**Submission to the Meta Oversight Board re: Post in Polish Targeting Trans People case**
***By Paige Collings (Electronic Frontier Foundation)***

**Introduction**

Just as Facebook can be used for positive advocacy, it is also routinely used with the intention to cause harm. That was clearly the case in April 2023, when a Polish user posted an image of a curtain in the colors of the transgender flag with the text overlay stating (in Polish), “New technology. Curtains that hang themselves” and “spring cleaning <3.” The intent behind this message is clear: To encourage self-harm by and violence toward transgender individuals.

While this content was recognized and reported by a number of users, Facebook’s automated systems failed to prioritize the content for human review. From our observations—and the research of many within the digital rights community—this is a common deficiency made worse during the pandemic, when Meta decreased the number of workers moderating content on its platforms. In this instance, the content was eventually sent for human review and was still assessed to be non-violating and therefore not escalated further. Facebook kept the content online despite 11 different users reporting the content 12 times and only removed the content once the Oversight Board decided to take the case for review.

This incident serves as part of the growing body of evidence that Facebook’s systems are inadequate in detecting seriously harmful content, particularly that which targets marginalized and vulnerable communities. Our submission will look at the various reasons for this shortcoming and make the case that Facebook should have removed the content—and should keep it offline.

**The Shortcomings of Automated Decision-Making and Poorly Trained Human Reviewers**

As EFF has demonstrated, Meta has at times over-removed legal LGBTQ+ related content whilst simultaneously keeping content online that depicts hate speech toward the LGBTQ+ community. This is often because the content—as in this specific case—is not an explicit depiction of such hate speech, but rather a message that is embedded in a wider context that automated content moderation tools and inadequately trained human moderators are simply not equipped to consider. These tools do not have the ability to recognize nuance or the context of statements, and human reviewers are not provided the training to remove content that depicts hate speech beyond a basic slur.

The lack of transparency only adds to the complexity of the issues as Meta does not disclose the detailed criteria for content moderation, including enforcement guidelines related to internal policies—making it difficult to assess the scale and contours of such bias as reflected in opaque internal policies, as well as any potential built-in bias regarding the moderation of LGBTQ+ content. Additionally, because algorithms can only be trained on known examples, they are more likely to remove similar kinds of content and can be blind to others. The challenges of

content moderation enforcement in languages other than English—such as Polish—further exacerbates these issues.

In countries like Poland where anti-LGBTQ+ hate speech and harassment is so prevalent both online and offline, Meta's inconsistent and inflammatory content removal systems are even more detrimental. As highlighted in GLAAD's 2023 Social Media Safety Index (SMSI) report, Meta's Facebook and Instagram are largely failing to mitigate dangerous anti-trans and anti-LGBTQ+ hate and disinformation, despite such content conflicting with the sites' policies. The June 2023 SMSI also made the specific recommendation to Meta and others that they better train moderators on the needs of LGBTQ+ users, and enforce policies around anti-LGBTQ content across all languages, cultural contexts, and regions.

Under international human rights law, restrictions to rights such as freedom of expression (article 19 ICCPR) and freedom of assembly and association (articles 21 and 22 ICCPR) can only be justified if there's a legal basis, a legitimate aim, and if they are necessary and proportionate. Without adequately taking into consideration the context in which words and audio-visual content is used, benign content is suppressed whilst hate speech and content inciting violence is able to remain online; thereby failing to meet the conditions to restrict freedom of expression, civic engagement, and activism under international human rights law.

**Recommendations**

It is of vital importance that online speech is put into its appropriate context. The Rabat Plan of Action provides guidance for companies seeking to remain in compliance with the UN Guiding Principles. Meta should consider (1) the social and political context prevalent at the time the post was uploaded; (2) the user's position or status in the society, specifically the individual's or organization's standing in the context of the audience to whom the post is directed; (3) the intent of the user in relation to their audience; (4) the content of the post; (5) the extent of the post, taking into account the post's reach, its public nature, its magnitude, and size of its audience; and lastly (6) the likelihood, including imminence, of harm to result from the post.

Meta's Trust and Safety teams at Facebook, Instagram, and Threads must also provide adequate training to human reviewers to recognize how hate speech and incitement to violence can appear in a more nuanced manner than basic slurs or hateful images. The image shared on Facebook in April 2023 was a clear illustration of anti-trans hate speech, and the arbitrary and inefficient content review—first by the automated system and second by the human reviewer that chose to keep the content online—has a particularly detrimental impact for the LGBTQ+ individuals using online platforms in countries like Poland, where hate and harassment is so prolific. The anti-trans content posted on Facebook in April 2023 must remain offline.