Appeal to the Meta Oversight Board on repeated failures by Meta to implement its election disinformation policies in Brazil

March 2023

Complainants

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

Netlab - UFRJ

- Address: Av. Pasteur, 250 - fundos, Urca - Rio de Janeiro, RJ, Brazil
- Website: http://www.netlab.eco.ufrj.br/
- Contact: Marie Santini, marie.santini@eco.ufrj.br

Table of contents

_Toc126834748Our complaint 2

What Global Witness found: evidence of repeated failures by Meta to implement its content moderation policies 2

How Meta responded to Global Witness' findings 3

Our comments on Meta's responses 4

What NetLab found: ads attacking electoral integrity during the elections and ads calling for a coup after Lula's election victory 6

An appeal to review content moderation failings 6

A note on the Oversight Board appeal process 7

Appendix – the text of the ads that Meta accepted for publication 9

Our complaint

The NetLab research laboratory of the Federal University of Rio de Janeiro and the Digital Threats to Democracy team at the non-profit organisation Global Witness have carried out investigations that indicate repeated failures by Meta to implement its content moderation policies in Brazil.

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. A number of Brazilian organisations support this ask; their names are listed at the end of this document.

In addition to this complaint, Global Witness is also submitting a second complaint that outlines further evidence from outside Brazil of widespread and repeated content moderation failings by Meta across a range of languages and jurisdictions.

What Global Witness found: evidence of repeated failures 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 Global Witness believes 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.

Global Witness has not made all of the text of the ads we used public in order to avoid in advertently spreading disinformation and inciting violence, but for reference, we have included the text in an appendix available to the Oversight Board.

Global Witness' findings:

Investigation A

In August, Global Witness submitted 10 ads in Portuguese containing blatant election disinformation ahead of the 2022 elections in Brazil. We posted the ads from outside Brazil from an account that had not been through the "ad authorisations" process that Meta says they require to be able to post election ads.

Not only did Meta allow us to post adverts from an unauthorised account, Meta accepted all 10 ads for publication.

Investigation B

In September, after Global Witness' first investigation had been published and Meta had responded to our findings (see below), we re-tested Meta's ability to detect election disinformation in ads, again using an account that had not been through the "ad authorisations" process.

We:

- Re-submitted the same 10 ads that we used in investigation A from a new Facebook account. Meta accepted 40% of them for publication.
- Submitted 10 new ads containing blatant election disinformation in Portuguese. Meta accepted 50% of them for publication.

Investigation C

Ahead of the presidential run-off elections in October, Global Witness re-tested Meta's ability to detect election disinformation in ads a third time.

We re-submitted the second set of ads that we used in investigation B. As before, Meta accepted 50% of them for publication, though the individual ads that were accepted differed between the tests.

Investigation D

After the violent anti-democratic attacks in Brasilia on January 8, 2023, Global Witness tested Meta's ability to detect ads calling for a violent overthrow of the government and death threats against Lula voters and their children in Portuguese. Meta accepted 14 of the 16 (87%) ads for publication.

How Meta responded to Global Witness' findings

Global Witness contacted Meta to give them the opportunity to comment on our findings.

In response to investigation A, Meta said that they are committed to protecting election integrity in Brazil and that they prepared for the election in Brazil by launching tools to label election-related posts and establishing a direct channel for the Superior Electoral Court to send them potentially harmful content for review. They cited figures for the number of posts they removed in the last election for violating their policies. (For the full text of their statement, see the endnote.[1])

On August 16, after a report in O Globo, one of Brazil's leading newspapers about our joint findings, Meta announced that they would "prohibit ads calling into question the legitimacy of the upcoming election". However, this apparently new policy was in fact a policy that they had already put in place.[2]

[1] "We cannot comment on these findings as we don't have access to the full report. However, we prepared extensively for the 2022 election in Brazil. We've launched tools that promote reliable information and label election-related posts, established a direct channel for the Superior Electoral Court to send us potentially-harmful content for review, and continue closely collaborating with Brazilian authorities and researchers. Our efforts in Brazil's previous election resulted in the removal of 140,000 posts from Facebook and Instagram for violating our election interference policies and 250,000 rejections of unauthorized political ads. We are and have been deeply committed to protecting election integrity in Brazil and around the world." – a Meta spokesperson

[2] On page 29, Meta states they will ban "ads suggest[ing] voting is useless or not to vote" (under 'Steps to fight voter suppression') and that there is "No newsworthy exemption for ads or content ... suppressing voting"

In response to investigation C, Meta said that they "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" and went on to say that their ad review process has several layers of analysis and detections, and that they invest many resources in their election integrity efforts.

In response to investigation D, Meta said "This small sample of ads is not representative of how we enforce our policies at scale. Like we've said in the past, ahead of last year's election in Brazil, we removed hundreds of thousands of pieces of content that violated our policies on violence and incitement and rejected tens of thousands of ad submissions before they ran. We use technology and teams to help keep our platforms safe from abuse and we're constantly refining our processes to enforce our policies at scale."

Global Witness' comments on Meta's responses

Global Witness' summary of Meta's response to our Brazil findings	**Global Witness'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 states that there are several layers to ad approval, including once the ads have been made live.[3] 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. NetLab's findings of large numbers of ads on the Meta Ad Library that violate the platform's policies provide evidence to indicate that the approval process that ads may go through post-publication are not sufficient to adequately detect violating content.
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 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:

under 'Labeling newsworthy content') This parallels their announcement that "we will prohibit ads calling into question the legitimacy of the upcoming election."

[3] The clarification about the possibility of checks after an ad has been made live was made in response to our investigation in Norway (see Global Witness' other appeal to the Meta Oversight Board)

	• 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 repeated, we believe the conclusions are even clearer. • 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 is 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 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 removed a lot of violating content and rejected a lot of ads before the 2022 elections in Brazil.	We believe that this statement misses the point. In the tests that we have set Meta their processes have proved overwhelmingly incapable of detecting violating content. The fact that Meta has detected some violating content– some of it possibly because it was flagged by users rather than their own systems – does not negate the findings of our investigations.

	It is impossible to put the numbers that Meta states for the amount of content removed and rejected into context. The amount removed might be a large number, but if the amount present on the platform is a significantly larger number it remains true that Meta's content moderation is not up to the task.

What NetLab found: ads attacking electoral integrity during the elections and ads calling for a coup after Lula's election victory

Netlab collected ads with content on the integrity of the Brazilian electoral system available in the Meta Ads Library that ran between June 26 and July 31, 2022. Of the 160 ads that we found, 27 (17%) attacked the country's electoral system.

Netlab's analysis indicated that language Jair Bolsonaro had used to attack the electronic voting machines, the electoral system and the TSE was reproduced in the ads run on Meta, paid for by pro-Bolsonaro pre-candidates.

Following the publication of these findings, Meta announced on 16 August 2022 that it would ban advertisements questioning the legitimacy of the Brazilian elections. To assess how well this policy was implemented, Netlab collected ads about the electoral system available in the Meta Ads Library that ran between August 16-31, 2022, including ads running on Facebook, Instagram, Messenger and Audience Network. Of the 1 ads that we found, 14 (10%) attacked the country's election system showing evidence of Meta's insufficient enforcement concerning material which is potentially damaging to the democratic process.

In the wake of the Brazilian antidemocratic acts that led to the invasion of the Praça dos Três Poderes (three important government buildings in the capital Brasilia, including the Supreme Court) on January 08, 2023, Netlab found 185 ads on Meta library related to the coup agenda which contested the election results, attacked the electoral process and encouraged antidemocratic demonstrations. We verified and evaluated the content of these ad pieces, with information on the dates of publishing and the pages responsible.

Therefore, Netlab concludes that Meta's declared efforts to remove content with antidemocratic messages, containing attacks on Brazilian institutions and against the electoral process were insufficient in the post-electoral period. In the case of ads containing this type of message, Netlab finds Meta's enforcement is even more questionable, considering that ads are reviewed by the platform using AI and human curation that approve the content before being published.

An appeal to review content moderation failings

It is our joint belief that Meta's content moderation policies are not being implemented adequately.

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 the extent to which Meta's claim that violating ads might have been detected after being published is correct.
- 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 repeated failings in Meta's content moderation systems in Brazil, 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.

Supporters of this appeal to the Meta Oversight Board

The following Brazilian organisations support this appeal and also request the Meta Oversight Board to review whether Meta is implementing its content moderation policies to their full and intended effect.

Aláfia Lab

Coding Rights

Conectas Direitos Humanos

*desinformante

Intervozes

Novelo Data

Rede Nacional de Combate à Desinformação

Sleeping Giants

Appendix – 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.