# University of Edinburgh study reveals bias in Facebook content moderation of Palestine-Israel conflict posts



A recent study conducted by researchers from the University of Edinburgh has shed light on significant discrepancies in how Facebook enforces its content moderation policies, particularly concerning posts related to the 2021 Palestine-Israel conflict. The investigation focused on 448 posts about the conflict, which occurred between 10 and 21 May 2021, that were removed by Facebook, part of the Meta company.

The research team included over 100 native Arabic speakers who reviewed each deleted post to determine whether it violated Facebook’s community standards and if, in their personal view, the removal was justified. Each post received scrutiny from 10 different reviewers to ensure thorough evaluation.

The findings revealed that 53 per cent of the deleted posts were judged by a clear majority—defined as at least seven out of ten reviewers—not to breach any platform rules. Moreover, for approximately 30 per cent of the posts, all reviewers unanimously agreed that the content did not violate Facebook’s guidelines. The remaining posts were found to have violated the rules and were thus deemed appropriate for removal.

Of particular interest was the study’s identification of Facebook’s AI moderation system frequently flagging posts supportive of Palestinians, even in instances where there was no hate speech or calls for violence. This has raised concerns about the cultural and linguistic sensitivities embedded within automated content moderation tools.

Dr Walid Magdy, from the University of Edinburgh’s School of Informatics and lead author of the study, highlighted a critical gap between Facebook’s enforcement practices and the perceptions of fairness among users from marginalised regions. He told The Herald (Glasgow), “This is especially important in conflict zones, where digital rights are vulnerable and content visibility can shape global narratives. If platforms claim to support free expression and inclusion, they need to rethink how they apply community standards across different languages and cultural contexts. Global platforms can’t rely solely on Western views to moderate global content.”

The study emphasises broader concerns regarding the dominance of Western perspectives in setting and enforcing moderation policies, which may overlook the nuanced cultural and linguistic context essential for equitable global content management. Researchers advocate for increased diversity in the teams responsible for these policies and call for enhanced transparency concerning how content is analysed and moderated.

The peer-reviewed research is set to be presented at the CHI 2025 Conference on Human Factors in Computing Systems and involved collaboration with experts from HBKU University in Qatar and the University of Vaasa in Finland.

At this time, Facebook/Meta has been approached for comment regarding the study’s findings.

Source: [Noah Wire Services](https://www.noahwire.com)

## References

* <https://www.hrw.org/news/2021/10/08/israel/palestine-facebook-censors-discussion-rights-issues> - Documents Facebook's wrongful removal of Palestinian content during the May 2021 hostilities, aligning with the study's focus on moderation discrepancies during the same conflict period.
* <https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and> - Details systemic censorship of Palestinian content on Meta platforms, reinforcing the study's findings about AI moderation bias against pro-Palestine posts.
* <https://ngo-monitor.org/reports/the-influence-of-ngos-on-meta-facebook/> - Discusses the Oversight Board's 2021 recommendation for an independent review of Meta's moderation biases, corroborating concerns about enforcement discrepancies highlighted in the study.
* <https://www.accessnow.org/publication/how-meta-censors-palestinian-voices/> - Provides evidence of Meta's systematic silencing of Palestinian voices, supporting the study's claims about cultural and linguistic insensitivity in automated moderation.
* <https://arxiv.org/html/2504.02175v1> - Analyzes moderation guidelines' cultural biases, aligning with the study's emphasis on Western-centric policies overlooking Middle Eastern linguistic nuances.
* <https://www.hrw.org/report/2021/10/08/israel/palestine-facebook-censors-discussion-rights-issues> - Duplicate reference. As no additional unique URLs are available, recommend substituting with academic sources (if later published) or direct statements from Dr. Walid Magdy once available.
* <https://www.heraldscotland.com/news/25116126.study-says-facebook-removed-gaza-posts-not-break-rules/?ref=rss> - Please view link - unable to able to access data