The Meta Oversight Board in the Trump Era

A Subtle Shift with Far-Reaching Consequences?

Juha Tuovinen

26.05.2025 doi: 10.59704/b0f5cab0ed937126

In late April 2025, the Meta Oversight Board ("Board") published its first six decisions since Meta implemented sweeping changes to its content moderation policy. While these decisions prompted concern among many over ongoing inconsistencies in its approach to freedom of expression, some also welcomed the Board's critical stance on certain recent changes (see here and here).

In this post, I argue that the Board appears to be shifting towards a more permissive stance on harmful or discriminatory expression. If this trend continues, it could significantly reshape the boundaries of acceptable speech on Meta's platforms – with far-reaching implications for online public discourse. This also raises broader questions about the Board's independence and legitimacy at a time when content governance is becoming increasingly politicised.

The Board and the january announcement

The Board is intended to be an independent body that "examine[s] whether Meta's [original content moderation] decision was in line with its policies, values and human rights commitments". In order to do this, the Board decides cases that (usually) have been appealed to it by users of Meta's platforms (Facebook, Instagram, Threads). Regardless of whether the Board actually improves content moderation or merely serves as a window dressing, its interpretation of the content moderation rules significantly shapes the kind of expression that Meta's approximately two billion users can post and are exposed to.

Following Donald Trump's election as President of the United States, Meta CEO Mark Zuckerberg announced far-reaching changes to the way content moderation would function on Meta's social media platforms. The details of these propositions have been analysed extensively. The general tenor was to be more permissive of expression, referring to undoing censorship and alleged liberal bias in the company's content moderation practices. The Board's initial reaction was relatively mild, stating that it "looked forward to working with Meta" on these issues. The six decisions in late April struck a more critical tone, expressing the Board's dissatisfaction with the speed at which the changes were implemented without undertaking a human rights compliance assessment.

However, in two controversial cases, the Board held that problematic posts could remain online. *The Apartheid Flag* case involved two posts that used South Africa's pre-democratic flag in a positive light – once in reference to military service, once as part of a collage of photographs from the era – though not in a way that explicitly endorsed Apartheid outright. The case concerning the Gender Identity Debate Videos involved two posts as well. The first showed an identifiable transgender woman using a women's bathroom and was captioned: "Male student who thinks he's a girl". The second video showed a transgender girl winning a race, named the athlete, and stated: "Boy who thinks he's a girl".

Tightening the standard for removing posts

How did the Board reach its conclusion that these posts were permissible? The Board bases its decisions on two different types of normative sources. First, Meta's own Community Standards, which provide a rationale for why certain content is banned, including examples of prohibited content and exceptions. Second, international law is applied as another set of norms, although the Board has repeatedly held that international law, as applied to private entities, does not have to mirror the same standards that apply to

states. In this post, I will focus on how the Board conceives the link between the expression in question and the potential harm it may cause – an issue central to interpreting both the relevant Community Standards and international law.

In both cases, the Board was split on whether the posts amounted to hateful conduct and/or violated international law. Regarding the flag posts, the majority of the Board stated that "[w]hile the flag is invoked positively in this post, that context is specific to military service and there is no sufficiently clear statement [...] that apartheid [...] should be reinstituted." Regarding the collage, the Board found that "[the pictures] feasibly evoke general nostalgia for the period they depict" and that "[positive statements and emojis], [...] are indicators of a racist message [...]". Nevertheless, racially insensitive messaging is not prohibited, and the post did not "rise to the level of advocacy for the reinstitution of apartheid [...]".

In both videos involving transgender individuals, the Board also concluded that the posts did not violate the hateful conduct standard. Although each post referred to the biological sex of the individuals and asserted that they "think" they are the opposite sex, the Board held that "it does not amount, even by inference, to a statement that transgender people or identities do not exist."

Both decisions reflect a high threshold for defining hateful conduct — perhaps influenced by Meta's policy shift toward greater protection of expression (even though there were minority opinions in each case). Where this high threshold should not play a role, however, is in the Board's application of international law. Yet here, too, the Board appears increasingly protective of expression.

When assessing the proportionality of permitting the Apartheid flag, the majority noted South Africa's "relatively stable representative democracy since the end of Apartheid and its robust legal framework for protecting human rights," as well as the absence of any "clear intent to advocate for future acts of discrimination or violence." They stressed that "[e]ven if considered to be 'deeply offensive,' this does not convert them to incitement [...]."

Similarly, in the Transgender Videos decisions, the majority underscored the lack of any explicit call to discrimination and emphasized the importance of public debate in a democratic society. As with the hateful conduct threshold, the Board appears to demand a very close connection between the expression at issue and any potential harm that could justify its removal.

A shift from a more protective approach?

The reasoning in both cases stands in stark contrast to some of the Board's earlier decisions, in which it highlighted that the cumulative effect of harmful stereotyping can, over time, foster an environment in which discrimination and violence are more likely.

For example, in Depiction of Zwarte Piet (2021), the Board reviewed a video showing a family with their faces painted black, following a Dutch tradition. Facebook had removed the post under its Community Standards, which prohibit caricatures of Black people. While the tradition might be practiced by some Dutch people without apparent racist intent, the Board concluded that the portrayal was discriminatory and reinforced harmful stereotypes. It justified the restriction by arguing that the accumulation of such harmful stereotypes could create an environment conducive to discrimination or violence.

Similarly, in South African Slurs (2021), the Board examined a case involving a presumably Black South African user who used racial slurs to criticize other Black South Africans for subjugating themselves to white South Africans. Although the post addressed important social and economic issues, the Board ruled that it must be taken down, as the use of racial slurs is "closely linked with discrimination and the history of apartheid in South Africa."

Contrasting these earlier decisions with the more recent ones, the Board's emphasis appears to have shifted. If the use of blackface in a private context is impermissible, it is difficult to understand how general nostalgia for Apartheid or calling transgender individuals delusional could be deemed acceptable. It also appears contradictory – and indeed, rather ironic – that a white South African is permitted to express

positive sentiments about Apartheid, while a black South African is not allowed to use forceful language to criticize the very system that oppressed them.

At the same time, while this shift may seem stark, we must be careful not to overinterpret it. Drawing conclusions from cases involving different forms of expression in different contexts can be challenging – visual communication can differ fundamentally from written language, for example. The real issue may lie in the ad hoc nature of the Board's balancing tests rather than in a shift in its overall philosophy. Also, both of these recent decisions were split, with dissenting members advocating for removal (often citing the Board's own prior decisions). Likewise, many earlier decisions featured dissents calling for stronger protection for expression. To some extent, the apparent shift may therefore reflect changes in the Board's composition rather than a fundamental alteration of its principles.

Conclusion

This post has argued that the Board may be rethinking how it weighs the relationship between social media expression and the potential harm it may cause. Such a change could have significant implications for the billions of users across Meta's social platforms, both on- and offline. What is particularly concerning is that this trend appears to protect the expressions of those who use vague or humorous forms of language to spread hateful messages. With the Board's approval, it becomes more difficult to regulate against these ambiguous forms of hate speech.

Beyond the immediate question of permissible content, these decisions also raise broader concerns about the Board's independence. If its reasoning increasingly aligns with Meta's prevailing interests, its legitimacy as an external check on corporate power may be seriously undermined.

ABOUT THE AUTHOR

Juha Tuovinen
Juha Tuovinen holds a PhD from the European University Institute (EUI) in Florence.

SUGGESTED CITATION

Tuovinen, Juha: The Meta Oversight Board in the Trump Era: A Subtle Shift with Far-Reaching Consequences?, Verfassungsblog, 26.05.2025, doi: 10.59704/b0f5cab0ed937126.

Licensed under CC BY-SA 4.0.

Langzeitarchiviert bei IntRechtDok unter: doi.org/10.17176/20250627-142159-0.