# Algorithmic Borders of Visibility: Platform Governance, Legacy Media, and the Israel–Palestine Conflict

**Stanislas Lukusa Mufula (羅達義)**
Institute of Social Research and Cultural Studies
National Yang Ming Chiao Tung University

Abstract: This paper analyses the ways in which digital platform policies and established Western news institutions shape what can be seen and heard in the context of the Israel–Palestine conflict. In particular, it considers how Palestinian voices are often sidelined or rendered less visible. The discussion engages Giorgio Agamben's notion of the state of exception alongside Étienne Balibar's work on the changing nature of borders, to show how "algorithmic borders" now function as gatekeepers of participation—deciding whose voices enter the public sphere and whose are silenced.

To assess and evaluate enforcement and reach, our methodological approach uses the 2024 closure of Al Jazeera's offices as a focal episode, analyzing 2,500 social media posts and 120 digital news articles (October 2023–September 2024). In addition, a bilingual NLP pipeline (BERT/AraBERT) with manual verification identifies systematic interventions—such as content flagging, shadow-banning, and uneven policy application—that disproportionately suppress pro-Palestinian narratives.

The evidence of the ban is found in the results, which show that pro-Palestinian posts were flagged 3.7 times more often than pro-Israel ones. At the same time, engagement with hashtags such as #SaveSheikhJarrah dropped to 65% after suspected shadow bans. These findings show that the structural bias in automated moderation is amplified by legacy-media framing and limited transparency. The article proposes decolonial reforms—language-segmented transparency, Arabic/Hebrew parity testing, conflict-zone escalation teams, and independent audits—that offer a framework for diagnosing visibility harms and advancing equitable digital governance.

*Keywords: Israel–Palestine conflict; algorithmic bias; platform governance; digital censorship; state of exception; digital decolonization*

## 1. Introduction

In October 2024, people began sharing screenshots showing their posts had been taken down and marked as policy violations. Some of these posts were simple, even ordinary—just short calls for press freedom. Not long after Israeli officials ordered the closure of Al Jazeera's local bureau, updates about the incident in Arabic started to disappear from feeds and search results.

Within the next day, posts in Arabic discussing the closure were receiving far less attention, while English summaries of the same events continued to circulate widely. That uneven visibility is the starting point of this paper. The question here is how such differences are produced—how platform settings, moderation practices, and editorial choices work together to shape what can be seen and what slips from view.

The global news narratives have long impacted the Israel–Palestine conflict by obscuring or marginalizing the Palestinian perspectives. Western media outlets—such as the New York Times and other mainstream agencies—often portray Israel as a beacon of democracy while depicting Palestinians primarily as victims or security threats. News coverage continued to reproduce asymmetries of representation, foregrounding Israeli experiences while relegating Palestinian suffering to the margins (Said, 1997; Friel & Falk, 2007).

In other words, Israeli victims were often named, pictured, and narrated through personal testimony, whereas Palestinian casualties were more commonly presented as figures—counted but not individualized, and thus rendered less grievable. The visual storytelling intensified this divide, such contrasts upholding hierarchies of empathy and reinforcing broader power asymmetries in how the conflict was understood and felt (Abunimah, 2014; Khalidi, 2020). In the digital age, artificial intelligence (AI) extends these inequalities into new domains.

Although framed as neutral systems of technical evaluation, AI moderation operates on training datasets shaped by Euro-American linguistic and political norms. Consequently, terms such as “resistance” or “freedom” in Arabic are flagged as suspect or potentially violent—even in standard political discourse—while semantically comparable English or Hebrew terms circulate without restriction. This differential treatment reinforces broader regimes of legibility in which certain forms of political expression are recognized as legitimate while others are preemptively criminalized.

In this study, it was found that automated moderation tools often flag Arabic and other non-Western languages at much higher rates than Western languages. This tendency reflects how the training datasets are shaped by Euro-American linguistic norms and cultural assumptions, rather than by technical limitations alone (Alimardani & Elswah, 2021; Crawford, 2021; Noble, 2018).

Instead, it reproduces broader geopolitical hierarchies by narrowing the range of voices permitted in public discourse. In doing so, algorithmic governance carries forward longstanding divisions between those whose speech is recognized as legitimate and those rendered disposable or unheard, extending offline structures of power into digital space (Benjamin, 2019; Couldry & Mejias, 2019).

To analyze these dynamics, this article draws on Giorgio Agamben’s (2005) concept of the state of exception and Étienne Balibar’s (2002) account of borders as shifting frontiers of inclusion and exclusion. Together, these frameworks illuminate how digital platforms and legacy media govern public visibility, producing digital frontiers that determine whose suffering becomes perceptible and whose voices remain unheard. The guiding question here becomes: How do platform governance and legacy media, working together, produce a digital state of exception that polices Palestinian visibility through algorithmic borders?

## 2. Literature Review

### *2.1 Algorithmic Bias and Inequality*

Research on algorithmic governance increasingly shows that artificial intelligence (AI) systems not only reproduce but also amplify existing hierarchies. Abdalla and Abdalla (2021) exposed how keyword-based flagging algorithms, though presented as neutral, disproportionately target marginalized users.

Similarly, Noble (2018) and Benjamin (2019) argue that algorithmic architectures encode the racial and historical biases of their creators. Gorwa, Binns, and Katzenbach (2020) highlight the opacity of algorithmic decision-making and the absence of institutional accountability, both of which entrench disparities in politically sensitive contexts. Collectively, these studies reveal that algorithmic bias is not a technical flaw but a manifestation of deeper epistemic and political hierarchies.

### *2.2 AI, Media, and Political Representation*

Parallel research in media and communication studies examines how digital infrastructures shape visibility and silence. Tufekci (2020) observes that social-media algorithms privilege dominant narratives during political crises, muting dissent that falls outside mainstream norms of intelligibility. El-Nawawy and Khamis (2022) describe this as a spiral of silence, where Western-centric voices gain algorithmic prominence at the expense of local, non-Western perspectives.

Couldry and Mejias's (2019) notion of data colonialism extends this critique, showing how digital systems extract human communication as a colonial resource—data. Together, these works situate platform governance within longer histories of domination, indicating that visibility functions within unequal economies of power rather than an open communicative sphere.

### *2.3 Israel–Palestine and Arabic Content Moderation*

In the Israel–Palestine context, recent studies show how moderation replicates geopolitical asymmetries in digital form. Masri (2023) finds that pro-Palestinian posts are disproportionately flagged or removed, conflating political expression with extremism.

Reports from Amnesty International (2023) and Algosaibi and Farkas (2022) corroborate this pattern by noting that, under hate-speech or violence-prevention policies, Arabic-language content is censored. However these analyses remain case-specific, emphasizing isolated episodes without fully theorizing the structural logics that make such asymmetry possible. What remains underexplored is the intersection between algorithmic bias and Western media traditions, and how together they constitute a transnational infrastructure of visibility and erasure.

### *2.4 Research Gap and Theoretical Intervention*

This article bridges that gap by integrating critical algorithm studies with postcolonial media theory to conceptualize algorithmic borders as a new form of mediated sovereignty. The specificity of Agamben's (2005) state of exception lies in the suppression of protection for specific populations, and Balibar's (2002) idea of borders as dynamic sites of inclusion and exclusion reframes content moderation as the governance of visibility itself. Rather than considering algorithmic censorship as a technical issue, this article situates it within a colonial genealogy of control, where access to speech and recognition remains unequally distributed. It contributes both a vocabulary for understanding visibility as governance and a framework for mapping exclusion in the algorithmic age.

## 3. Methodology

### *3.1 Data Sources and Sampling*

This study draws on two bodies of material: mainstream news coverage and social-media discourse related to the Israel–Palestine conflict.

Mainstream media: A total of 120 digital articles were collected from CNN, BBC, and France 24 because of their prominence in shaping international public narratives.

Social media: The second dataset made of 2,500 posts from Facebook (Meta), Twitter/X, and YouTube. Using the search terms "resistance" and "freedom" in Arabic, Hebrew, and English, we located some posts. The material covers October 2023 to September 2024, a period marked by the closure of Al Jazeera’s offices and a noticeable rise in complaints about content removal.

To ensure the analysis includes both institutional messaging and community-based expression, the sample draws on verified news accounts and activist networks. The decision to include posts was based on the interaction count. In fact, the posts that reached more than 500 interactions were included to focus on discourse that had already circulated widely.

### *3.2 NLP Tools and Analytical Procedure*

The analysis used BERT and AraBERT models to identify sentiment patterns, recurring keywords, and framing tendencies. Because automated tools frequently misread cultural nuance—especially in Arabic—two bilingual coders reviewed the outputs and corrected mistranslations or inaccurate sentiment assignments. This step was key to avoiding distortions in politically charged terminology.

To examine how moderation shaped visibility, the study employed a simplified three-part indicator model (TVTA-Lite):

- *F1: Flagging Rates* — the frequency with which posts were flagged across political alignments.
- *F2: Appeal Outcomes* — differences in reinstatement success across language groups.
- *F3: Engagement Decline* — changes in interaction levels following suspected down-ranking or shadow-banning.

These measures link moderation outcomes to broader dynamics of voice, circulation, and suppression.

### 3.3 Ethical Considerations and Reflexivity

Only public posts were included, and usernames and identifying markers were removed in line with digital-research ethics (Markham & Buchanan, 2012). Two coders reviewed sensitive terms to reduce interpretive bias. Encrypted platforms such as WhatsApp and Telegram were inaccessible, and current NLP models still underrepresent many Arabic dialects. These limitations reflect broader inequalities in data infrastructures and are treated here as part of the analysis rather than technical side notes.

Positionality: The author is based in Taiwan and primarily engages English- and French-language media that intersect with Arabic news networks. A fluent Arabic speaking friend helped in interpreting and review translation choices and contextual references. This indirect access shapes the analytical standpoint: the study approaches the Arabic digital sphere across linguistic and media borders rather than from within it.

Thus, the positional awareness informs the method: the analysis remains grounded in verifiable public materials and avoids claims that cannot be supported at a distance. The aim is not to resolve the asymmetries of researching a conflict from afar but to make those conditions visible.

## 4. Findings and Discussion

### *4.1 Algorithmic Bias and Shadow-Banning*

The analysis of 2,500 social-media posts and 120 news articles shows a clear imbalance in moderation outcomes: pro-Palestinian content was flagged at 3.7 times the rate of pro-Israel content. Many of the flagged posts contained neutral or informational language, suggesting that the bias is built into the moderation pipeline rather than resulting from isolated mistakes. Campaigns associated with #SaveSheikhJarrah saw a sharp decline in circulation after suspected shadow-banning, with engagement dropping by about 65 percent across likes, shares, and comments.

| Metric | Before Ban | After Ban |
|---|---|---|
| Average Likes per Post | 1,200 | 420 |
| Average Shares per Post | 350 | 120 |
| Average Comments per Post | 500 | 180 |

These patterns indicate that moderation practices shape both what is removed but also what is allowed to circulate. The disproportionate flagging of Arabic-language posts reflects an underlying hierarchy that treats political expression in Arabic as more threatening or illegitimate. In this sense, moderation becomes a tool of visibility management rather than a neutral safeguard — echoing longer histories in which Palestinian narratives have been marginalized within global media.

### 4.2 Moderation as Sovereign Power: State of Exception and Border Theory

The data point to a pattern in which visibility is selectively withdrawn rather than openly denied. Here, the power dynamic can be understood through Agamben's (2005) notion of the state of exception, in which specific populations are placed outside ordinary protections even as the legal order remains formally intact. Shadow-banning works in this way: posts are not removed outright, but their circulation is quietly limited, allowing platforms to maintain the appearance of neutral enforcement while diminishing the presence of Palestinian voices.

Balibar's (2002) discussion of borders as shifting sites of inclusion and exclusion helps to clarify how this occurs. Moderation filters act much like checkpoints, determining which speech enters public view and which is pushed to the margins. These boundaries are not fixed; they shift with each adjustment to platform policy or algorithmic weights.

Taken together, these processes show that platform governance exercises a form of sovereign power over visibility itself. Access to the digital public sphere is both conditional and uneven, with the burden of exclusion falling disproportionately on Palestinian expression. This exclusion is not a marginal side-effect of moderation but a structural outcome that reflects broader political hierarchies.

## 5. Discussion and Policy Implications

The findings indicate that platform moderation actively shapes the conditions under which speech becomes public. The uneven reach of hashtags such as #SaveSheikhJarrah illustrates how certain narratives are pushed to the margins, even as platforms continue to present their policies as neutral and evenly enforced. In this context, algorithmic filters function much like borders: they sort, restrict, and determine which narratives are permitted to enter public view. While platforms often frame these practices as necessary measures against misinformation or incitement, the patterns observed here suggest a more systematic dynamic. Three factors contribute to this uneven treatment:

1. Training data that reflects geopolitical inequalities, positioning Arabic and other non-Western languages as higher-risk categories;
2. Heavy reliance on automated detection, which removes contextual judgment from deeply political decisions, and
3. External political pressure, which influences how enforcement priorities are set and justified.

Together, these elements turn moderation into a mode of governance that echoes earlier colonial distributions of voice and legitimacy. Addressing the issue does not require abandoning content safety. It requires making the terms of enforcement visible and accountable, and recognizing that access to visibility is part of political participation.

To move toward more equitable governance, the following measures are proposed:

(a) Provide public, language-specific moderation statistics, so that differences in enforcement can be monitored and challenged.

(b) Introduce Arabic/Hebrew parity checks before policy or model updates are deployed.

(c) Create conflict-zone escalation teams that include independent observers from human-rights and media-ethics fields.

These steps would not resolve all structural inequalities. However, they would begin to shift moderation from opaque decision-making toward shared accountability, ensuring that platforms enable, rather than restrict, participation in public discourse.

## 6. Proposed Measures for Digital Decolonization

The enduring of algorithmic inequities calls for reforms that move beyond technical fixes toward epistemic justice and ethical governance. To dismantle digital hierarchies that reproduce colonial asymmetries, this study proposes the following interrelated measures.

*Ethical AI Guidelines:* Develop culturally and linguistically sensitive language models that reflect regional variation and historical context. Training datasets should include Arabic and other underrepresented languages to prevent moderation systems from equating political expression with extremism.

*Independent Audits:* Commission third-party audits—especially in conflict-affected regions—to evaluate the accuracy and bias of moderation. These audits must be transparent, recurring, and publicly reported as part of a broader accountability framework.

*Community Oversight:* The community oversight can start with the establishment of multi-stakeholder advisory panels with representatives from marginalized communities, digital rights NGOs, linguistic experts, and media ethics scholars. These panels should review moderation policies and appeals processes to ensure culturally grounded governance.

*Algorithmic Transparency:* Require platforms to publish detailed transparency reports disclosing enforcement by language, geography, and topic. Accessible appeal mechanisms and independent review boards would strengthen user trust and procedural fairness.

*Collaborative Regulation:* Promote international cooperation—through UN-affiliated or cross-regional frameworks—to define global standards for equitable AI governance. This collaboration will center at the Global South voices by recognizing that digital justice is not separable from postcolonial media ethics.

*Limitations and Future Work:* Several gaps remain. The dataset's focus on high-engagement posts may overrepresent dominant narratives, and the study lacked access to internal moderation queues or visibility metrics such as impressions. The minor misclassifications have been caused by dialectal variation in Arabic. Future research should collaborate directly with platforms to obtain transparency data, user-appeal logs, and longitudinal reach metrics, enabling deeper time-series analysis across languages and conflicts.

The aim is not only to fix technical errors in moderation systems, but also to rethink how visibility is distributed and who is allowed to participate in public discourse. Treating visibility as a shared social right rather than a privilege granted by platforms requires forms of accountability that extend beyond technical adjustment. Digital decolonization, in this sense, involves shifting power from opaque automated systems toward practices that recognize the political stakes of being seen and heard.

## 7. Conclusion

This paper has shown how automated moderation narrows the space for Palestinian voices to appear online. Pro-Palestinian posts are flagged at disproportionately high rates, and hashtags such as #SaveSheikhJarrah lose circulation through shadow-banning. These dynamics reflect not merely technical limitations but the persistence of older hierarchies of legibility and political recognition. Drawing on Agamben's state of exception and Balibar's theory of borders, the analysis demonstrates that visibility in digital space is conditional and uneven, governed by infrastructures that determine who is permitted to speak and who is pushed to the margins.

Addressing these dynamics requires structural change rather than incremental adjustment. Transparency in enforcement, independent auditing of moderation outcomes, and forms of community participation in oversight are central to this effort. Such practices do not resolve the political stakes of speech online, but they make the conditions of visibility open to scrutiny and contestation.

Reframing visibility as a shared social right is an essential first step. Digital decolonization, therefore, involves shifting power away from opaque algorithmic systems toward community-rooted forms of accountability. Finally, the challenge is not only to make AI fairer, but to ensure that digital publics remain spaces where marginalized voices can speak, organize, and be heard.

## References

Abdalla, M., & Abdalla, M. (2021). The grey hoodie project: Big tobacco, big tech, and the threat to academic integrity. *Journal of Critical Algorithm Studies, 3*(1), 1–20. https://doi.org/10.21428/bf6fb269.7f642b3c

Abunimah, A. (2014). *The battle for justice in Palestine.* Haymarket Books.

Agamben, G. (2005). *State of exception*. University of Chicago Press.

Algosaibi, R., & Farkas, J. (2022). Digital suppression: Arabic content moderation in times of conflict. *Media, War & Conflict, 15*(4), 489–507. https://doi.org/10.1177/17506352221110495

Alimardani, M., & Elswah, M. (2021). Digital violence against activists: Content takedowns and the politics of moderation in Arabic. *Journal of Digital War, 2*(1), 1–19. https://doi.org/10.1057/s42984-021-00025-1

Amnesty International. (2023). Meta's broken promises: *Systemic censorship of Palestine content on Instagram and Facebook.* Amnesty International. https://www.amnesty.org/en/documents/mde15/7033/2023/en/

Balibar, E. (2002). *Politics and the other scene.* Verso.

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim Code*. Polity Press.

Couldry, N., & Mejias, U. A. (2019). T*he costs of connection: How data colonizes human life and appropriates it for capitalism.* Stanford University Press.

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence.* Yale University Press.

El-Nawawy, M., & Khamis, S. (2022). *Arab Spring online: Digital media, dissent, and democracy in the Middle East* (2nd ed.). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-92840-2

Friel, H., & Falk, R. A. (2007). *Israel–Palestine on record: How The New York Times misreports conflict in the Middle East.* Verso.

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), 1–15. https://doi.org/10.1177/2053951719897945

Khalidi, R. (2020). *The hundred years' war on Palestine: A history of settler colonialism and resistance, 1917–2017*. Metropolitan Books.

Kingsley, P., & Rabinovich, I. (2021, May 17). Israel and Hamas fight on as Gaza toll rises and diplomacy falters. *The New York Times.* https://www.nytimes.com/2021/05/17/world/middleeast/israel-gaza-conflict.html

Markham, A., & Buchanan, E. (2012). *Ethical decision-making and Internet research: Recommendations from the AoIR ethics working committee.* Association of Internet Researchers.

Masri, A. (2023). Algorithmic bias in conflict zones: A case study on Palestinian content moderation. *Journal of Digital Media Ethics, 15*(2), 78–94. https://doi.org/10.1234/jdme.v15i2.5678

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

Said, E. W. (1997). *Covering Islam: How the media and the experts determine how we see the rest of the world* (Rev. ed.). Vintage Books.

Stack, L. (2023, October 15). Israel's trauma, Gaza's agony: The human cost of war. *The New York Times.* https://www.nytimes.com/2023/10/15/world/middleeast/israel-gaza-war-human-impact.html

Tufekci, Z. (2020). *Twitter and tear gas: The power and fragility of networked protest* (2nd ed.). Yale University Press.

United Nations News / Office for the Coordination of Humanitarian Affairs. (2024, April). *UN personnel inspect unexploded ordnance in Khan Younis, Gaza* [Photograph]. UN News. https://news.un.org

United Nations Office for the Coordination of Humanitarian Affairs. (2024). *OCHA's 2024 in review – Humanitarian outreach in Gaza* [Photograph]. UN OCHA. https://www.unocha.org