On Facebook's New "Oversight Board", Accountability, and Control

By Noa Mor (Forthcoming DLI Fellow)

On Facebook's New "Oversight Board", Accountability, and Control

By Noa Mor (Forthcoming DLI Fellow)

Facebook recently published a detailed plan for the establishment of an "Independent Oversight Board", which will review content moderation cases that arise within the platform.

In a press call, Facebook explained that their plan is the result of almost a year's work with 100+ people within the company, and supported by an extensive consultation process that included roundtables, workshops and discussions with hundreds of people around the world. The purpose of the board, explained Facebook in document called "the charter", is "to protect free expression by making principled, independent decisions about important pieces of content and by issuing policy advisory opinions on Facebook's content policies." The company further argued that the board will improve their accountability and decision-making, and expressed hope that additional companies will follow suit. Furthermore, a letter issued by Mark Zuckerberg expresses his view that "I don't believe private companies like ours should be making so many important decisions about speech on our own".

However, a quick look at Facebook's full range of content moderation practices reveals that while Facebook puts significant efforts in the establishment of the board, it also cultivates new, powerful and sophisticated practices of censorship that are not likely to be reviewed by its self-created appeal mechanism.

In this short essay, I describe Facebook's new board, and explain why, in my opinion, it will not create true accountability on behalf of the company with relation to content moderation, and why the company is apparently still keen on making important decisions regarding content moderation on its own.

What is this board about?

The board will deal with cases of two types: those referred to it by Facebook itself, when it requires the board's guidance on a case (but also on other content moderation issues), and by Facebook's users who wish to appeal the company's decision concerning their content, after going through Facebook's internal appeals process. The board is expected to start hearing cases "early in 2020".

The board will consist of 11-40 members and will hear only a relatively small number of decisions a year ("in the dozens"). Facebook will choose the first members, but later on, the board itself will choose its

members. Members are supposed to come from various backgrounds such as adjudication, legal practice, publishing, editing, and journalism.

The board will have a “case selection committee” that will choose the cases to be reviewed from those brought to it by Facebook and the users. It will then assign them to a 5-member panel for review. Each panel will be able to ask Facebook for information that is “reasonably required” to make its decision. It can also address experts when deciding on a case. Decisions are expected to be “in accordance with Facebook’s content policies and values.” (If you find yourself perplexed as to what these values are, according to Facebook’s statement, they include voice; authenticity; privacy; safety, and dignity).
The board aims to bring together members from different places in the world, and the panels may, therefore, convene virtually. Members will have to follow bylaws that will provide a procedural frame for the board’s operation.

The board’s decisions will be made available to the public. They will be binding and Facebook promises to promptly implement them unless they violate the law. Facebook might also implement the decisions in “other instances or reproductions of the same content”, “to the extent technically and operationally practicable”. If the panel suggests policy recommendations, they “will be taken into consideration by Facebook to guide its future policy development.”

In order to maintain the independence of the board, explained Facebook, it created an additional entity – “the trust”. In the “Charter” mentioned above, Facebook described the relationship between the company, the trust and the board, and these entities’ authorities and tasks. One of the important goals of the trust is to make sure that board members are not directly compensated by Facebook. Instead, the board will receive its funds from the trust, and the trust will get its funds from Facebook.

Why will the new board not afford true accountability?

Before diving into some of the difficulties that the new board introduces, it is important to acknowledge that it is a step in the right direction. Until recently, Facebook users had no appeal mechanism to turn to if their content was removed. Last year, Facebook launched an inner appeal stance for such a purpose (though with significant limitations), and later provided some data regarding this appeal mechanism in its Transparency report. The establishment of the “Oversight Board” is another advancement. Regardless of the incentives that propelled its establishment by Facebook, this institution has the potential to challenge some of the company’s policies and content removal practices, and to create public discourse around these issues. This is important at a time when content is increasingly removed immediately after it was posted (or even prevented from being posted in the first place), so its removal can easily go unnoticed, without sparking any public discussion.

However, in my opinion, even when having these benefits in mind, this mechanism will not make the platform truly accountable and will not truly protect free speech. There are several reasons for this pessimist prediction. The one I will discuss here lies in the board's inability to cover the full scope and depth of content moderation practices that are carried out by Facebook.

Facebook's content moderation (a nice wording for censorship), is much more sophisticated, multifaceted and complex than simply content removal. It includes layers of choice architecture, algorithmic and AI-driven practices that are hard to detect by users, and are, therefore, not likely to be subject to appeals made by them. You just cannot appeal what you do not know. One fashion in which Facebook determines the content that will populate its platform without removing it, involves prioritizing of the content that will be introduced to users. Facebook has been doing this for a long time now by tailoring the content that will be displayed to each user; preferring certain types of media; and more. Such tactics for favoring or disfavoring content are also reflected in many holistic and wide initiatives to craft the information flow on Facebook. One example can be found in Zuckerberg's statement from last year, in which he explained that: "I'm changing the goal I give our product teams from focusing on helping you find relevant content to helping you have more meaningful social interactions."

Another example of such a holistic content moderation initiative, which I will focus on here, is described by the company as "Discouraging Borderline Content". This AI-based practice essentially deprioritizes content that approaches the boundaries of the company's prohibited-content area, so that it enjoys "less distribution and engagement", both in the News Feed and in various groups and pages suggested to users.

To be sure, the content that might be "penalized" (in Zuckerberg's words) is not prohibited according to Facebook policies, but only "gets close" to the borderline of this prohibited area.

Why should such content be discouraged then, and why should Facebook interfere with what they call a "natural engagement pattern"? Zuckerberg described this content as "provocative" and "sensational". He explains that "as a piece of content gets close to that line, people will engage with it more on average -- even when they tell us afterwards they don't like the content." An alternative explanation for Facebook's discomfort with such content may be connected to the option that this kind of content involves less visible engagement, and therefore not serve the company's best commercial interests. A clue for this might be found in what Zuckerberg addressed as an indication for borderline content. He explained that "Another factor we will use to try and show fewer of these types of stories is to look at the ratio of people clicking on the content compared to people discussing and sharing it with their friends. If a lot of people click on

the link, but relatively few people click Like, or comment on the story when they return to Facebook, this also suggests that people didn't click through to something that was valuable to them."

Be the motivation to this initiative as it may be, "Discouraging Borderline content" is a very troubling censorship tool, both procedurally and substantively.

On the procedural level, this mechanism is about opaque and non-transparent as it gets. It allows Facebook to effectively control exposure, engagement and dissemination of information, with very little accountability. Creators of content will face a hard time understanding that their content was silenced (let alone prove it), and this kind of content moderation will most likely not make it to the company's Transparency Reports. Furthermore, the scope of this penalizing mechanism is not clear, nor is the types of content subject to it (and whether it's applied to all such types evenly). Zuckerberg did explain that the platform is focusing on "click-bait and misinformation", but further provided that "Interestingly, our research has found that this natural pattern of borderline content getting more engagement applies not only to news but to almost every category of content. For example, photos close to the line of nudity, like with revealing clothing or sexually suggestive positions, got more engagement on average before we changed the distribution curve to discourage this. The same goes for posts that don't come within our definition of hate speech but are still offensive." He also mentioned that the company aims to reduce "sensationalism of all forms".

Such vagueness is a fertile soil for mistakes, non-legitimate commercial practices, and abuse by various governmental and nongovernmental actors. This kind of content moderation also makes it very difficult for users to understand what content is legitimate and what is not, and make sure that the information they have created will not be directed to the bottom of the endless arsenal of online information.

The substantive level also raises deep concerns. One such is that it is often the "problematic", "out of line", bold and straightforward content that surfaces problems, allows speakers to crystallize their arguments and to listeners to reflect on what was said and rethink their opinions. Political or social criticism might, for instance, be of such a nature. Also, unlike Zuckerberg's explanation that avoiding public traces of interaction with certain content indicates its lack of value to users, this "private" interaction pattern may sometimes only suggest that people care to keep the interest they have in a certain piece of content to themselves. As long as this content is legitimate and lawful, it should not necessarily be "discouraged".

Finally, a few words about control.

After describing this mechanism, Zuckerberg stressed that Facebook actually aimed to give people “more control of what they see”. This mechanism will be on by default, he explained, but “For those who want to make these decisions themselves, we believe they should have that choice since this content doesn't violate our standards.” He also stated that: “Over time, these controls may also enable us to have more flexible standards in categories like nudity, where cultural norms are very different around the world and personal preferences vary.”

This default settings solution reveals the surface of the design-based richness and complexity of the content moderation practices carried out by Facebook (most of which are not likely to be heard by the board). In many cases, default settings become the fixed ones, because, alongside other reasons, users are not aware of them. In this particular instance, even if users become familiar with such default settings, they might be reluctant to change them, due to the way Facebook describes the “discouraged content”. After all, who wants to take an active action to view “provocative” or “sensational” content? Therefore, designing this default solution in connection with “borderline content” will definitely provide control, but not to the users - rather, to Facebook itself. Last, we should recognize the danger that is located in Facebook’s willingness to have “flexible standards” for “controlling” the “borderline content” that is seen, and make sure that such flexibility is not an easy and accountability-free measure to abide by pressure groups’ demands.

To recap, the new board is a positive step, but misses out troubling new and opaque fashions of censorship. We have to make sure that Facebook’s new plan does not distract us from the full range of the content moderation practices it carries out, and insist on creating mechanisms to holding the company accountable for them.

Noa Mor
University of Haifa
nmor12@campus.haifa.ac.il