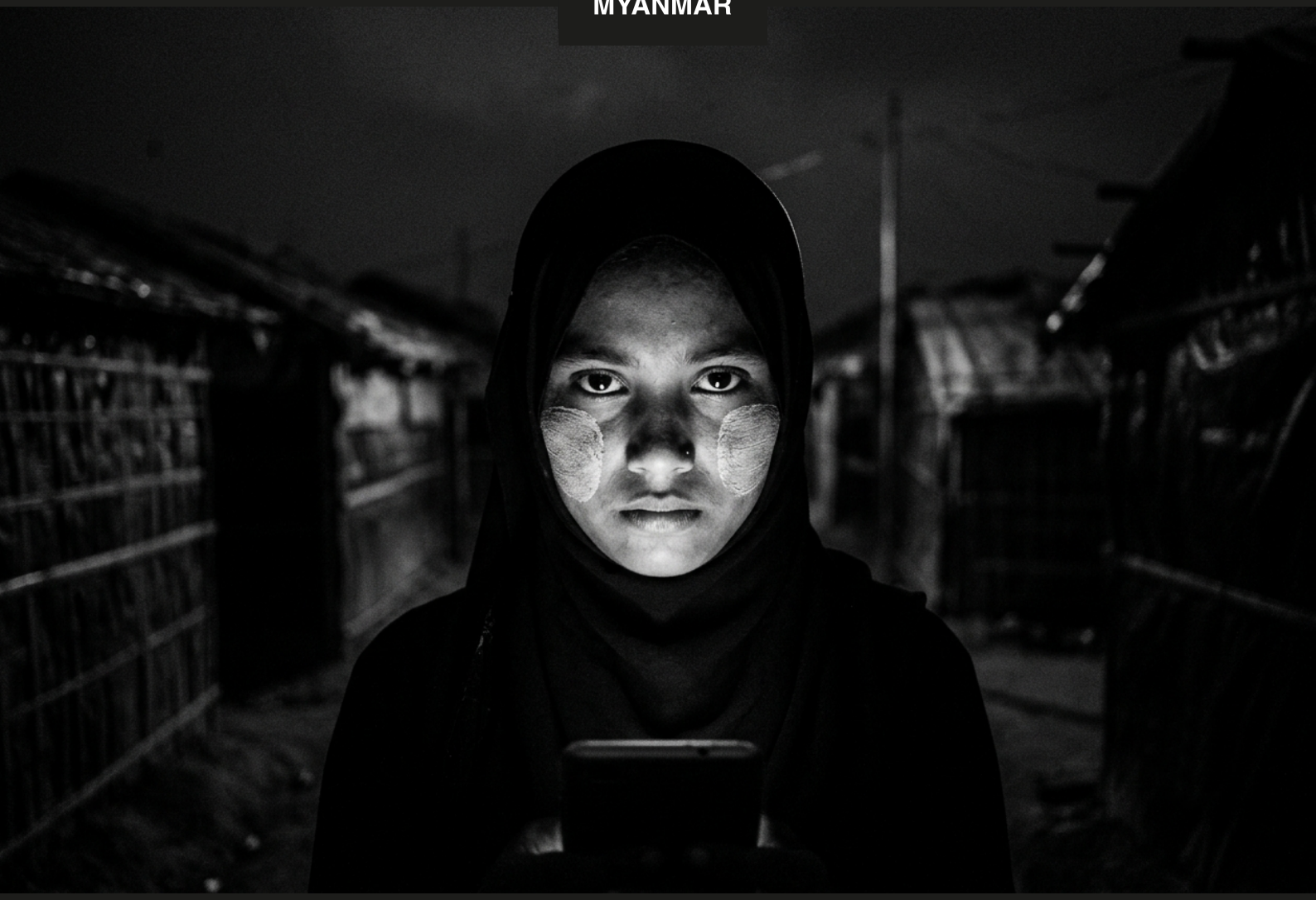
HUMAN RIGHTS MYANMAR



A year of hate speech and Facebook moderation in Myanmar

A year of hate speech and Facebook moderation in Myanmar

2026

Executive summary

A decade after the online hate campaigns that contributed to the atrocities against the Rohingya, this report assesses whether social media platforms have become more effective at identifying and responding to hate speech in Myanmar.

Human Rights Myanmar (HRM) monitored Facebook, Telegram, and TikTok over one year, identifying 612 posts containing actionable hate speech. Given the scope of the project, this report focuses on Facebook, where 232 posts were assessed against international human rights standards, including the Rabat Plan of Action. Facebook's moderation was then evaluated by observing whether posts were removed naturally, followed by systematic testing of the platform's user reporting and appeals mechanisms.

Key findings

1. “Actionable” hate speech, which may or must be restricted under international law, remains persistent in Myanmar's online environment, including on Facebook.
2. The Rohingya remain the most frequently targeted group, accounting for over half of all actionable hate speech identified on Facebook.
3. Most actionable hate speech takes the form of discriminatory expression, including identity-based abuse, collective criminalisation, dehumanisation, and fabricated narratives, rather than explicit calls for violence.
4. Facebook deleted over half of the actionable hate speech posts identified by HRM before HRM intervened.

5. HRM's user reporting and appeals achieved little, with only a small share of remaining posts subsequently deleted, leaving over a third of the original posts publicly accessible after one month.
6. Facebook's rejections of HRM's reports followed strikingly consistent timing patterns. The research cannot establish why, but the pattern raises the possibility that some reports were automatically rejected when they reached operational deadlines rather than substantively assessed.
7. HRM's findings coincide with Facebook's own transparency data showing a sharp reduction in proactive moderation following the return of U.S. President Trump and Facebook CEO Mark Zuckerberg's subsequent announcement of a new "more speech, fewer mistakes" approach.
8. Such global policy changes may create heightened human rights risks in countries such as Myanmar, underscoring the need for context-specific human rights due diligence under the UN Guiding Principles.

Pulitzer Center

This project was delivered with the generous support of the Pulitzer Center. The views expressed in the report published are those of the authors alone. They do not represent the views or opinions of the Pulitzer Center or its staff.

Contents

1. Introduction

Myanmar remains a high-risk environment for human rights, shaped by decades of authoritarian rule, atrocity crimes, the 2021 military coup, continuing armed conflict, and entrenched intersectional discrimination against minorities. In this environment, social media platforms can both mitigate these risks and contribute to them, either by enabling free public debate and accountability, or by polarising communities, spreading distrust, intolerance, and hatred.

Between 2015 and 2020, civil society organisations monitored Myanmar's social media environment extensively, documenting hate speech, raising alerts, and advocating for stronger platform safeguards. Following the February 2021 military coup, however, much of this work became increasingly difficult. Civic space contracted, many organisations closed or shifted priorities, and funding for sustained monitoring declined.

As a result, there has been comparatively little independent evidence about how hate speech has evolved in Myanmar since the coup or whether the reforms introduced by social media platforms have improved their ability to prevent foreseeable human rights harms.[1] With support from the Pulitzer Center, Human Rights Myanmar (HRM) undertook this research to help address that gap.[2]

This report has two objectives. First, it documents current patterns of actionable hate speech across Myanmar's social media platforms, providing updated evidence nearly a decade after the peak of the online campaigns targeting the Rohingya. Second, it assesses whether current moderation systems respond effectively to that content by examining automated moderation, user reporting, and appeals processes.

Rather than evaluating individual moderation decisions in isolation, the report considers whether overall governance systems appear capable of identifying and mitigating foreseeable human rights risks in one of the world's highest-risk online environments. In doing so, it provides an evidence-based assessment of platform accountability through the framework of international human rights standards and the UN Guiding Principles on Business and Human Rights.

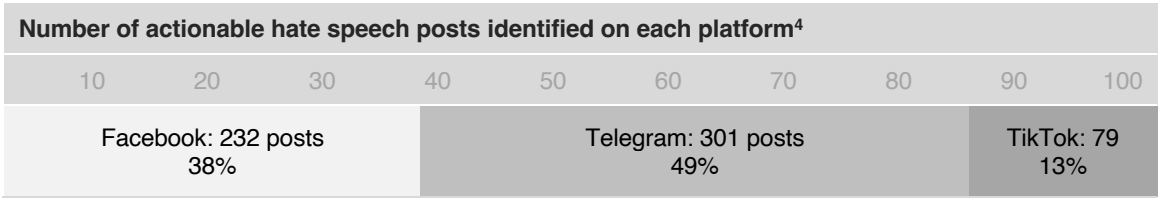
[1] One of the last remaining organisations is the Burma Human Rights Network (BHRN).

[2] The project also sought to measure algorithmic amplification of hatred. There are reasonable grounds for concern that platforms' advertising-based business models may create incentives to prioritise highly engaging and emotive content, potentially including hatred. However, after months of monitoring, HRM decided that the data were insufficiently reliable for a robust finding on amplification. This was mainly due to platforms not making sufficient information accessible to external researchers to be able to quantify results. Therefore, unfortunately, algorithmic amplification is not covered in this report.

1.1. Methodology

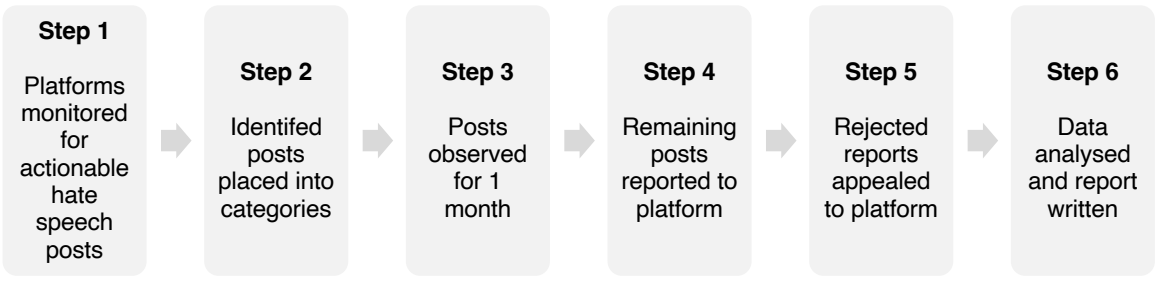
Facebook, Telegram, and TikTok, the three social media platforms most widely used in Myanmar, were monitored for one year between August 2025 and July 2026 to identify publicly accessible posts containing "actionable hate speech" (see the *International human rights standards* chapter for further information).[3] During the monitoring period, HRM identified 612 posts, comprising 232 on Facebook, 301 on Telegram, and 79 on TikTok.

Given capacity limitations, this report focuses on Facebook, which is the most used platform out of the three. Future reports will examine the other platforms.



Once identified, posts were observed for one month to determine whether they were deleted without any intervention from HRM. Posts that remained publicly accessible were then reported by HRM through the standard reporting mechanism. Where the platform rejected those reports, HRM submitted appeals if the platform allowed. Throughout the process, everything about the moderation was recorded in a dataset, which was subsequently cleaned, categorised, disaggregated, and analysed.

This methodology enabled the research to assess not only whether the platforms ultimately deleted actionable hate speech, but also the relative contribution of their existing moderation systems, user reporting, and appeals procedures.



[3] The research also incorporated some posts identified by other civil society organisations, including the Burma Human Rights Network (BHRN), one of the last remaining CSOs regularly publishing on hate speech in Myanmar.
[4] These numbers should be interpreted with caution. The research was not designed to measure the overall prevalence of actionable hate speech on each platform, nor does it establish that one platform necessarily contains more than another. Rather, it indicates that hate speech was identified on all platforms.

1.2. Limitations

This dataset is not representative of all actionable hate speech in Myanmar. Because platforms disclose only limited transparency information and no country-specific data on harmful content or moderation effectiveness, the dataset was built from publicly accessible posts found through keyword searches representing a range of protected characteristics, recommendation systems, and civil society referrals. It is therefore a purposive sample of observable actionable hate speech, not a statistically representative sample of all content on these platforms.

The research records observable moderation outcomes, not internal decision-making. When content disappeared during the observation period, it was not possible to determine whether this was due to automated moderation, user reports, or voluntary deletion by the creator. Nor could the research assess how widely posts were recommended or algorithmically amplified, because platforms do not disclose enough information about their recommendation systems.

Similarly, platforms do not disclose enough information to assess every element of the Rabat Plan of Action (see the *International human rights standards* chapter). Moderators may have access to contextual information unavailable to external researchers, including information about anonymous or pseudonymous users, their influence, previous conduct, or intent. Assessments in this report are therefore necessarily based on publicly available information. All these limitations affect conclusions about the prevalence of hate speech and the internal operation of moderation systems, but not the research's ability to assess the moderation outcomes directly observed.

1.3. Ethical and security considerations

All aspects of this research were subject to a human rights risk assessment to prevent and mitigate foreseeable harm. The research involved monitoring sensitive content in a conflict-affected environment where creators face arbitrary arrest and violence, and where targeted groups remain at risk of retraumatisation. Republishing hate speech can reinforce and spread discriminatory narratives. Even anonymised screenshots are easily identifiable. For these reasons, the report does not reproduce posts or identify the accounts that published them. This was unnecessary to the research objectives and would have created avoidable risk.

Where particularly dangerous posts were identified during monitoring, HRM reported them immediately rather than leaving them online for observation. Those posts were excluded from the dataset to avoid creating foreseeable human rights risks through the research itself.

2. International human rights standards

This research assesses "hate speech" through the framework of international human rights law rather than platform policies. That framework serves three purposes. First, "hate speech" is widely used but has no agreed definition.[5] Second, it provides a consistent basis for identifying actionable hate speech within the dataset. Third, where platforms state that their policies are informed by international human rights law, it provides an independent benchmark for evaluating whether the platform's moderation responses were proportionate to the human rights risks present in Myanmar.

International human rights law protects freedom of expression while also protecting the rights to equality and non-discrimination. The challenge is therefore not simply identifying harmful speech, but determining when expression crosses the threshold at which restrictions become justified or required. The following sections explain the standards applied throughout this report.

2.1. Equality and protected characteristics

Freedom of expression is protected by Article 19 of the International Covenant on Civil and Political Rights (ICCPR). That protection extends to expression that may be controversial, offensive, or disturbing. At the same time, Article 26 guarantees equality before the law and protection against discrimination. These rights are complementary and must be interpreted together.

For that reason, the "hate" in hate speech is not simply hatred as an emotion or language that offends. International human rights law is concerned with discriminatory expression directed against individuals or groups because of characteristics protected under international law, including race, ethnicity, religion, nationality, sex, disability, and other "protected characteristics".

This distinction is important. Criticism of a person's actions, beliefs, or conduct is generally protected, even when strongly expressed. By contrast, expression that attacks people because of who they are may justify or require restriction where it reaches a sufficient level of seriousness. Restrictions are permitted only where they satisfy the strict requirements of Article 19(3) that they must be lawful, pursue a legitimate aim, and be necessary and proportionate.

Where discriminatory expression reaches a sufficient level of seriousness, international law allows or requires restrictions. Article 19 allows restrictions where

[5] The term is subjective, meaning different things to different people. One person may regard an opinionated expression about a group as "hate speech" while another person believes it is legitimate criticism.

they are necessary to protect the rights of others. Article 20(2) requires States to prohibit advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence. International law therefore recognises three broad categories of expression:

1. Expression that remains protected (Article 19)
2. Expression that may be restricted (Article 19(3))
3. Expression that must be prohibited (Article 20(2)).

The main framework for distinguishing between these categories is the Rabat Plan.

2.2. The Rabat Plan of Action

The Rabat Plan of Action was developed by the United Nations to help determine when harmful expression reaches the threshold at which restrictions may be allowable or when it must be prohibited. Rather than relying on individual words or phrases, it requires expression to be assessed through six factors:

1. The social and political context in which the expression occurred
2. The speaker's position and influence
3. The speaker's intent
4. The content and form of the expression
5. The extent of its dissemination
6. The likelihood, including imminence, that it will contribute to harm.

The framework recognises that identical language may create very different risks depending on where, when, and by whom it is used. Context is therefore central to any human rights assessment of hate speech.

2.3. Context is key in the six-part test

The first factor in the Rabat Plan, social and political context, is particularly significant in Myanmar. Decades of armed conflict, discrimination, and serious human rights violations mean that discriminatory expression cannot be assessed in isolation. Identity-based abuse, dehumanisation, collective criminalisation, and fabricated narratives may each contribute to a broader environment of hostility even where they do not explicitly advocate violence.

The Independent Investigative Mechanism for Myanmar's (IIMM) 2024 report concluded that coordinated online hate campaigns formed part of the wider

environment preceding and accompanying the military's 2017 atrocities against the Rohingya.[6] Since the 2021 military coup, continuing conflict and widespread human rights violations have further increased the importance of context when assessing foreseeable risks arising from online expression.

Recognising this context does not reduce protection for freedom of expression. Political criticism, disagreement, and offensive speech remain protected unless the requirements for restriction under Article 19(3) are met. Rather, the Rabat Plan requires context to be taken seriously when determining whether discriminatory expression is likely to contribute to discrimination, hostility, or violence.

For platforms operating in Myanmar, this means that effective moderation cannot depend solely on identifying explicit threats or violent language. It also requires understanding how discriminatory narratives function within a society that has already experienced atrocity crimes and continuing armed conflict.

2.4. Platform responsibilities under the UN Guiding Principles

While the ICCPR establishes obligations for States, the UN Guiding Principles on Business and Human Rights (UNGPs) establish the responsibility of businesses to respect human rights.

Under the UNGPs, companies should avoid causing or contributing to adverse human rights impacts, undertake human rights due diligence to identify and mitigate foreseeable risks, and monitor whether their responses are effective. The UNGPs also recognise that businesses operating in conflict-affected environments should apply heightened human rights due diligence because the risks of severe harm are greater.

For social media platforms, these responsibilities extend beyond publishing content policies. They require governance systems capable of identifying context-specific risks, adapting moderation accordingly, and demonstrating that those systems effectively reduce foreseeable human rights harms.

This is especially relevant in Myanmar. Following criticism of Facebook's role during the 2017 atrocities against the Rohingya, the company acknowledged that it had not done enough to prevent its platform from contributing to violence and committed itself to strengthening its governance, moderation systems, and human rights due diligence.

[6] The IIMM concluded after reviewing 43 Facebook pages and 10,000 posts published between July and December 2017 that coordinated online hate campaigns formed part of a broader strategy to dehumanise the Rohingya, spread false narratives, and encourage hostility before and during the military's 2017 crackdown. The military's so-called "clearance operations" in 2017 involved many human rights violations and are currently the subject of ongoing proceedings before the International Court of Justice, the International Criminal Court, and other accountability mechanisms.

Throughout this report, the UNGPs provide the framework for assessing not only individual moderation decisions, but whether overall governance systems appear capable of responding appropriately to Myanmar's heightened human rights risks.

2.5. Platform rules and international standards

Most social media platforms regulate content through their own policies, commonly called Community Standards, Community Guidelines, or Terms of Service. These are private rules rather than legal obligations, although some companies state that they are informed by international human rights law.

Facebook, which is the focus of this report, has Community Standards that prohibit "hateful conduct" directed at people based on protected characteristics and state that they are informed by international human rights standards, including the Rabat Plan of Action. In practice, Facebook's rules are broader than international law and permit the company to remove some content that would remain protected under freedom of expression. Nevertheless, because Facebook publicly states that its moderation framework is informed by the same international standards applied in this research, those standards provide an appropriate benchmark against which to evaluate the platform's moderation decisions.

Moderation decisions are made through a combination of automated systems, human reviewers, and internal appeals processes rather than independent courts. Assessing whether those systems operate consistently and effectively is therefore as important as examining the content policies themselves.

2.6. Applying the standards in research

HRM used the framework described in this chapter to identify "actionable hate speech". The starting point throughout the research was that freedom of expression should be protected. Offensive, controversial, or discriminatory expression was not automatically treated as actionable. A post was included in the dataset only where, applying the Rabat Plan of Action and giving particular weight to Myanmar's social and political context, HRM concluded that it was likely to meet the threshold at which restriction would be justified or required under international human rights law Articles 19 or 20(2).

These assessments represent expert human rights analysis rather than judicial determinations. Applying a consistent methodology across the dataset nevertheless provided an objective benchmark against which Facebook's moderation decisions could be evaluated.

Because Facebook states that its Community Standards are informed by international human rights law, posts meeting this higher threshold would also generally be expected to violate Facebook's own rules. The findings in the following chapters therefore assess whether Facebook's moderation systems responded proportionately to the foreseeable human rights risks presented by those posts.

3. Patterns of hate speech in Myanmar

This chapter examines the characteristics of the actionable hate speech identified through the research. It outlines what forms actionable hate speech takes and which groups are most frequently targeted. Together, these findings describe the online environment in which Facebook's moderation systems were operating. They also provide important context for assessing whether those systems responded appropriately to foreseeable human rights risks.

3.1. Multiple forms of hate speech

The research identified 232 Facebook posts containing actionable hate speech. Each post was categorised according to six forms of discriminatory expression. Because many posts contained more than one form, individual posts could be assigned to multiple categories.

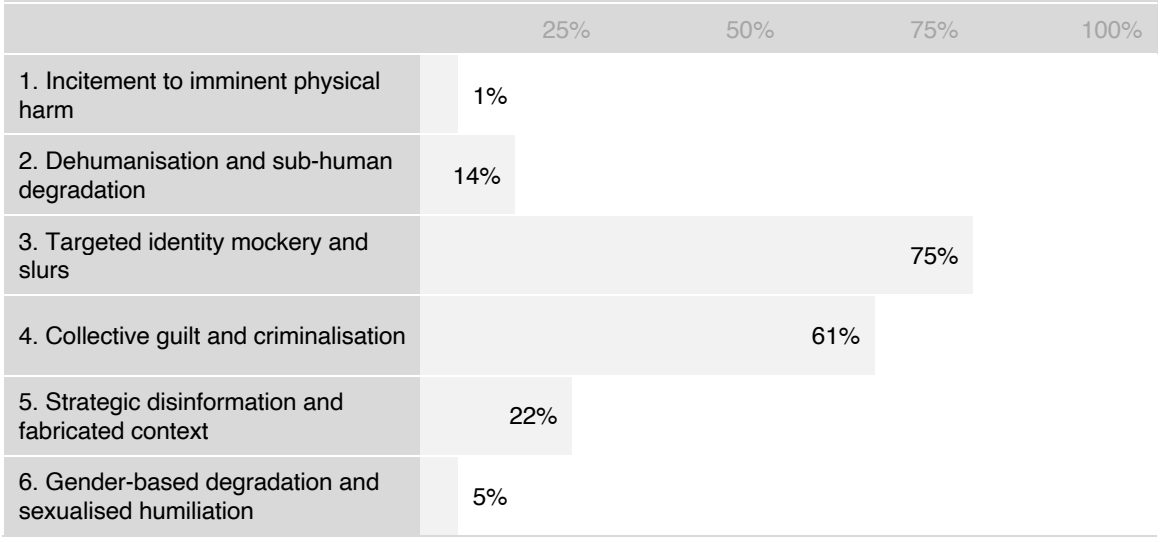
The categories were developed during the monitoring process to reflect the principal forms of actionable hate speech observed in Myanmar. Although created for analytical purposes, they are grounded in international human rights standards, particularly the Rabat Plan of Action and the principles of equality and non-discrimination.

Categories of actionable hate speech		
Category	Source	Diagnostic question
1. Incitement to imminent physical harm	Grounded in the Rabat Plan, likeliness and imminence of physical harm	Does the post explicitly or implicitly call for, justify, celebrate, or demand physical violence, death, arrests, or military/armed crackdowns against an individual or group with a protected characteristic?
2. Dehumanisation and sub-human degradation	Grounded in international humanitarian law, stripping a group of human rights protections	Does the post reduce an individual or group with a protected characteristic to animals, insects, diseases, filth/garbage, or inanimate objects?

3. Targeted identity mockery and slurs	Grounded in the right to equality and non-discrimination	Does the post use derogatory slurs, hostile stereotypes, or systematic mockery targeting a specific protected characteristic (ethnicity, religion, disability, gender)?
4. Collective guilt and criminalisation	Grounded in the legal principle of individual criminal responsibility	Does the post sweepingly accuse an entire civilian population, demographic group, or political identity of being inherent criminals, terrorists, or security threats?
5. Strategic disinformation and fabricated context	Grounded in Rabat Plan, context, speaker, and intent to mislead to cause public hostility	Does the post rely on a completely fabricated event, fake quote, or media stripped of its original context specifically to generate panic or hatred?
6. Gender-based degradation and sexualised humiliation	Grounded in the Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW)	Does the post weaponise sexualised insults, vulgarity, or patriarchal norms to humiliate women or LGBTQ+ individuals, often to strip them of their political voice?

Targeted identity mockery and slurs were by far the most common category, followed by collective guilt and criminalisation. By contrast, explicit incitement to imminent physical harm was extremely rare.[7]

Share of posts containing each category of actionable hate speech[8]



[7] As covered in the methodology, posts containing the most severe forms of hate speech were immediately reported to the platform and not included in this research. They all fell into the category of incitement to imminent physical harm. If the posts were included, the category order based on share would remain the same.

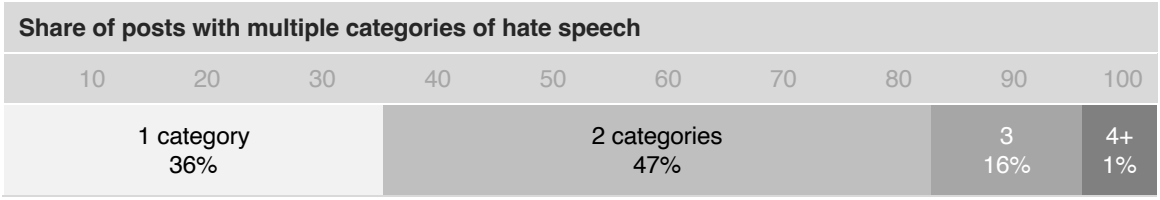
[8] As covered in the methodology, posts containing the most severe forms of hate speech were immediately reported to the platform and not included in this research. They all fell into the category of incitement to imminent physical harm. If the posts were included, the category order based on share would remain the same.

The findings demonstrate that actionable hate speech in Myanmar rarely consists of explicit calls for violence. Instead, it is primarily expressed through discriminatory narratives that attack protected groups by ridiculing identity, attributing collective criminality, dehumanising communities, or spreading fabricated information designed to increase hostility.

This has important implications for content moderation. Systems focused primarily on identifying explicit threats or violent language are likely to miss much of the actionable hate speech circulating online. Effective moderation therefore depends on understanding context rather than individual words alone.

3.2. Hate speech rarely takes only one form

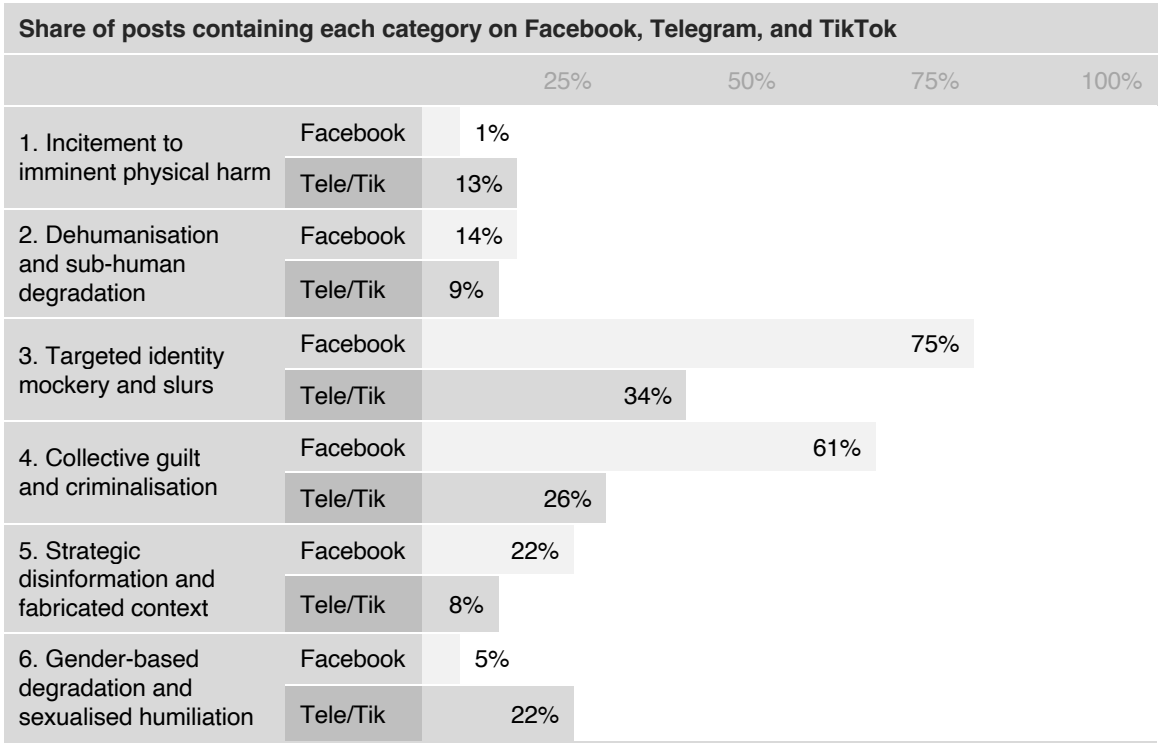
Actionable hate speech rarely appeared in isolation. Almost two-thirds of posts combined several forms of discriminatory expression within. A single post might ridicule a protected group's identity, accuse the group collectively of criminality, dehumanise its members, and reinforce those claims through fabricated information.



This finding reinforces one of the central principles of the Rabat Plan of Action, that harmful expression should be assessed in context rather than sentence by sentence. Different forms of hostility may reinforce one another, increasing the overall severity of a post even where no single statement explicitly advocates violence. The findings therefore support moderation systems that evaluate posts holistically rather than relying solely on keywords or isolated phrases.

3.3. Differences between social media platforms

The characteristics of actionable hate speech differed considerably between Facebook and the other platforms included in the research. Facebook posts were substantially more likely to contain identity mockery, collective criminalisation, and fabricated narratives. By contrast, the posts identified on Telegram and TikTok more frequently contained explicit incitement to physical harm and gender-based degradation.



The research was not designed to measure the prevalence of hate speech across different platforms, and these differences should therefore be interpreted cautiously. Nevertheless, the findings suggest that different platforms foster different forms of discriminatory expression. These differences may reflect variations in user communities, recommendation systems, moderation practices, or the ways users choose to communicate on each platform.

The findings also suggest that moderation strategies influence the character of harmful content. Platform design and governance therefore shape not only how hate speech is moderated, but also how it develops.

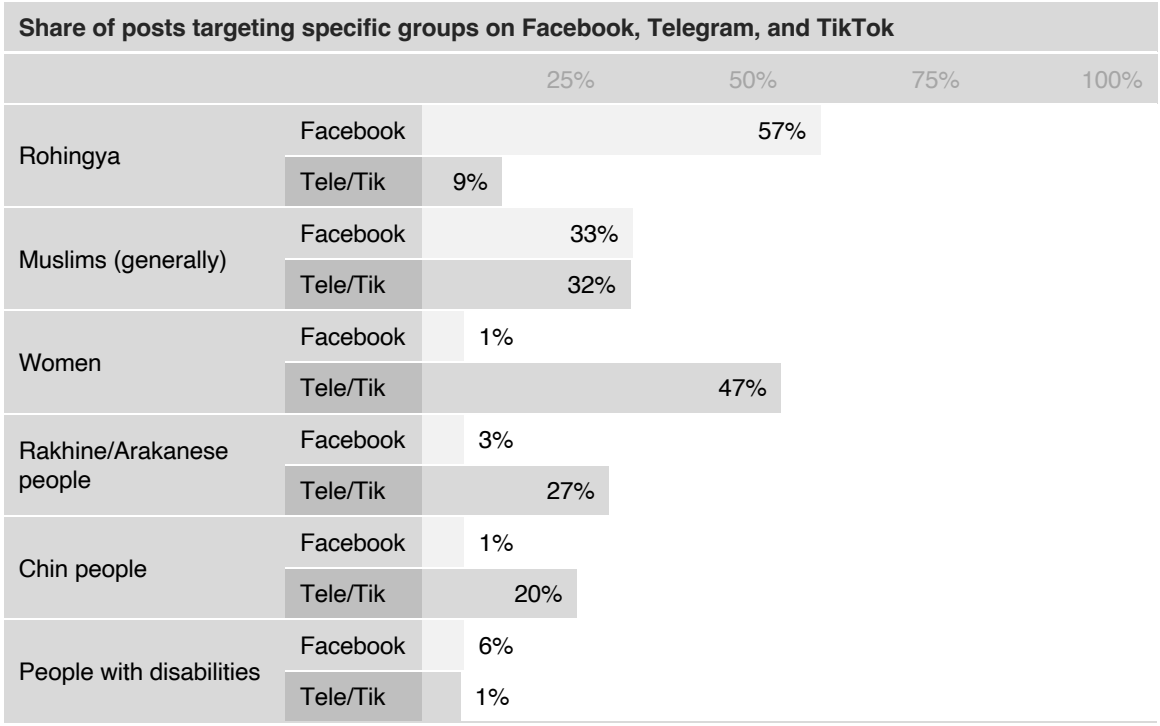
3.4. Rohingya remain the primary target

The Rohingya remained by far the most frequently targeted protected group on Facebook. More than half of all Facebook posts identified in the research targeted the Rohingya, while one-third targeted Muslims more generally. Many posts targeted both groups simultaneously.

The continued prominence of anti-Rohingya hate speech is particularly significant given Myanmar's recent history. The Independent Investigative Mechanism for Myanmar concluded that Facebook played an important role in spreading hatred before and during the 2017 atrocities against the Rohingya. Although this research

does not identify a comparable coordinated campaign or establish any causal link between individual posts and offline violence, it demonstrates that discriminatory narratives directed at the Rohingya continue to circulate nearly a decade later.

Other protected groups targeted on Facebook included people with disabilities, Rakhine/Arakanese people, Chin people, Christians, immigrants, and women.



The research was not designed to compare prevalence across platforms, but the distribution differed markedly elsewhere. While Facebook remained dominated by anti-Rohingya and anti-Muslim hate speech, Telegram and TikTok contained substantially higher proportions of content targeting women, Rakhine/Arakanese people, and Chin people.

These findings suggest that discrimination in Myanmar continues to be shaped by both long-standing patterns of ethnic and religious persecution and newer forms of hostility directed towards other protected groups. Different platforms therefore appear to present different human rights risks.

3.5. Human rights implications

The findings show that actionable hate speech remains a persistent feature of Myanmar's online environment. Although explicit calls for violence were

uncommon, discriminatory expression directed at protected groups remained widespread. Most of the identified posts relied instead on identity-based abuse, collective criminalisation, dehumanisation, and fabricated narratives that contribute to hostile online environments.

The findings also reinforce the importance of context. Myanmar's history of armed conflict, systematic discrimination, the atrocities committed against the Rohingya, and continuing human rights violations means that discriminatory narratives cannot be viewed in isolation. Their cumulative effect may reinforce prejudice, exclusion, hostility, and fear even where individual posts stop short of directly advocating violence.

The persistence of anti-Rohingya hate speech is particularly concerning because it demonstrates that the narratives associated with the 2017 atrocities remain active almost a decade later. At the same time, the research identified discriminatory expression directed at a wider range of protected groups, demonstrating that online discrimination in Myanmar extends well beyond any single community.

Taken together, these findings describe the environment that Facebook's moderation systems were expected to manage. The next chapter therefore examines whether the platform's moderation systems identified and responded effectively to this content.

4. Testing Facebook's content moderation

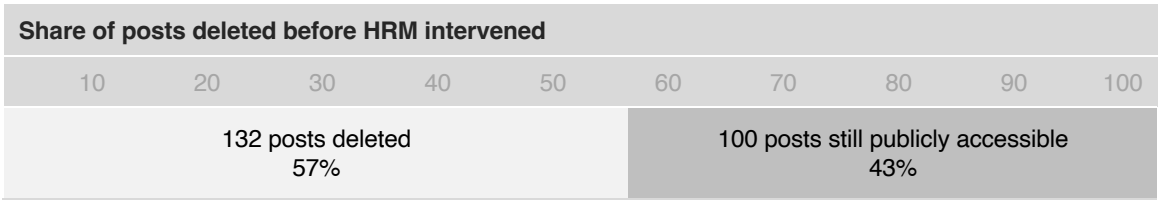
This chapter assesses how Facebook responded to the actionable hate speech identified through the research. The objective is not to determine whether individual moderation decisions were correct. Rather, it evaluates whether Facebook's moderation systems, including proactive moderation, user reporting, and appeals, appeared capable of identifying and responding to content that HRM assessed as actionable hate speech under the framework set out in the *International human rights standards* chapter.

4.1. Facebook's systems deleted half of actionable hate speech

Before any reports were submitted by HRM, the 232 Facebook posts identified through the research were observed for one month. This allowed Facebook's existing moderation systems, including proactive automated moderation, moderation resulting from reports by other users, and voluntary deletion by creators, to operate without intervention from HRM.

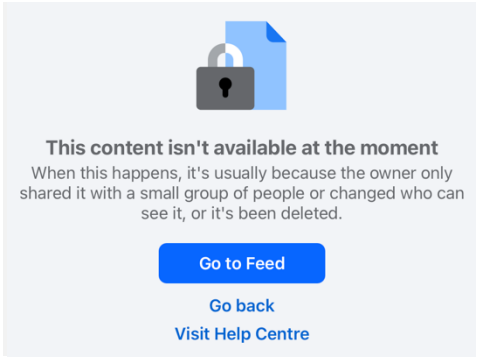
Facebook's moderation infrastructure	
Proactive automated moderation	Facebook states that it invests in automated systems designed to identify potentially harmful content proactively. These systems use artificial intelligence (AI) to analyse text, images, and videos for indicators that content may violate Facebook's Community Standards.
Reactive user-reported moderation	Facebook also relies on a reactive mechanism in which users to report content they believe violates the Community Standards. User reports are assessed through automated systems, human reviewers, or a combination of both. Where Facebook concludes that a post violates its policies, it is deleted.

After one month, over half of the 232 posts were no longer publicly accessible, while just under half remained online.



Facebook displays the same generic "tombstone" notice whenever content becomes unavailable, making it impossible to determine why individual posts disappeared.[9] They may have been removed by Facebook's automated systems, following reports from other users, or deleted voluntarily by the creator.

Nevertheless, the findings show that Facebook's existing moderation systems removed a substantial proportion of actionable hate speech before HRM intervened.



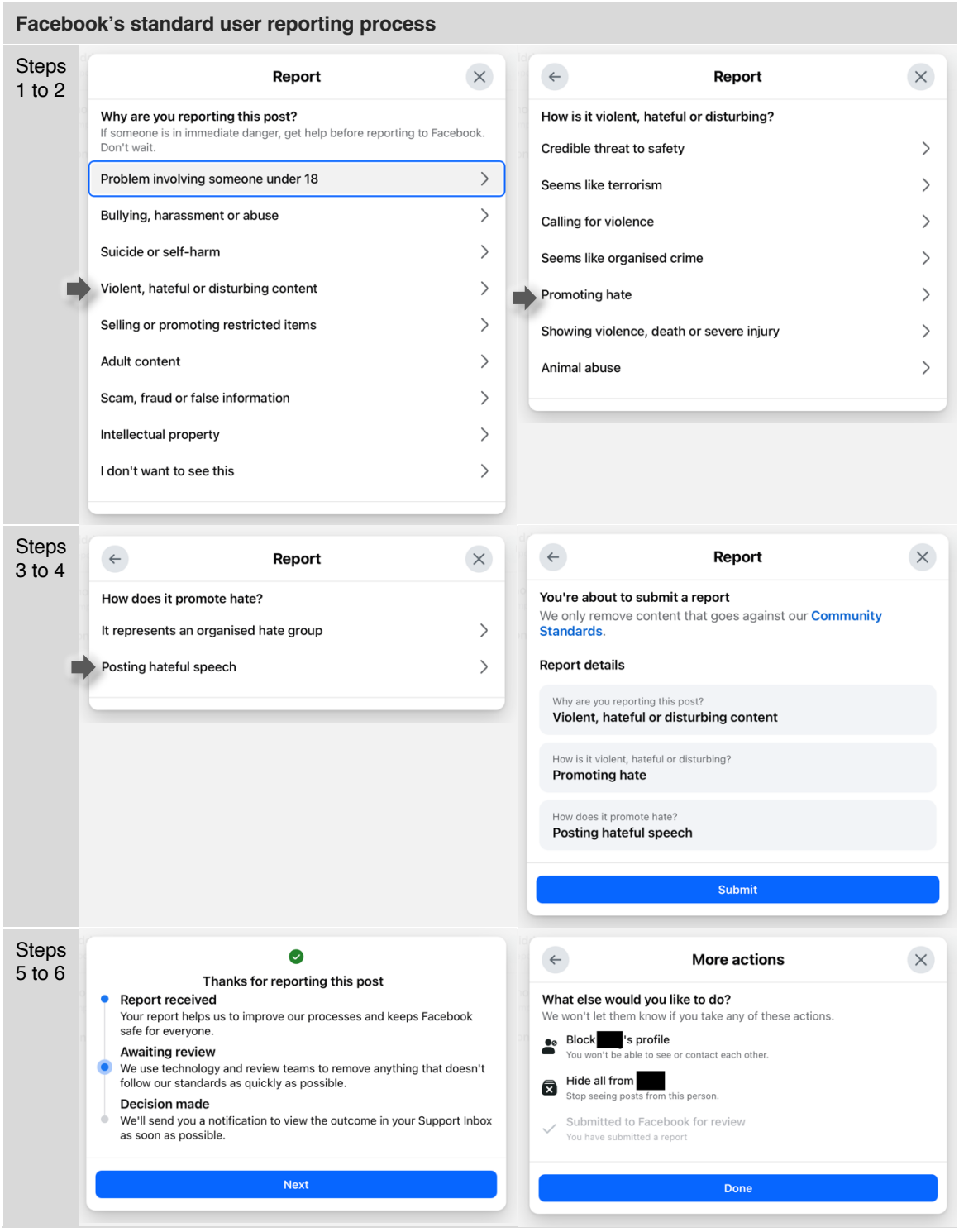
At the same time, the fact that more than four in ten posts remained publicly accessible for at least a month demonstrates that Facebook's existing moderation systems did not identify all actionable hate speech. Those remaining posts formed the basis for testing Facebook's reporting and appeals mechanisms.

4.2. User reporting had little additional effect

HRM reported all 100 posts that remained publicly accessible after the one-month observation period using Facebook's standard reporting mechanism. Reports were

[9] These notices are called "tombstones" is software development. The posts were monitored several times over the project period and they were never restored.

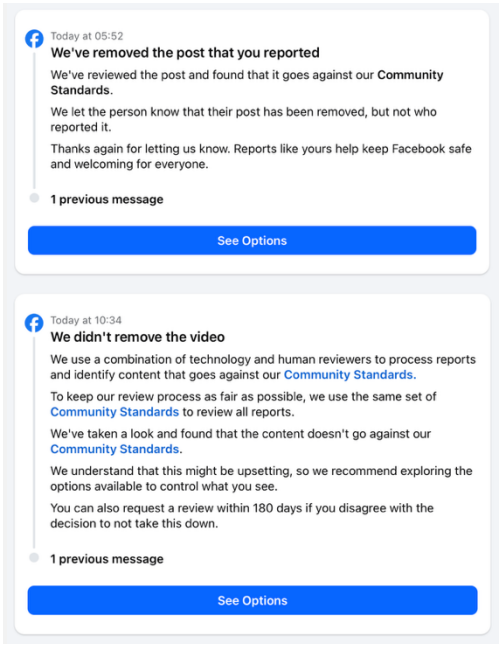
submitted from within Myanmar using multiple ordinary user accounts and private browser sessions with no identifiable link to HRM. All reports were submitted under Facebook's category of "posting hateful speech".



Within 24 hours, Facebook had processed 26% of reports, although none resulted in the deletion of content. Within 48 hours, Facebook had processed 32% of reports. Only three reports (3%) were accepted and the posts deleted.

After one month, Facebook had processed a total of 92% of reports. The remaining 8% continued to appear as pending within Facebook's reporting interface and never received a decision. Overall, Facebook accepted 3% of HRM's reports, rejected 89%, and failed to process the remaining 8%.

These findings suggest that Facebook's user reporting mechanism added very little once posts had remained online for an extended period. Despite HRM identifying each post as actionable hate speech that would generally also be expected to violate Facebook's Community Standards, almost every report was rejected. The findings therefore indicate a substantial gap between the harmful content identified through this research and Facebook's moderation outcomes.

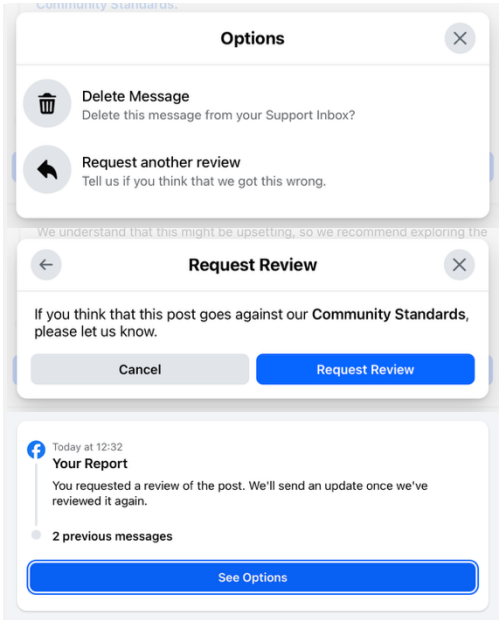


4.3. Appeals provided only a limited safeguard

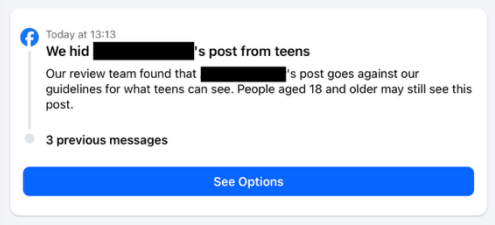
HRM appealed every report that Facebook rejected. Within 24 hours, Facebook had processed 32% of appeals. No additional appeals were processed during the following day.

After one month, Facebook had processed 91% of appeals. The remaining 9% continued to appear as pending without receiving any decision. Overall, Facebook accepted 15% of appeals, rejected 76%, and left 9% unanswered. Appeals therefore resulted in the removal or restriction of a small number of additional posts but altered relatively few moderation decisions overall.

The findings suggest that Facebook's appeals process functioned as a limited safeguard rather than a robust corrective mechanism. Once Facebook rejected an initial report, that decision was rarely overturned.



4.4. Overall moderation effectiveness



The moderation process can be understood as three successive stages. Before HRM intervened, Facebook's existing moderation systems deleted over half of the 232 actionable hate speech posts identified through the research.

HRM's reports resulted in the deletion of a very small number of posts, while appeals deleted a few more. Overall, two-thirds of the original posts had been deleted by the end of the research, while a third remained publicly accessible.

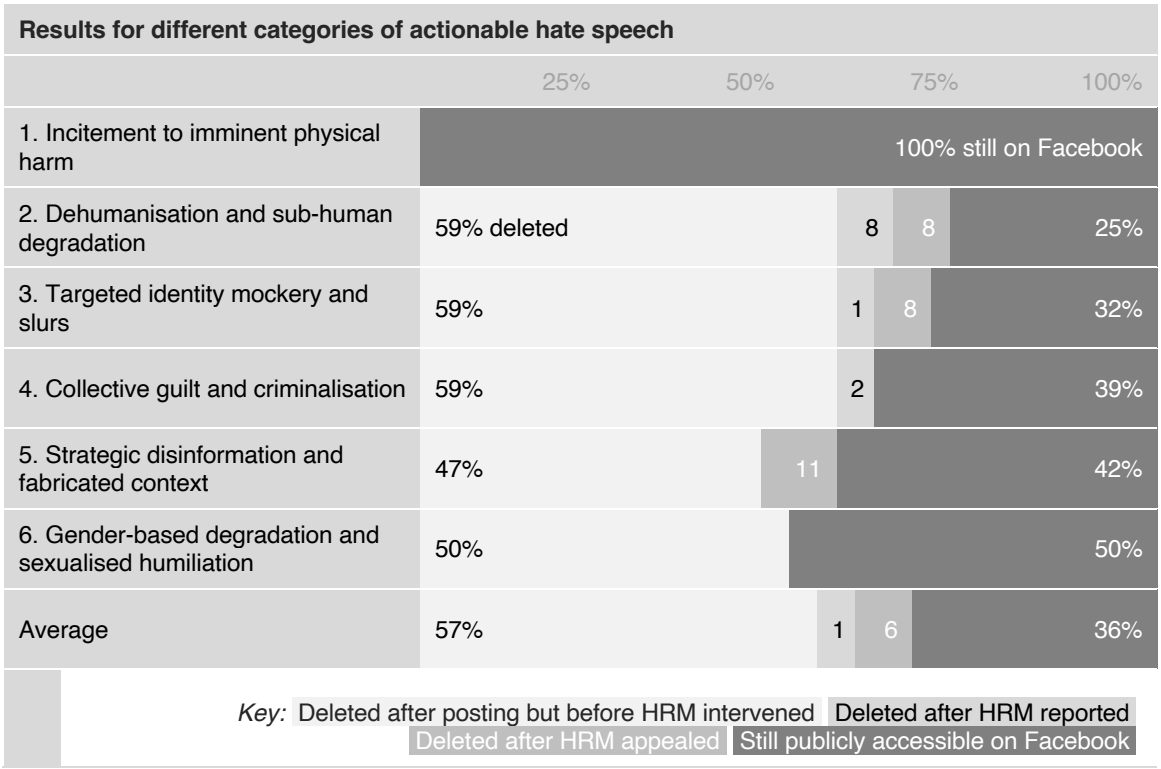
Facebook moderation result flowchart		
	25% 50% 75% 100%	
Initial identification	232 actionable hate speech posts on Facebook	
After one month	100 posts still on Facebook (43%)	132 posts deleted (57%)
After HRM first-instance reports	97 posts still on Facebook (42%)	140 posts deleted (58%)
After HRM appeals	84 posts still on Facebook (36%)	148 posts deleted (64%)

These findings show that almost all successful moderation occurred before HRM intervened. Once actionable hate speech remained online for a month, Facebook's reporting and appeals mechanisms produced relatively few additional deletions.

This suggests that Facebook's moderation operates in layers. Existing moderation systems identify a substantial proportion of harmful content, but posts that pass through those systems appear considerably less likely to be removed through subsequent user reports or appeals. From the perspective of users seeking to report harmful content, the practical value of Facebook's reporting mechanisms therefore appeared limited.

4.5. Moderation was inconsistent across different forms of hate speech

The research also examined whether Facebook responded differently to different categories of actionable hate speech.



The results should be interpreted cautiously because some categories contained relatively few posts. Nevertheless, one finding is clear. No category of actionable hate speech achieved consistently high moderation rates after HRM reported the content.

Acceptance rates remained low across almost every category. Very few first-instance reports resulted in deletion, and appeals produced relatively few additional deletions. Perhaps the most notable finding is not that Facebook performed better or worse in relation to particular forms of hate speech, but that no clear moderation pattern emerged at all.

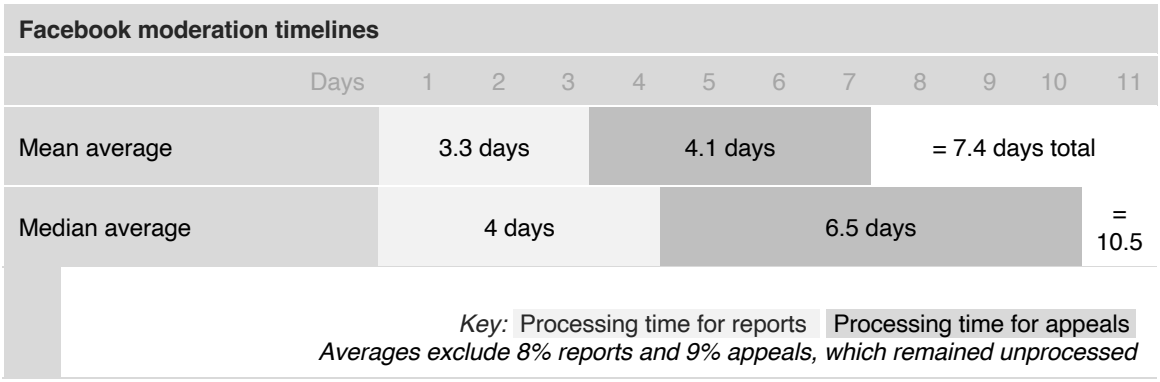
The findings therefore suggest that the effectiveness of Facebook's reporting and appeals mechanisms was systemic rather than confined to particular forms of discriminatory expression.

4.6. Long response times for moderation

The research also examined how long Facebook took to process reports and appeals. The mean average processing time was 3.3 days for first-instance reports and 4.1 days for appeals, giving a combined average of 7.4 days. Because a small number of decisions were made within one hour, median processing times provide a more

representative measure of the research experience. The median processing time was 4 days for reports and 6.5 days for appeals, giving a combined median of 10.5 days.

Accepted reports and appeals were processed slightly more quickly, with a combined median of six days. However, the number of successful reports was too small to support reliable conclusions.



These processing times have particular significance in the context of social media. Posts remained publicly accessible throughout the review process and could continue to be viewed, shared, and recommended by Facebook's systems. Because most engagement occurs shortly after publication, posts deleted after a week or more are likely already to have reached most of their potential audience.

In high-risk environments such as Myanmar, moderation that occurs only after most engagement has already taken place is unlikely to prevent much of the foreseeable harm associated with discriminatory content.

4.7. Overall assessment

Taken together, the findings present a mixed picture of Facebook's moderation performance. Facebook's existing moderation systems removed a substantial proportion of actionable hate speech before HRM intervened, demonstrating that the platform retains a significant moderation capability.

However, more than one-third of the actionable hate speech identified through this research remained publicly accessible after observation, reporting, and appeals. Once content had escaped Facebook's initial moderation systems, user reporting and appeals rarely resulted in its removal.

The findings do not establish why apparently similar posts received different moderation outcomes. Nor do they identify whether moderation decisions were made by automated systems, human reviewers, outsourced moderators, or other

internal processes. Nevertheless, they raise important questions about the effectiveness, consistency, and transparency of Facebook's moderation systems in Myanmar.

The next chapter therefore considers what these findings suggest about Facebook's broader human rights governance and whether the company demonstrates the level of due diligence expected under the UN Guiding Principles on Business and Human Rights.

5. Facebook's human rights governance

The previous chapter examined how Facebook responded to posts identified through this research as actionable hate speech. This chapter considers what those findings suggest about Facebook's broader governance systems and human rights due diligence. It does not seek to determine how Facebook's internal moderation systems operate or attribute individual decisions to particular technologies or reviewers. Rather, it draws evidence-based inferences from observable moderation outcomes and considers their implications under the UNGPs.

5.1. Responses may suggest backlog-driven auto-rejection

One of the most striking findings concerned the timing of Facebook's moderation decisions. Many of HRM's reports and appeals were rejected after exactly the same interval from submission. Most reports were rejected after precisely 96 or 120 hours (four and five days respectively), while most appeals were rejected exactly 156 hours (6.5 days) after submission. This pattern occurred repeatedly across more than 100 submissions. By contrast, accepted reports and appeals, together with some rejected cases, were processed at a range of different times and showed no comparable pattern.

<table>
<tr><th colspan="5">Facebook moderation timelines</th></tr>
<tr><th></th><th>25%</th><th>50%</th><th>75%</th><th>100%</th></tr>
<tr><td>Reports</td><td colspan="2">Rejected at exactly 96 hours (4 days)</td><td colspan="2">Rejected exactly 120 hrs (5 days)
Random acceptance/ rejection decisions</td></tr>
<tr><td>Appeals</td><td colspan="2">Rejected at exactly 156 hours (6.5 days)</td><td colspan="2">Random acceptance/ rejection decisions</td></tr>
</table>

The research cannot determine why these identical rejection times occurred. Possible explanations include batch processing, queue-based workflows, internal service-level

targets, or other operational procedures. However, the repeated timing pattern raises legitimate questions about whether some reports were rejected through administrative workflow rather than substantive assessment.

One possible explanation is that many reports reached a predetermined operational deadline and were automatically rejected without either human review or meaningful automated assessment. This may be due to a lack of employees or compute time, with the intention of stopping the backlog queue from getting too large. The research cannot establish that this occurred, but neither can it be ruled out from the available evidence.

If moderation outcomes are influenced by operational capacity rather than substantive assessment, this would raise important questions about the effectiveness of Facebook's reporting and appeals mechanisms. Under the UNGPs, grievance mechanisms should be capable of identifying and addressing adverse human rights impacts. Facebook should therefore explain how reports and appeals are prioritised, reviewed, and closed, particularly where repeated timing patterns emerge. Greater transparency would enable independent assessment of whether these patterns reflect ordinary workflow or shortcomings in the moderation system.

5.2. Human rights due diligence extends beyond removing posts

The findings also highlight a broader point about platform governance. Under the UNGPs, companies are expected not only to respond to individual instances of harm but also to identify, prevent, and mitigate foreseeable human rights risks arising from their services. For social media platforms, that responsibility extends beyond deciding whether individual posts should remain online.

Platforms have a range of tools that may reduce harm without deleting content entirely, including limiting recommendations, reducing distribution, restricting monetisation, and excluding content from recommendation systems. Facebook has publicly stated that it uses such measures in some circumstances.

This research did not assess whether those interventions were applied. However, accounts created specifically for the research, with no prior browsing history or social connections, quickly began receiving recommendations for discriminatory content after searching for material relating to protected characteristics. Although this observation is not sufficient to evaluate Facebook's recommendation systems systematically, it suggests that users seeking such content may still be directed towards substantial quantities of discriminatory material.

Whether or not individual posts should ultimately have been deleted, the cumulative effect observed during the research was the continued circulation of discriminatory

narratives directed at protected groups. In line with the Rabat Plan's emphasis on context and the UNGPs' requirement to mitigate foreseeable human rights risks, reducing the amplification and accumulation of discriminatory content may be as important as removing individual posts, particularly in conflict-affected settings such as Myanmar.

5.3. Commercial relationships create governance challenges

The research also identified examples of accounts publishing actionable hate speech that displayed Facebook verification badges. It was not possible to determine whether these badges reflected paid verification subscriptions or free verification granted to notable public figures. Nor was it possible to determine whether verification affected moderation decisions.

Similarly, advertising and promoted content appeared alongside posts containing actionable hate speech. The research could not determine how advertisements were selected or whether creators received any share of advertising revenue associated with those posts.

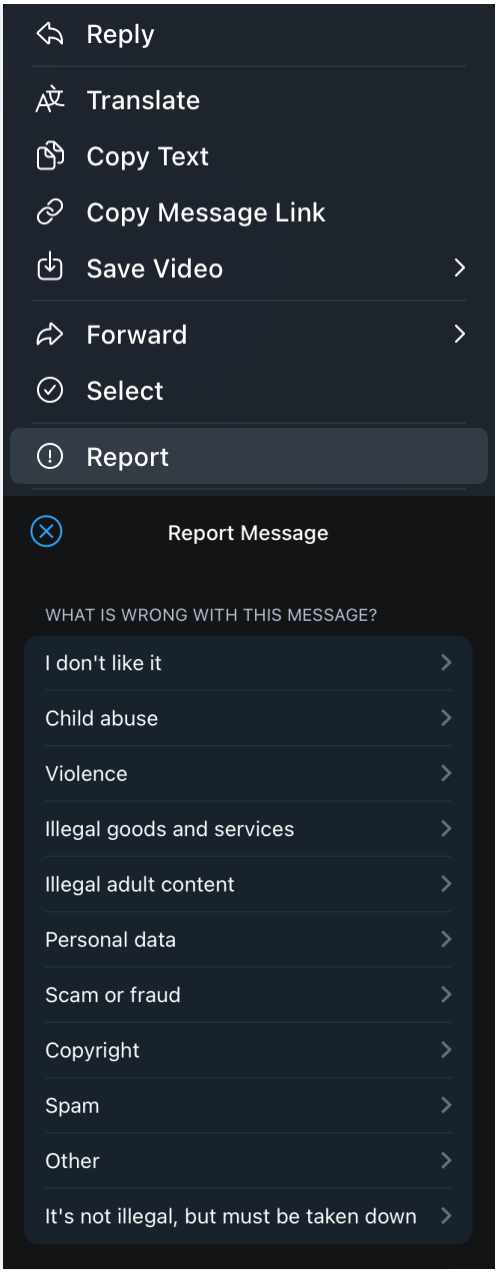
These observations do not establish that Facebook intentionally profits from hate speech or that commercial relationships influence moderation outcomes. They do, however, illustrate an inherent governance challenge. Social media companies simultaneously maintain commercial relationships with creators and advertisers while also acting as regulators responsible for enforcing their own rules.

The UNGPs require businesses to identify and address situations in which commercial incentives may contribute to foreseeable human rights risks. For that reason, platforms should be able to demonstrate that commercial interests do not undermine the impartial enforcement of their moderation systems. Greater transparency about verification, monetisation, and enforcement would help provide that assurance.

5.4. Transparency remains insufficient for independent oversight

Compared with some other platforms, Facebook publishes relatively extensive information about its moderation systems, including its Community Standards, Community Standards Enforcement Reports, prevalence estimates, and global statistics on proactive detection, user reports, and appeals. This provides a stronger basis for independent scrutiny than is available for many competing platforms.

For instance, although Telegram included a user reporting mechanism, its rules were unclear, and the reporting process was basic and opaque. Telegram provided no information about whether reports had been reviewed, whether decisions had been made, or why. Nor did it offer a comparable appeals process. As a result, HRM could not meaningfully evaluate Telegram's moderation performance using the same methodology.



Nevertheless, important gaps remain in Facebook's transparency. Enforcement data is reported globally rather than by country, making it impossible to assess moderation performance in high-risk contexts such as Myanmar. Although Facebook previously stated that its enforcement reporting would receive regular independent assessment, only one public assessment appears to have been published, in 2021. The company has also reduced the frequency of its reporting from quarterly to half-yearly without clearly explaining why.

These gaps matter because independent researchers can observe moderation outcomes but cannot determine why individual decisions were made. Throughout this research, it was often impossible to establish whether posts disappeared because of automated moderation, human review, user reports, voluntary deletion by creators, or other internal processes. Without this information, external stakeholders cannot meaningfully evaluate whether Facebook's moderation systems are functioning effectively where human rights risks are greatest.

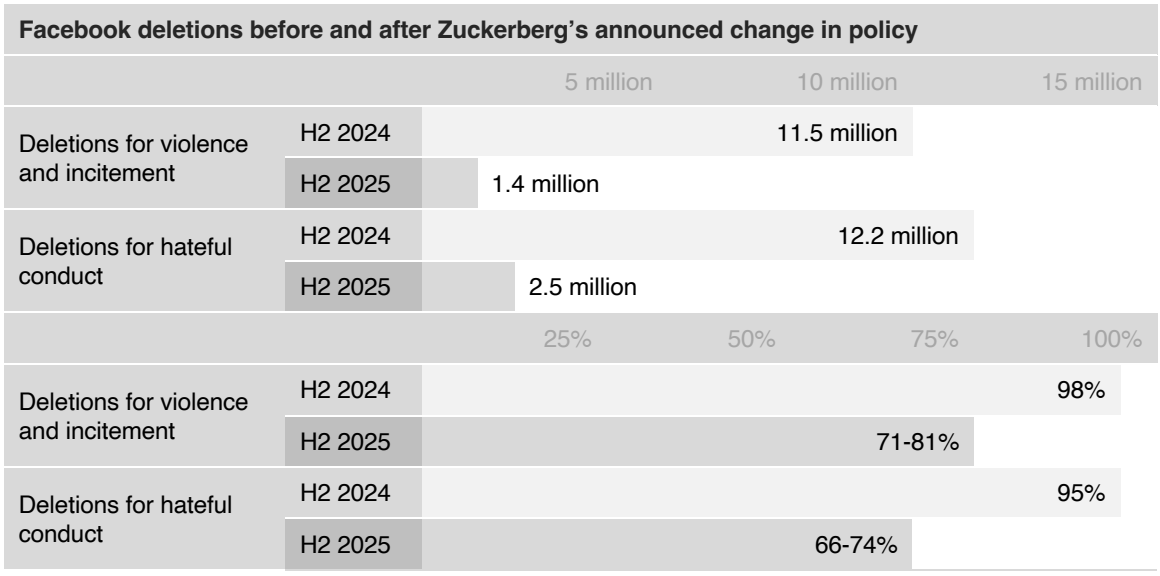
Under the UNGPs, companies should be able to demonstrate that they are identifying and addressing adverse human rights impacts. Greater transparency on country-specific enforcement, moderation workflows, quality assurance, outsourcing arrangements, and governance changes would substantially improve independent oversight and accountability.

5.5. Global policy changes may have reduced protection in Myanmar

In January 2025, immediately following U.S. President Trump's return to office, Facebook CEO Mark Zuckerberg announced that the company would adopt a "more speech and fewer mistakes" approach to moderation, reducing proactive enforcement and focusing automated systems on "illegal and high-severity violations".

This represented a significant change in Facebook's public approach to content moderation. Previously, the company had argued that expanding proactive automation improved its ability to identify harmful content at scale, including in Myanmar. The revised approach placed greater emphasis on avoiding mistaken removals, while accepting substantially lower levels of proactive enforcement.

Facebook's own transparency reports show a sharp global reduction in proactive moderation following the announcement. Overall removals for violence and incitement fell dramatically between the second half of 2024 and the second half of 2025, while removals for hateful conduct declined a similarly significant amount. The data also indicates that the platform became substantially more reliant on user reporting than on proactive automated detection.



HRM's findings are consistent with this broader trend. Throughout the research, substantial amounts of actionable hate speech remained publicly accessible despite Facebook's automated moderation, user reporting, and appeals systems. This research cannot establish that Facebook's January 2025 policy changes caused those outcomes in Myanmar. However, Facebook's own transparency reporting

demonstrates that they coincided with a substantial global reduction in proactive moderation.

Given Myanmar's well-documented history of online hate contributing to serious human rights violations, that relationship deserves particular scrutiny. Under the UNGPs, companies should assess the human rights implications of significant governance changes before implementing them and demonstrate how foreseeable risks will be mitigated in high-risk contexts. Facebook has not published sufficient information to allow independent assessment of whether such context-specific analysis was undertaken for Myanmar.

6. Conclusion

Nearly a decade after the IIMM documented Facebook's role in the atrocities against the Rohingya, this research shows that actionable hate speech, especially against the Rohingya, remains a persistent feature of Myanmar's online environment. Although Facebook deleted many of the harmful posts identified by HRM, more than one-third survived, remaining publicly accessible after observation, reporting, and appeals. The research raises serious questions about the transparency, consistency, and effectiveness of Facebook's moderation systems in one of the world's highest-risk human rights contexts.

This report does not conclude that Facebook has failed. It previously invested heavily in moderation infrastructure and provides more information about its systems than most competitors. But the UNGPs require more than evidence of effort. They require companies to show that their systems effectively identify, prevent, and mitigate foreseeable human rights harms. Facebook has not yet met that test.

The broader challenge extends beyond Facebook alone. As social media companies increasingly rely on global moderation systems, there is a risk that the specific needs of conflict-affected and high-risk countries receive less attention. Myanmar illustrates why moderation cannot be assessed solely through global averages or generic policies set in response to U.S. political changes. Human rights due diligence requires companies to understand how the same governance decisions may produce very different consequences in different contexts.

Myanmar has repeatedly demonstrated the consequences of allowing online hatred to circulate unchecked before contributing to wider human rights abuses. The findings in this report suggest that those lessons remain highly relevant.

6.1. Recommendations

To States

1. **Require meaningful transparency from platforms**: States should require large social media platforms to publish enough detail for independent assessment of moderation systems, including country-specific enforcement data, moderation workflows, the use of automated systems and human reviewers, and the human rights impacts of major governance changes, especially in high-risk contexts.
2. **Embed the UNGPs in digital regulation**: Laws should require companies to conduct ongoing human rights due diligence, including assessing the foreseeable human rights impacts of major moderation, governance, or algorithmic changes before implementation and mitigating identified risks.
3. **Support independent oversight and research**: States should protect researchers' and civil society organisations' ability to monitor online harms, access platform data where safeguards allow, and support evidence-based oversight while respecting human rights.

To social media companies

4. **Adopt context-specific human rights due diligence**: Moderation systems should reflect heightened risks in conflict-affected countries. Governance decisions, moderation thresholds, and enforcement systems should account for documented discrimination and violence against protected groups.
5. **Strengthen transparency and accountability**: Platforms should publish country-specific enforcement data for high-risk contexts, explain major moderation policy changes and their human rights impacts, disclose the roles of automated systems, human reviewers, and contractors, and resume regular independent assessments of transparency reporting.
6. **Demonstrate moderation effectiveness**: Platforms should regularly assess whether their moderation systems identify and mitigate foreseeable human rights risks in different country contexts. These assessments should cover content removals, recommendation systems, amplification, distribution, monetisation, and grievance mechanisms, and be transparent enough for independent scrutiny.
7. **Assess governance changes before implementation**: Before reducing moderation or significantly changing automated enforcement, platforms should conduct and publish human rights impact assessments explaining how foreseeable risks will be mitigated in countries where online discrimination has contributed to violence or other serious abuses.