The Facebook Oversight Board and ‘Context’

Juncal Montero Regules 2021-02-16T09:37:09

The Facebook Oversight Board (FOB) has announced its first five decisions. After (ongoing) controversies on its creation (see here and here), and criticism on every step of the Board’s path (see here and here), the first decisions shed some light on the FOB’s approach to its cases and to content moderation more generally. The FOB overturned four Facebook’s takedown decisions and upheld one. That is: the Board concluded that in 80% of the cases the content had to be put back up, and in 20% of them its removal was correct and the content had to stay down. At the outset, these numbers can be seen as a strong stand on freedom of expression: “more free speech on Facebook! Put up those posts!” the Board seems to say (for a summary of the decisions, see here).

The FOB uses three sets of rules for examining all of them: Facebook’s Community Standards, Facebook’s Values (“Voice, Authenticity, Safety, Privacy, Dignity”), and International Human Rights. Two cases were removed for violating Facebook’s Hate Speech Community Standards. One of them is the only case where the removal decision was upheld.

I argue that the standout conclusion of the two hate speech decisions is the Board’s heavy reliance on context, in assessing the content’s removal. This is only reasonable, as any speech issue is context-dependent. But the FOB’s context assessment is incomplete, just as its decisions further highlight Facebook’s content moderation flaws, which likewise fail to consider context.

Hate speech in Myanmar

The first case concerns a post in Burmese in a private group, self-described as “a forum for intellectual discussion”. A Facebook user in Myanmar posted two widely-shared photographs of a Syrian toddler or Kurdish ethnicity who drowned in the Mediterranean in September 2015, accompanied by a text stating that “there is something wrong with Muslims psychologically”. It questioned the lack of response by Muslims generally to the treatment of Uyghur Muslims in China, compared to killings in response to cartoon depictions of the Prophet Muhammad in France. The post concludes that recent events in France reduce the user’s sympathies for the depicted child, and seems to imply the child may have grown up to be an extremist.

Facebook categorized the content as hate speech and removed it, considering the written statement as a generalization of mental deficiency regarding Muslims. On appeal, the user explained that the post was sarcastic. The statement is obviously inflammatory and raises a number of questions. Do the privacy settings of a group affect the content’s scrutiny? What about the group’s title or label? It also brings up

a difficult topic for content moderation: sarcasm. Identifying sarcastic comments is highly context-dependent and deemed to fall out of AI's capabilities. Myanmar is a particularly delicate context for content moderation (to say the least).

The Board's Reasoning

The FOB overturned Facebook's removal decision – the content had to be restored. Reading the case as a whole, the Board concludes that the post is better understood as a commentary pointing to alleged inconsistencies between Muslims' reactions to events in France and China. The statement, as interpreted by the Board, is protected under the Community Standards and does not reach the level of hate speech that would justify removal. Likewise, under Facebook's Values, the FOB finds that the content did not pose a risk to the "Safety" value that would justify displacing "Voice". Voice is Facebook's paramount, overarching value, whose purpose is to "build community and bring the world closer together". Voice means people posting content – the core of Facebook's business. Per the Safety value, "expression that threatens people, has the potential to intimidate, exclude or silence others […] isn't allowed on Facebook". No explanation is given for why the content does not threaten people enough to justify a restriction of Voice, nor the Board's method of weighing up the values. Rather, the FOB's assessment is solely grounded in the Community Standards. Lastly, the statement is assessed in the light of Human Rights Standards, under which the Board does not consider the statement's removal to be necessary to protect the rights of others, despite some considering it offensive and insulting towards Muslims.

Context in Myanmar

In this case, the assessment of the context is limited to a new translation of the statement. Nuance in translations can totally change a statement's meaning, intent and potential impact – as it seems has happened here. The actual context of Muslims in Myanmar is not further considered, despite Facebook's generally-acknowledged role in the Rohingya genocide. It can easily be argued that in such a context of violence and discrimination, the statement in question could eventually contribute to real violence, harm and discrimination. Offending and insulting a protected category of people is not the same in all geographical and social situations, something the Board does not consider here.

It is also surprising that the nature of the group is not relevant for the assessment. Arguably, as a private group, its reach and impact is low to negligible for influencing real-life actions. But the opposite is also true. Private groups are a breeding ground for hateful and harmful speech, leading to radicalization and polarization. Remarkably, group membership numbers are unknown in this case. When assessing controversial content published in a private group, the number of members, and members' reactions to the content in question should be considered relevant context.

Novel Information

In this case, the FOB reveals something new about Facebook, which was not in the public domain: For the company, generalizations are "unqualified negative

statements, with no room for reason, factual accuracy, or argument and they infringe on the rights and reputations of others". While Facebook should be more transparent about its moderation mechanisms and internal rulebooks, disclosure through the FOB's cases is a positive outcome in the meantime.

Demeaning slur in the Armenian-Azerbaijan context

The only case in which the Board upheld Facebook's removal involves a public post, showing photos of churches in Azerbaijan, accompanied by a text in Russian, claiming that Armenians built Baku and that this heritage has been destroyed. The post was uploaded during the recent armed conflict between Armenia and Azerbaijan in November 2020, and received more than 45k views. It used the term "taziks" to describe Azerbaijanis, who are called nomads with no history compared to Armenians. Facebook removed the post for violating its Community Standards on hate speech, claiming the post used a slur – "taziks" – to describe a person or group of people on the basis of a protected characteristic, namely national origin.

Prohibited slurs and context

We learn that Facebook has an internal list of prohibited slurs, "which it compiles after consultation with regional experts and civil society organizations". But there remains uncertainty – are there different slur lists for different geographical and sociopolitical contexts? A slur is so context-dependent that a worldwide list would definitely enter censorship territory.

The term "taziks" features in this list. Under the Community Standards, the Board concluded that its use was meant to dehumanize its target, considering the context in which it was used. It emphasized that the prohibition of slurs targeting national origin is intended to prevent users from posting content meant to silence, exclude, harass or degrade others. The ongoing, long-standing armed conflict between Azerbaijan and Armenia is especially relevant – the content in question was posted shortly before a ceasefire went into effect: "the danger of dehumanizing slurs proliferating in a way that escalates into acts of violence is one that Facebook should take seriously". In this case, the values of "Safety" and "Dignity" surpass the supreme value of "Voice" because of the context of latent, ongoing conflict and because the statement includes a slur targeting national origin, whose use is prohibited by Facebook.

Disagreement on human rights

The assessment under Human Rights standards leads to disagreement within the FOB. The majority found the removal proportionate, as less severe interventions would not have provided the same protection: that the content would stay down. A minority found it disproportionate, arguing that the risks cited by the majority were too remote and unforeseeable, and that alternative, less-intrusive enforcement options should have been considered. A separate minority argued that the reference to an inanimate object ("tazik" is directly translated as "wash basin") was offensive but not dehumanizing, and that the slur would not flame violent action. This is

the only case where the FOB (openly) disagrees, namely in the proportionality assessment, showing that the case was far from crystal-clear.

Context proves crucial, again. The ongoing armed conflict and the term used in the post, posted in the run-up to a ceasefire, contribute to its categorization as dangerous. The broad viewership of the post was not mentioned in the FOB's assessment, even though 45k views would support the Board's argument for keeping it down.

Context moderating?

Context is the turning point in the two hate speech cases, with different results. The context of the post itself – the user's intent – and the general sociopolitical context in which it was posted, meaning a post's place within a conflict and its potential to cause actual risk or harm are crucial. Surprisingly, the nature and reach of a post is not relevant for the FOB. This approach is questionable and leaves the Board's assessment incomplete: an oversight mechanism for online content should, by any logic, consider spread numbers as well as the post's nature (public or private, in a group, in the user's page or in comments), which are also *context*.

These decisions do not tell us which direction the FOB is taking: but it shouts that "context matters!". Context, however, is never clear-cut. In fact, the Board does tell us that: 1 – Facebook's global rules do not work because they are not context-dependent, and 2 – large-scale moderation is flawed because it fails to consider context (in most cases). The degree of detailed examination needed for circumstantial understanding of each piece of content, to be able to contextualize and assess them, is just impossible to apply at scale.

Final thoughts

These decisions mark a relevant moment for online speech moderation, whether you choose to consider them the Marbury v Madison of platform governance, or not. This is the first time an oversight mechanism – set up by a social media platform – has a say in the platform's content moderation decisions. While the Board does lean into the "more free speech" direction, it mostly relies on the circumstances around that speech, implying that the current content moderation mechanisms are flawed. The FOB asks for more transparency, communication with the users, and rule clarification. Although not new, it is good that the FOB points these recurring criticisms out to Facebook, who has to listen. We will see to what extent and how.

But let us not forget about the implications of the Board's existence. While it is an interesting one-of-a-kind experiment in content moderation, it is first and foremost a PR exercise from Facebook. It is a body created, set up and financed by the company, created as a response to legislative, civil society and user concerns about the decision-making processes and outcomes of controversial content moderation decisions. The Board appeases the public and serves as corporate whitewashing. While its decisions and existence may be interesting and even, to a certain extent,

positive, it shouldn’t prevent us from talking and debating the many other issues involving content moderation, platform regulation and online speech.
