot] skop [at] uni-siegen [dot] de

Sarah Rüller is a research associate at the University of Siegen.
E-mail: sarah [dot] rueller [at] uni-siegen [dot] de

Konstantin Aal is a research associate of the Department for Business Informatics and New Media of the University of Siegen.
E-mail: konstantin [dot] aal [at] uni-siegen [dot] de

Houda Elmimouni is a Computing Innovation Fellow and postdoctoral researcher in the Department of Informatics at Indiana University Bloomington.
E-mail: houdaelmimouni [at] fulbrightmail [dot] org

Notes

5. Savolainen, 2022, p. 1,094.
7. The report indicates that the Israeli government’s Cyber Unit has been actively flagging content for removal during the events, often bypassing legal processes. Facebook has not disclosed the nature and
number of these requests.

8. To circumvent moderation, participants used various strategies, a topic we will explore in future work and is beyond the scope of this paper.

9. Additionally, Israeli journalists have pointed out that Meta maintains strong ties with the Netanyahu regime. For instance, Jordana Cutler, a former advisor and campaign strategist for the Likud party, was appointed as the Public Policy Director for Israel and the Jewish Diaspora at Meta (Gil’ad, 2016). In 2016, the Israeli government announced a collaborative effort with Facebook to address ‘incitement’ on the social media platform (Associated Press in Jerusalem, 2016).


11. The active role of Meta’s Oversight Board in events after 7 October 2023, is commendable. However, it’s important to note that this does not directly relate to the data in our current study, which focuses on earlier events. The continuous evolution of content governance on these platforms remains a vital area for further research and scrutiny.

12. Our study did not analyze Hebrew content, which limits our ability to compare moderation practices across different languages within the Israel-Palestine context. Given the politically sensitive nature of the Israel-Palestine conflict, such comparison could offer valuable insights into content moderation’s fairness and impartiality. Future research addressing this gap could provide a more holistic understanding of content moderation in politically sensitive areas.

References


Opaque algorithms, transparent biases: Automated content moderation during the Sheikh Jarrah Crisis


doi: https://doi.org/10.1057/9781137020925_1, accessed 10 March 2024.


Jason Kelley, 2020. “A decade after the Arab Spring, platforms have turned their backs on critical voices in
Opaque algorithms, transparent biases: Automated content moderation during the Sheikh Jarrah Crisis.


Renkai Ma and Yubo Kou, 2022. “‘I’m not sure what difference is between their content and mine, other than the person itself’: A study of fairness perception of content moderation on YouTube,” Proceedings of the ACM on Human–Computer Interaction, volume 6, number CSCW2, article number 425, pp. 1–28.


doi: https://doi.org/10.1109/ASONAM.2014.6921619, accessed 10 March 2024.


Opaque algorithms, transparent biases: Automated content moderation during the Sheikh Jarrah Crisis


Kate Starbird and Leysia Palen, 2012. “(How) will the revolution be retweeted? information diffusion and


Hibby Thach, Samuel Mayworm, Daniel Delmonaco, and Oliver Haimson, 2022. “(In)visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit,” *New Media & Society* (18 July).


Opaque algorithms, transparent biases: Automated content moderation during the Sheikh Jarrah Crisis


---

**Editorial history**

Received 21 September 2023; accepted 10 March 2024.

---

This paper is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/).


doi: [https://dx.doi.org/10.5210/fm.v29i4.13620](https://dx.doi.org/10.5210/fm.v29i4.13620)