Joan Mas Castella
501 Jelepang RD, Singapore

November 18th, 2024

Mark Zuckerberg

Meta

601 Willow Road, Menlo Park, California

Subject: Recommendation Report on Facebook's Role in the Rohingya Genocide and Content Moderation Algorithms

Dear Mr. Zuckerberg,

I am content in submitting the following attached recommendation report, '*A Recommendation Report on Facebook's Role in the Rohingya Genocide and Content Moderation Algorithms*'. This report includes a feasibility study regarding Facebook's (Meta's) role in amplifying violent rhetoric, which led to real-world violence through its misconduct of its content moderation algorithms. The findings and recommendations provided in this report are formulated through extensive research and analysis, which I believe should be looked into for Meta to uphold its responsibility to protect its users and the societal harm Facebook might have.

The analysis provided in the report focuses on the fact that Meta needs to reevaluate its content moderation system and structure. This should be done particularly in countries where conflicts are occurring, which it has been shown that Meta’s algorithms have amplified hate speech and incite violence. Meta should also consider refocusing the intentions behind their algorithms instead of focusing on driving engagement and making more money, considering what harm these algorithms could have on the users that use them.

I believe that you will find the report’s findings invaluable, and I am looking forward to actions Meta should take in response to this report. If you need further information and/ or clarification, please feel free to contact me.

Sincerely,

Joan Mas Castella

Researcher and Writer

Encl.

Facebook’s Role in the Rohingya Genocide and Content Moderation Algorithms: A Recommendation Report

Prepared for: Mark Zuckerberg

CEO

Meta

601 Willow Road, Menlo Park, California

Prepared by: Joan Mas Castella

Researcher and Writer

501 Jelepang RD, Singapore

January 18, 2024

Table of Contents

LIST OF ILLUSTRATIONS

Figures

Introduction

Social media platforms such as Facebook have a significant impact on public discourse and sentiments. This is a double-edged sword; it can also pose significant risks when it is misused. One of the most recent and well-documented occurrences of social media platforms, in this case Facebook, playing a role in amplifying harmful content, is during the 2017 Rohingya genocide in Myanmar. This report examines how Facebook's profit/engagement-driven algorithms contributed to the escalation of violence by promoting hate speech and inflammatory content, and what changes should be made to prevent this from occurring again. Despite internal and external warnings, Meta, Facebook's parent company, did not address or prevent the spread of anti-Rohingya rhetoric, which played a key role in the violence that unfolded (Amnesty International, 2022). The motivation behind this report arises from the growing impact that social media platforms and their profit-driven motives and algorithmic amplification can have disastrous consequences in the real world if left unregulated.

The purpose of this report is to evaluate Facebook's current content moderation practices and algorithms, focusing on how they are applied in sensitive geopolitical contexts like in Myanmar, and to offer recommendations for improving these systems. This will also include an ethical analysis of the responsibilities that social media platforms should hold when curbing hate speech, misinformation, and violence. I will address the following tasks:

- Examine how Facebook's algorithms contributed to the spread of hate speech.
- Analyze the ethical implications of Facebook's content moderation model, mainly in non-English speaking regions.

- Identifying potential improvements in Facebook's content moderation policies to better counteract harmful content.

The principal findings from the research conducted showed that Facebook's algorithms prioritize engagement at the expense of content quality, leading to the amplification of hate and harmful narratives. This was particularly detrimental in the case of Myanmar, where the spread of anti-Rohingya sentiment on Facebook had direct contributions to the violence that followed. Based on these findings, I recommend Facebook (Meta) to reassess its content moderation practices, mainly in adjusting its existing systems to adapt to cultural and language differences in different regions and not prioritize engagement at the expense of the safety of its users.

In the following sections, I will provide the research methods used, the results I obtained, and the conclusions and specific recommendations for action. This should provide Meta with a clear idea of which practices should be improved to mitigate further occurrences of its involvement in Myanmar.

Research Methods

This report is based on primary and secondary sources to identify and understand the role of Facebook's algorithms played in amplifying hate speech and contributing to violence in the 2017 Rohingya genocide in Myanmar. The research methods that were employed are to gather data from multiple perspectives, including social media analysis, ethical considerations, and content moderation practices, in order to value the implications of Facebook's content moderation algorithms and provide recommendations for improvement.

The research began with reviewing academic sources, including reports by Amnesty International (2022) and scholarly articles written on the topic, such as Khennour (2023), which explores the impact Facebook had on ethnic violence. Khennour's report provided a foundational understanding of how Facebook's engagement-driven algorithms amplify harmful content. These secondary sources provided detailed evidence of Facebook's complicity in the genocide, in its actions of spreading anti-Rohingya content. Additionally, I reviewed reports from the Ifo Institute (2024), which had a different perspective; it analyzed Facebook's broader role in ethnic conflicts and its implications in global social media practices (Ifo Institute, 2024). These sources provided the base knowledge needed for understanding how Facebook's algorithms operate, specifically in conflict zones and how they can be weaponized.

Furthermore, I also reviewed Meta's own internal documents, including the Facebook Papers (2021), which highlighted how content moderation was insufficient in addressing the spread of harmful content in regions like Myanmar. These documents illustrated the pitfalls that Meta's content moderation efforts faced. This was especially present in non-English-speaking regions, where local dynamics and languages were not fully considered in the platform's automated systems. The research also included a comparative analysis on how content moderation practices differ on other platforms, helping to contextualize Meta's failures.

Specifically, the research tasks were as follows:

- Task 1: Investigate the role of Facebook's algorithms in amplifying hate speech and inflammatory content, focusing on the Myanmar case study.
- Task 2: Evaluate the ethical implications of Facebook's content moderation model, particularly in conflict zones where social media can contribute to real-world violence.
- Task 3: Assess potential improvements in Facebook's content moderation policies, with a focus on algorithmic changes and localized moderation to better address harmful content.

In the following sections, I will present my findings and conclusions drawn from these findings, along with specific recommendations for Meta's future actions to address these issues directly.

Results

In this section, I will cover my key findings from my research, where I examine Facebook's role in amplifying hateful content, in particular in the Rohingya 2017 genocide. Each of the tasks mentioned in the research methods has been analyzed below with the relevant data and findings.

Task 1: The role of Facebook's Algorithm in Amplifying Hate Speech

This first task involved examining how Facebook's algorithms contributed to the spread of hate speech, specifically in relation to the Rohingya genocide and other conflict-prone regions. Research revealed that Facebook's engagement-driven algorithms amplified anti-Rohingya content in order to increase engagement. Meta's algorithms prioritized content based on user engagement, such as likes, comments, and shares, which inadvertently promoted that type of content (Amnesty International, 2022). These findings are also shared in the work of Khennour (2023), who explained how Facebook's amplification model inadvertently created echo chambers, in which users were more likely to interact with content that reinforced their biases. These algorithms and their structure increased tensions and divides in ethnic and religious lines. According to the Amnesty report, Facebook's failure to moderate this content, despite the early warnings, exacerbated the violence experienced by the Rohingya minors, contributing to atrocities (Amnesty International, 2022).

A key piece of evidence that further proves Facebook's involvement and its algorithm's failures came from Amnesty International's (2022) report, which explained how Facebook's algorithm prompted divisive content that outright violated community standards. For instance, videos from hate figures like U Wirathu were circulating on the platform, with over 70% of the views being driven by “chaining”, which is where content automatically plays after a previous video. This mechanism allowed harmful content to spread quickly (Amnesty International, 2022). These findings demonstrate the link between algorithmic amplification, the spread of harmful content, and its resulting in real-world violence.

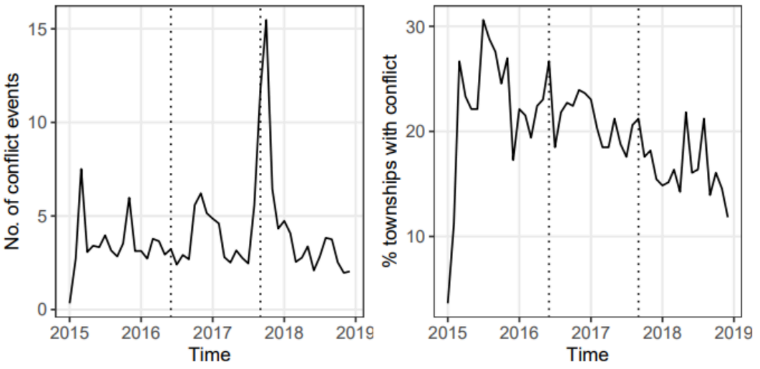

Figure 1: The chart on the left shows the average monthly number of violent incidents in townships in Myanmar, and the chart on the right shows the percentage of townships with violent incidents. The vertical dotted lines represent the beginning and end of the Facebook campaign. Data source: GDELT

Additionally, Figure 1 below shows the average monthly number of violent events (left) and the percentage of townships experiencing violent conflict (right) between 2015 and 2019. The vertical dotted lines mark the beginning (June 2016) and end (September 2017) of Facebook’s “Free Basics” campaign, which made Facebook the primary source of internet access in Myanmar. As the figure illustrates, there is a significant spike in both the number of violent events and the percentage of affected townships beginning in mid-2017. This increase correlates with the time frame when Facebook’s algorithms were amplifying divisive content, underscoring the connection between the platform’s algorithmic structure and the rise in violence. This data further substantiates the argument that Facebook’s algorithms, by promoting inflammatory content, contributed to the escalation of violence in Myanmar.

Task 2: Ethical Implications of Facebook’s Content Moderation Practices

The second task that was examined focuses on the ethical implications of Facebook’s content moderation practices, focusing primarily on their failure in adapting to linguistic and cultural differences/nuances that can be found in Myanmar. Meta’s content moderation system has been in place for a while and currently is automated. The issue becomes that the automated system struggled to address the spread of hate speech in languages that were not English. One of the solutions to this issue would be to hire moderators that speak the local language, but Facebook again failed to hire enough moderators that spoke the local language and its dialects. As noted by Amnesty International (2022), Meta had only one Burmese-speaking moderator stationed in its Dublin office in 2014, which persisted for years, leaving the platform ill-equipped to moderate content from Myanmar. The neglect of local contexts and languages, lack of investment in content moderation, allowed for hate speech to run rampant on the platform (Amnesty International, 2022).

The whistleblower from the Facebook Papers, Frances Haugen, has alleged that Facebook's algorithms have been stoking ethnic violence in countries such as Ethiopia. Frances claimed that 87% of the investment to prevent misinformation is for English content, while non-English-speaking countries receive 9% (Milmo, 2021). This further highlights how Meta’s content moderation efforts lack support for non-English-speaking regions, leaving countries like

Myanmar vulnerable to the spread of harmful content due to the lack of prevention resources at hand.

Task 3: Evaluating Potential Improvements in Content Moderation Policies

The third and final task focused on finding improvements to Facebook's Content Moderation Policies. Due to significant role that algorithmic amplification plays currently in spreading hate speech, the first thing to be done would be to reassess the approach used in content moderation. One of the ways it could be improved is instead of prioritizing engagement, prioritize user's safety and their rights.

A challenge that is often encountered in content moderation, especially in non-English-speaking religions, is the lack of resources that are dedicated to local language moderation. The figure below, " How Does Facebook Moderate Content," shows the effort of hiring content moderators, but the majority of them are English-speaking content moderators. In conflict regions like Myanmar, fewer resources can leave them vulnerable to harmful content (Hughes, 2021).

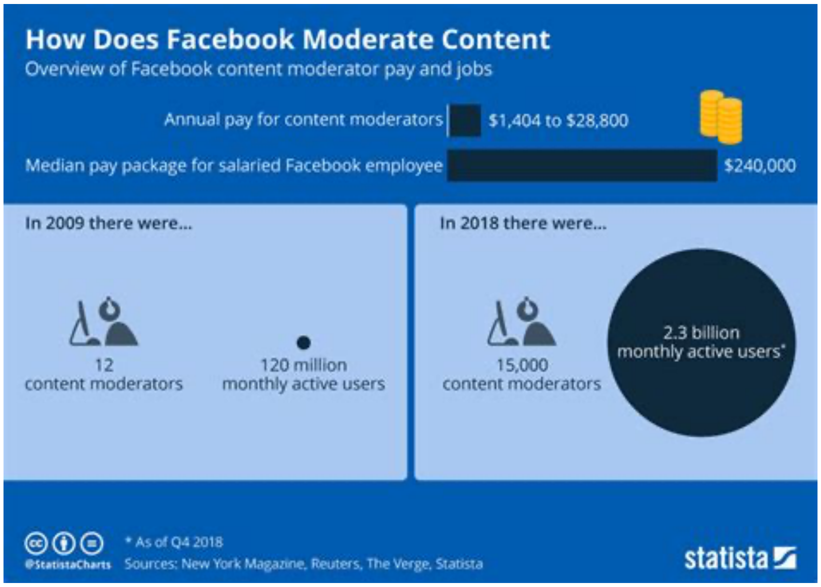

Figure 2: How Facebook Moderates Content. Source: Statista

To improve content moderation on the platform, it would require further investment in AI tools and human moderators that are proficient in local languages and contexts, especially in conflict regions. This investment would allow for better prevention and detection of harmful content on the platform. Additionally, when looking at other platforms, Facebook could implement solutions like in Twitter, where community notes are there to provide context and fight misinformation and hate speech.

Conclusion

Based on the research, it is clear that Facebook's algorithms are designed to prioritize user engagement rather than protecting their rights, and how they contributed significantly to the amplification of harmful content during the 2017 Rohingya genocide in Myanmar. The spread of anti-Rohingya narratives and engagement-driven algorithms exacerbated the violence in Myanmar. Meta's failure to properly address and tackle this issue despite numerous external and internal warnings has highlighted the ethical responsibility social media platforms and companies have to regulate harmful content.

The findings also showed that Facebook's content moderation practices, in particular in non-English-speaking regions, were insufficient to say the least in mitigating the spread of harmful content. The lack of local language moderators, which was shared in the amnesty report and in Milmo's article, highlights how further investment needs to be made to ensure that AI tools and local moderators are there for content moderation, especially in conflict-prone regions like Myanmar.

In conclusion, Facebook needs to reassess its content moderation policies, focusing on ethical aspects and implications rather than its current engagement/profit-driven algorithms. The company must prioritize user safety over engagement and invest in AI and human moderators who are culturally aware and proficient in the local dialects. The changes are essential in preventing another catastrophe like the 2017 Rohingya genocide in Myanmar.

Recommendation

Based on the findings, I recommend the following actions to Meta and to its CEO, Mark Zuckerberg:

1. **Prioritize User Safety Over Engagement:**

Facebook should adjust their algorithms to not be engagement-driven but to be more conscious of content quality/authenticity and user safety. This shift would help dramatically in reducing the possibilities of a situation like the one that occurred in Myanmar from occurring again.

2. **Increase Investment in Local Content Moderation:**

Facebook should invest more in AI tools and local human moderators that are more well-prepared and aware of local contexts and proficient in local languages. This would allow for more effective content moderation. This should specifically be done in conflict-rich zones like Myanmar.

3. **Implement Transparent Content Moderation:**

Meta as a whole should also be more transparent in how its algorithms work, allowing users to be more aware of how they receive the content they view on a day-to-day basis. This can allow users to make more informed choices and help combat the spread of misinformation and hate speech.

By implementing these recommendations, Meta can ensure that its platform is not complicit and is not used to contribute to violence and human rights violations through the spread of harmful content. It would also elevate Facebook in the global standing as a more ethically aware company that prioritizes its users' safety, protecting the vulnerable.

References

Amnesty International. (2022, September 6). Myanmar: Facebook's systems promoted violence against Rohingya – Meta owes reparations, new report. Amnesty International. https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/

Amnesty International. (2022, September 6). The social atrocity: A report on Facebook's complicity in Myanmar's genocide. Amnesty International. https://www.amnesty.nl/content/uploads/2022/09/Updated-Final-Report_The-Social-Atrocity-_06092022-1.pdf?x88970

Adapt Peacebuilding. (2019, December 20). The weaponization of social media: How digital platforms have enabled hate speech and violence. Adapt Peacebuilding. https://adaptpeacebuilding.org/2019-12-20-the-weaponization-of-social-media/

GDELT Project. (2019). Global database of events, language, and tone (GDELT). https://www.gdeltproject.org/

Ifo Institute. (2024). When Facebook's internet role: Social media's role in ethnic conflict. Ifo Institute. https://www.ifo.de/en/publications/2024/working-paper/when-facebook-internet-role-social-media-ethnic-conflict

Khennour, M. (2023). Examining the role of social media in fostering ethnic violence in light of Facebook's complicity in the Rohingya crisis.

Milmo, D. (2021, December 6). Rohingya sue Facebook for £150bn over Myanmar genocide. The Guardian. https://www.theguardian.com/technology/2021/dec/06/rohingya-sue-facebook-myanmar-genocide-us-uk-legal-action-social-media-violence

Haugen, F. (2021, October 5). Testimony before the U.S. Senate subcommittee on consumer protection, product safety, and data security. U.S. Senate. https://www.congress.gov/116/meeting/house/110235/witnesses/HHRG-116-JU10-20201005-SD001.pdf