Appeal to the Meta Oversight Board on widespread and repeated failures by Meta to implement its content moderation policies

March 2023

Complainant

Global Witness

Address: The Green House, 244-254 Cambridge Heath Rd, London E2 9DA, UK

Website: www.globalwitness.org

Contact: Rosie Sharpe, rsharpe@globalwitness.org

Table of contents

Our complaint

The Digital Threats to Democracy team at the non-profit organisation Global Witness has carried out investigations that indicate widespread, repeated failures by Meta to implement its content moderation policies across a range of topics, languages and jurisdictions.

In this complaint we describe our findings and ask the Meta Oversight Board to review whether Meta is implementing its content moderation policies to their full and intended effect.

In addition to this complaint, Global Witness is also submitting a second complaint together with Brazilian organisations that outlines further evidence from Brazil of content moderation failings by Meta.

What we found: evidence of widespread failure by Meta to implement its content moderation policies

Global Witness tested Meta's ability to implement its content moderation policies: we submitted content to Facebook that definitively breached the platform's Community Standards in the form of adverts and recorded whether Meta accepted or rejected them for publication.

Submitting content that violates Meta's Community Standards in the form of an advert - which could be removed prior to publication - allows us to test the company's content moderation systems without posting the violating content ourselves.

Meta has stated that it holds advertisements to an 'even stricter' standard than organic posts. Therefore, if violating content in an ad is not detected by Meta, we believe it is reasonable to assume that the same content is even less likely to be detected in an organic post.

In all cases we believe that we made the test as easy as possible for Meta to pass by using content that wildly breached the Community Standards and was written in clear language that is easy to understand. By design, none of the ads contained coded expressions or dog whistles. As examples, several of the ads we submitted for publication said that people of a certain ethnicity should be killed or raped or said that people of a certain ethnicity weren't human.

We have not made all of the text of the ads we used public because of the violent and offensive nature of their content, but for reference, we have included the text in an appendix available to the Oversight Board.

Our findings:

- In **Myanmar**, we submitted eight ads in Burmese containing real-life examples of hate speech inciting violence and genocide against the Rohingya taken from a UN fact finding mission. Meta accepted all of the ads for publication.
- In **Ethiopia**, Global Witness, working in partnership with the legal non-profit Foxglove and Ethiopian researcher Dagim Afewerk Mekonnen, submitted 12 ads in Amharic containing hate speech inciting violence and genocide during the ongoing civil war. Meta accepted all of the ads for publication. After informing Meta of this serious problem with their content moderation in Ethiopia, and a spokesperson acknowledging that the ads "shouldn't have been approved in the first place as they violate our policies," we submitted another two examples of real-life Amharic-language hate speech. Both ads were, again, accepted by Meta for publication.

- In **Kenya**, Global Witness, working in partnership with the legal non-profit Foxglove, submitted 20 ads, half of them in Swahili, half in English, containing hate speech and ethnic-based calls to violence ahead of elections in the country. Our English language hate speech ads were initially rejected for failing to comply with Meta's Grammar and Profanity policy. Meta invited us to update the ads, and after making minor corrections to the grammar and removing swear words, Meta accepted all of the English language ads for publication. All of the Swahili language ads were accepted for publication without any editing.
- In the **US**, Global Witness, working in partnership with the Cybersecurity 4 Democracy team at New York University, submitted 20 ads, half of them in English, half in Spanish, containing blatant election disinformation ahead of the 2022 midterm elections. We posted the ads from outside the US from an account that had not been through the "ad authorisations" process that Meta says they require to be able to post election ads. Meta approved 30% of the ads in English and 20% of the ads in Spanish. We tested the same ads again two days later, this time posting from a different account within the USA. This time, Meta approved 20% of the ads in English and 50% of the ads in Spanish.
- In the **US,** Global Witness, again working in partnership with the Cybersecurity 4 Democracy team at New York University, submitted 20 ads, half of them in English, half in Spanish, containing real-life death threats against election workers. The ads were submitted on the day of and the day before the 2022 midterm elections. Meta approved 90% of the ads in English and 60% of the ads in Spanish.
- In **Norway**, Global Witness, working in partnership with SumOfUs, submitted 12 ads, nine in Norwegian, three in English, containing extreme hate speech and disinformation including racist, anti-immigrant and anti-LGBTQ+ hate speech, text from the manifesto of far-right terrorist Anders Behring Breivik who murdered 77 people in Norway in 2011, health disinformation and extreme dieting messaging. Meta accepted all of the ads for publication.
- For investigations in **Brazil**, see our other complaint.

As well as providing evidence of widespread failings with Meta's systems for content moderation, we believe that these findings also reveal how Meta treats people differently according to where in the world they are. In Myanmar, Ethiopia, Kenya and Norway there was not a single ad we submitted that Meta rejected for publication.[*] In the US, however, Meta rejected at least some of the ads we submitted in both investigations that we have carried out there.

This finding holds true no matter what language the ads were submitted in: Meta accepted all ads for publication in Kenya no matter whether they were in Swahili or English and accepted all ads for publication in Norway no matter whether they were in Norwegian or English. Similarly, in the US, Meta rejected at least some of our ads no matter whether they were in English or Spanish.

We believe that this demonstrates that Meta puts more effort into content moderation in the US than it does in the other countries, despite the fact that the risks posed by hate speech and election disinformation are extremely high in some of the countries where we tested Meta's implementation of its policies.

[*] In addition, as described in our other complaint, Meta accepted all of the initial round of ads containing election disinformation that we submitted in Brazil. They accepted 14 of the 16 anti-democratic ads that we submitted in Brazil.

How Meta responded to our findings

For each of the above investigations, we contacted Meta to give them the opportunity to comment on our findings.

- Responding to our investigation on Myanmar, Meta did not reply to us. However, when the Associated Press put the same questions to them, they said what they had done to improve content moderation in Myanmar, including building a team of Burmese speakers and investing in Burmese language technology. (Full text of the statement published by AP in the footnote.[†])

- Responding to our investigation in Ethiopia, Meta said:
  "While these ads were removed before anyone saw them, they shouldn't have been approved in the first place as they violate our policies. We've invested heavily in safety measures in Ethiopia, adding more staff with local expertise and building our capacity to catch hateful and inflammatory content in the most widely spoken languages, including Amharic. Despite these investments, we know that there will be examples of things we miss or we take down in error, as both machines and people make mistakes. That's why ads can be reviewed multiple times, including once they go live, and why we have teams closely monitoring the situation and addressing these errors as quickly as possible."

- Responding to our investigation in Kenya, Meta said:
  "We've taken extensive steps to help us catch hate speech and inflammatory content in Kenya, and we're intensifying these efforts ahead of the election. We have dedicated teams of Swahili speakers and proactive detection technology to help us remove harmful content quickly and at scale. We've also created a team of subject matter experts working on the election, including individuals with expertise in misinformation, hate speech, elections and disinformation. Despite these efforts, we know that there will be examples of things we miss or we take down in error, as both machines and people make mistakes. That's why we have teams closely monitoring the situation and addressing these errors as quickly as possible."

- After we alerted them to our investigation, Meta then put out a public statement on its preparations ahead of the Kenya elections highlighting their apparent action taken to remove hateful content in the country. We then submitted two more ads to see if there had indeed been any improvement in Meta's detection of hate speech ads. Once again the ads we resubmitted in Swahili and English were approved.

- Responding to our investigation into election disinformation in the US, Meta said:
  "These reports were based on a very small sample of ads, and are not representative given the number of political ads we review daily across the world. Our ads review process has several layers of analysis and detection, both before and after an ad goes live. We invest

[†] "We've built a dedicated team of Burmese speakers, banned the Tatmadaw, disrupted networks manipulating public debate and taken action on harmful misinformation to help keep people safe. We've also invested in Burmese-language technology to reduce the prevalence of violating content. His work is guided by feedback from experts, civil society organizations and independent reports, including the UN Fact-Finding Mission on Myanmar's findings and the independent Human Rights Impact Assessment we commissioned and released in 2018." – Rafael Frankel, director of public policy for emerging markets at Meta Asia Pacific in an e-mailed statement to AP on March 17 2022

significant resources to protect elections, from our industry-leading transparency efforts to our enforcement of strict protocols on ads about social issues, elections, or politics – and we will continue to do so."

- Responding to our investigation into death threats against election workers in the US, Meta said:
  "This is a small sample of ads that are not representative of what people see on our platforms. Content that incites violence against election workers or anyone else has no place on our apps and recent reporting has made clear that Meta's ability to deal with these issues effectively exceeds that of other platforms. We remain committed to continuing to improve our systems."

- Responding to our investigation in Norway, Meta said:
  "Hate speech and harmful content have no place on our platforms, and these types of ads should not be approved. That said, these ads never went live, and our ads review process has several layers of analysis and detection, both before and after an ad goes live. We continue to improve how we detect violating ads and behavior and make changes based on trends in the ads ecosystem."

- Concerning investigations in Brazil, kindly refer to our other complaint, "Appeal to the Meta Oversight Board on repeated failures by Meta to implement its election disinformation policies in Brazil".

Our comments on Meta's responses

Our summary of Meta's response to our findings	**Our comments on the validity of this response as an explanation for our findings**
Meta's policies do not allow the type of content they approved for publication.	We agree, but our point is that they did not - or could not - implement these policies fully.
Meta has taken steps to improve content moderation.	We do not dispute this. Our contention is that whatever steps Meta has taken pre-date our investigations and therefore were not sufficient to enable Meta to implement their policies fully.
Meta states that there are several layers to ad approval, including once the ads have been made live. That is, they suggest that had our ads been published they could have been subject to further scrutiny and taken down.	Once an ad with content that blatantly breaches Meta's Community Standards goes live, it is liable to cause harm. We believe that Meta's automated and human review mechanisms ought to be able to detect such clear policy violations prior to publication.
Meta states that we only submitted a small number of ads.	It is true that compared to the number of ads that Meta accepts globally, our sample in each test is small; we do not however accept that this implies our findings cannot be used to draw conclusions about the company's content moderation systems as a whole. In particular:

	• One test with a relatively small number of ads provides a stark indication of content moderation failings when a large proportion of those ads are accepted for publication. • When the test findings are repeatedly found across a number of languages and jurisdictions, the conclusions are even clearer. The cumulative number of ads we submitted across investigations is over 100, each one of which has the potential to cause harm. • We believe that the text we submitted ought to have been easy for Meta to detect as, in each case, it wildly violated the platform's Community Standards and was written in clear language. We assume that in real-life, it will be substantially harder for Meta's content moderation systems to detect violating content than in the tests we posed. If they cannot pass these easy tests, we believe we are justified in concluding that they are likely to do even worse with real-life hate speech and election disinformation. • The methodology we have used to test Meta's content moderation is one of the few that is available to outside organisations. There is no way for an outside organisation to be able to submit substantially more ads than we have done as it would involve setting up a substantial number of Facebook accounts, which Meta does not permit.
Meta states that the ads we submitted were not representative of political ads.	We believe this is irrelevant. We hope that the extreme speech we submitted is not representative of most other ads on the platform; the point however is that we believe Meta's content moderation systems should be able to detect it.
Meta states that they are better at dealing with issues such as incitement to violence than other platforms.	When asked for the evidence that supports the claim that the platform is better at dealing with incitement to violence than other platforms, Meta provided quotes from technology experts published in the media. These experts stated that Meta has more resources devoted than other platforms and that it does better at moderation than some alt-right platforms (for the details of the quotes, see Appendix II). While these assertions may be factual they do not constitute evidence that Meta is better at

	detecting incitement to violence than other mainstream platforms. In addition, there should be no tolerance for failure before a major election, when tensions and potential for harm are high.

An appeal to review content moderation failings

It is our belief that Meta's content moderation policies are not being implemented adequately across the markets the company operates in.

We are therefore submitting this appeal to the Oversight Board to request that you review and report on whether Meta's content moderation policies are being implemented to their full and intended effect.

In doing so, we request that you:

- Establish if the failings we have uncovered indicate a systemic failure by Meta to protect users from hate speech, disinformation, and incitement to violence and genocide.
- Establish the cause of the failings in how Meta implements its content moderation policies, including by reviewing a) the efficacy of the machine learning systems that flag potentially violating content; b) the efficacy, resourcing, support, and working conditions of human reviewers who flag violating content; and c) Meta's human rights due diligence processes for meeting their requirements under the UN Guiding Principles on Business and Human Rights.
- Establish whether the efficacy of content moderation in different languages and jurisdictions is proportionate to the risks faced in those places. Our results imply that content moderation in the US (in either English or Spanish), while not good enough is nevertheless more effective than content moderation in Myanmar, Ethiopia, Kenya and Norway.
- Review the effectiveness of Meta's ad review processes after ads have been published.
- Make your research and conclusions public.

A note on the Oversight Board appeal process

This appeal to the Oversight Board concerns widespread and repeated failings in Meta's content moderation systems, rather than a decision on whether a specific post should be allowed or not. We have therefore not been able to submit our concerns via the formal appeals procedure as stated on your website.

In addition, after we reported to Meta the identities of the accounts used to submit our test ads, the company banned the accounts for violating their policies. We are therefore unable to provide you with the name of a specific Facebook account or reference number of a complaint. However, we believe we have fulfilled the spirit of your policies in that we have alerted Meta to the issue and have failed to receive a satisfactory response.

Appendix I – the text of the ads that Meta accepted for publication

The text of the ads that we submitted to Facebook was made available to the Meta Oversight Board.

Appendix II – Meta's private response to us on our US death threats investigation

In response to an email from us asking Meta to direct us towards the source material for their statement that Meta is better at dealing with incitement to violence than other platforms, Meta responded with the two bullet points below:

- "There is a lot of anger and noise on the mainstream platforms like Twitter and Facebook, but the most aggressive statements on the day of the midterms, including calls to violence, are found on the alt platforms including Gab, Parler and Telegram." – Alex Stamos, Stanford Internet Observatory
- "TikTok is absolutely grappling with the same issues…They tried to take more of a hard-line policy against disinformation but they have nothing like the staff and capacity and experience that you see at companies like [Facebook parent company] Meta and [Google parent company] Alphabet for dealing with these kinds of things." – Samuel Woolley, program director of the propaganda research team at the Center for Media Engagement at the University of Texas at Austin