Interview

Prof. Catalina Botero Marino, Erik Tuchtfeld*

Quasi-Judicial Oversight Mechanisms for Social Platforms

– A Conversation with Catalina Botero Marino, Co-Chair of the Oversight Board –

Content regulation on social networks is controversial. Some would like to counter misinformation and hate speech by demanding that Facebook (now Meta)[1], Twitter and others, take a more active role in curating content posted on their platforms. Others fear censorship and arbitrariness. In this stormy debate, Meta decided to share responsibility for enforcement of its community guidelines by creating the Oversight Board (OB). This is an independent body for reviewing Meta's decisions on content moderation, which steps in when users appeal such decisions. It is free to pick cases that it considers crucial and of global significance, while Meta itself can refer "significant and difficult" cases to the OB. In a way, then, the OB serves as "Facebook's Supreme Court" – as it is colloquially known.

Professor Botero, you are a member of the Oversight Board (OB), the "Facebook Supreme Court", as some call it. Users who think that Meta unjustifiably removed their posts or blocked their accounts on Facebook or Instagram can ask the OB to

* Professor Catalina Botero Marino, a Colombian human rights scholar, holds the UNESCO Chair on Freedom of Expression at Universidad de los Andes. In addition, she holds an adjunct professorship at American University Washington College of Law's Academy on Human Rights and International Humanitarian Law, and advises Columbia University's Global Freedom of Expression and Information Project. From 2008–2014, she served as Special Rapporteur for Freedom of Expression for the Inter-American Commission on Human Rights (IACHR). In February 2020, Catalina Botero Marino was named one of the first four members of the Oversight Board, which she now co-chairs. In 2021, she was a visiting scholar at the Max Planck Institute for Comparative Public Law and International Law, Heidelberg. The interview was conducted by Erik Tuchtfeld, research fellow at the Max Planck Institute for Comparative Public Law and International Law, and board member of the German think tank D64 – Center for Digital Progress.

1 Facebook, Inc., the parent company of *inter alia* Facebook, Instagram and Whatsapp, was rebranded as "Meta" in October 2021. Accordingly, "Meta" refers in this interview to the company, "Facebook" to the concrete social media service.

 DOI: 10.5771/2699-1284-2021-3-254

review the company's decision. Let's talk about institutional set-up for a moment: How many "judges" does the OB consist of, and how were they (s)elected?

Currently, the Oversight Board has 20 members. Over time, it will grow to a maximum of 40 members. The Board was established through an extensive global consultation that took more than a year and involved hundreds of individuals and organizations with expertise in freedom of expression and platform governance. The first four members, including myself, were selected by Facebook (now Meta) and serve as the Board's co-chairs. We played a decisive role in selecting the other 16 members. We are in the process of increasing the number of members to increase diversity, which involves careful evaluation. Once we complete the first board composition, Meta will not be involved in the selection of new members anymore. Instead, the sitting members will select the new ones.

Meta – similar to other big social media companies like Twitter and YouTube – acts globally. In your opinion, does the OB's composition reflect the global community?

It is not easy to reflect the diversity of the billions of users on these platforms. The Board cannot realistically include people from every country, language group and culture. However, diversity – of thought, background, and experience – is essential for selecting members. At the moment, the Board is composed of people from all regions (North and South America, Asia Pacific, Oceania and Southeast Asia, Middle East and North Africa, Sub-Saharan Africa, Western and Eastern Europe). As we select new members, we are seeking greater sub-regional representation and diversity.

Since the OB reviews decisions made by another body, it is tempting to compare it to judicial institutions. Would you call the OB a "court"? Are members of the OB independent from Meta?

These are two different questions. With regard to independence and autonomy, the Board's governing instruments (the Charter and Bylaws) establish institutional, functional, budgetary, and personal guarantees that members can act with complete autonomy from the economic, political, or reputational interests of the company. For example, we operate with a non-revocable endowment of US$130 million that is administered by a trust independent of the company; our appointment is for a fixed term that the company cannot interrupt; the Board's administration is completely independent; and members do not depend in any way on the company. Beyond these institutional elements, the decisions that we have taken thus far – the majority of which overturn those made by Meta – clearly show that we are not shy about holding it to account.

The other question is whether the Board is a "court". Although this metaphor can generally explain our function and the conditions under which we operate, I actually

prefer to speak of an external, autonomous supervisory board as a self-regulatory mechanism. This description is longer and less sexy – but more fitting to our nature.

The "Oversight Board Charter", the OB's foundational document, grants the Oversight Board power to determine whether the decisions of removing or maintaining content "were consistent with Facebook's content policies and values". While the Charter also refers to "human rights norms protecting free expression", the main yardstick for the assessments of Meta's decisions seems to be the company's own policies, particularly its Community Standards. Would you agree? Is it Meta's laws that apply in the "state of Meta", or do international human rights norms take effect?

The founding instruments of the Board refer to international human rights law, and the company itself has committed to comply with the UN Guiding Principles on Business and Human Rights (UNGPs), endorsed by the UN Human Rights Council in 2011. In fact, just in March this year the company announced[2] its Corporate Human Rights Policy[3] in accordance with the UNGPs. In that sense, we understand that the company's procedures, its internal guidelines and its decisions must respect international human rights law. The UNGPs create a framework for the human rights responsibilities of private businesses, and the Board's analysis in all our cases is informed by those principles. You will see in our decisions: We have held that the Community Standards applicable to the respective case did not respect international human rights law and have, in some cases, overturned FB's original decision for this reason.

Let me give you a concrete example. Some of the rules in the Community Standards are really ambiguous and do not send a clear signal to users about what is permitted and what is prohibited in the community. In doing so, they depart from the principle of legality which is one of the essential elements of international human rights law on freedom of expression. When the user's behavior is not reasonably prohibited, we have reversed the company's decision and recommended that it bring its Community Standards in line with the principle of legality.

However, your question prompts a more fundamental reflection: The international human rights system cannot simply be transplanted to the private sphere without further consideration. Businesses have legitimate rights – interests which States do not have (such as for-profit motives, for example). When evaluating a decision on content moderation, it is essential to take this into account. The right to corporate autonomy cannot simply be removed from the equation. How far does this autonomy – reflected in the Community Standards and their application – go, and where does the international human rights system come into play? That is one of the most challenging questions the Board faces.

2 https://about.fb.com/news/2021/03/our-commitment-to-human-rights, last accessed on 12 December 2021.

3 https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf, last accessed on 12 December 2021.

How would you describe the relationship between the OB and national courts in deciding whether content was lawfully removed or not? Will the first supersede the latter in the long run?

Great question. First of all, cases which render binding judicial decisions for the company do not reach the Board. If we make a decision, and an independent judge holds otherwise, it is for the company to decide – not for the Board. As far as I know, this has not occurred yet.

I do believe that, in most countries, the law has not given courts sufficient tools for the moderation of content online. Doing so is complex, because of the global scale of the platforms, the impact that a judicial decision can have on the architecture of the Internet, and the fact that we have more actors than the two parties which are traditionally confronted in a conflict regarding freedom of expression. This complexity means that the weighing of conflicting rights and interests cannot simply be carried out with the instruments that exist to date to resolve conflicts between those who express themselves and those who feel affected by such expression.

I believe that one of the Board's fundamental tasks is to build a doctrine which is coherent and consistent with international human rights law (and, therefore, with democratic legal systems). It would help develop a doctrine that will serve judges and other operators when making decisions on moderation of digital online content.

The discussion about moderation of content, particularly hate speech, is currently focused on the most restrictive measures such as blocking of user accounts or removal of content. However, social networks also employ measures like demonetization[4] or warning messages.[5]. How do you evaluate their impact on freedom of speech? Will the OB assess such "softer" approaches, too?

Another key principle of international human rights law that the Board examines is the principle of necessity and proportionality. According to this principle, when an expression may cause disproportionate harm (there are harms that we are obliged to bear) to a fundamental right of a third party, it is legitimate to restrict that expression. However, the restrictive measure must be necessary, i.e. it must – among the ones which satisfy the desired purpose – be the least costly for freedom of expression. This is where alternative measures play a fundamental role in the Board's decision: if the harm can be avoided with a label or any other of the aforementioned measures, then such a measure should be preferred over the removal of the content. There are cases where content removal is indispensable, but this conclusion must be reached only after ruling out less costly measures for freedom of expression, such as those you rightly mention.

4 YouTube shares advertising revenues with content creators based on the success of videos and the revenue it created. When videos violate certain criteria, the content creators might be banned from their advertisement program (YouTube Partner Program).

5 For example, Twitter has marked then-President Trump's remarks on the result of the 2020 presidential election as "disputed" and "misleading".

When it comes to "harmful content", the "Facebook Files" by the former Meta employee Frances Haugen have brought some insights from Meta's own research[6] into light which shows that Meta's photo and video sharing service Instagram is dangerous for the mental health of its young users, especially teenage girls. Haugen also stated[7] Facebook is "literally fanning ethnic violence" by promoting harmful content during ethnic tensions, e.g. in Myanmar and Ethiopia. Do you see any chance that the Oversight Board will also look into these issues which are not about blocking lawful content but about promoting dangerous posts? How should a social network, in your opinion, treat legal but harmful content? Should it differentiate between these categories?

What to do with legal, yet harmful content is exactly the space in which the Board operates. In fact, the Board is currently reviewing a case related to the conflict and ethnic tensions in Ethiopia and we've already issued two decisions related to similar issues in Myanmar. We are constantly examining the relationship between content on the platform, particularly hate speech, and the potential for imminent harm. As I have mentioned early in this conversation, in cases where we have found that content can actually cause harm in people's real lives, we have found it essential to take steps to restrict that content. Currently, the Board has the scope to review cases where content has already been removed by the company and where the content remains on the platform and users appeal to us to have it removed. There is an incredible amount of nuance to this work, but an important first step for social media companies is to become more transparent. We believe that the Oversight Board's method of transparent moderation, that looks to ensure platform accountability, could be highly effective in promoting change within platforms regarding their moderation of legal yet disproportionately harmful content and we'll keep pushing in this direction. We will publish transparency reports after each quarter ends and annual reports, to provide a detailed qualitative assessment of how the company is implementing the Board's case decisions and recommendations.

Meta's Vice President for Integrity, Guy Rosen, recently clarified[8] that if "something might be hate speech but we're not confident enough that it meets the bar for removal, our technology may reduce the content's distribution or won't recommend groups, pages or people that regularly post content that is likely to violate our policies." This policy is called "shadow-banning" and also addresses

6 https://www.wsj.com/articles/the-facebook-files-11631713039, last accessed on 12 December 2021.

7 https://www.theguardian.com/technology/2021/oct/07/facebooks-role-in-myanmar-and-ethiopia-under-new-scrutiny, last accessed on 12 December 2021.

8 https://about.fb.com/news/2021/10/hate-speech-prevalence-dropped-facebook, last accessed on 12 December 2021.

the issue of (potentially) harmful but not (clearly) illegal content.[9] It is criticized because users are mostly unaware of being targeted by these measures and, in consequence, are unable to challenge them. Would you say that more transparency when executing such actions is needed, e.g. a notification by Facebook or Instagram that one's visibility is reduced? Also, these measures are often taken automatically by algorithms, without any human interference. What is your opinion on such automatic content moderation?

First, on transparency: the Oversight Board strongly advocates that the company provides as much transparency as possible to users when making decisions that impact their content. Most of our recommendations are aimed at substantially increasing the transparency of the company. Second, on algorithms: It's important to mention that our mandate is to review and improve Meta's content moderation decisions but it´s not within the scope of the Board to review its coding and algorithms, or their business model. However, there are clearly areas where our decisions and recommendations will engage with the company's design decisions and algorithmic treatment of content, for example on changes to classifiers or automated enforcement tools. We are already seeing that content decisions and the algorithmic treatment of content cannot be separated. For example, in our decision on a case regarding adult nudity in a post about Breast cancer awareness, the Board recommended that the company improve the automated detection of images with text-overlay to ensure that posts raising awareness of breast cancer symptoms are not wrongly flagged for review. In that case, the Board also recommended that it assures more and best human moderation and expand transparency reporting to disclose data on the number of automated removal decisions per Community Standard, and the proportion of those decisions subsequently reversed following human review.

In general, freedom of speech is an individual right against governmental interference. Recently, we have seen a fierce discussion on the treatment of governmental institutions on social networks. Then-President Trump's social media accounts were famously blocked, but there are other governmental accounts, for example in Nigeria and Myanmar, which have been shut down. Current developments in Afghanistan might lead to similar actions there. Should social media platforms be allowed to decide on the dissemination of governments' information? Might they even have a responsibility to limit the potential damage which irresponsible governments can inflict?

On the one hand, international human rights law establishes special limits for public officials and government institutions. Since Meta has committed to respect international human rights law, it must abide by these rules. The Board recognized this in some

9 "Shadow banning" describes a technique used by social networks to not remove content, but to limit the visibility, e.g. by making it more difficult to find via search functions (such as hashtags).

of its decisions such as one on Mr. Trump's account[10] and in a more recent decision regarding a medical council in Brazil[11]. In the first decision, the Board recommended that the company, "*escalate content containing political speech from highly influential users to specialized staff who are familiar with the linguistic and political context and who are independent to the interest of the company; dedicate adequate resourcing and expertise to assess risks of harm from influential accounts globally; produce more information to help users understand the application of the newsworthiness allowance, including how it applies to influential accounts*". However, one must tread carefully when talking about intermediary liability. This liability cannot lead to the assertion that a company should eliminate the assessment of the importance of the public interest implicit in a statement. Nor can it lead to an affirmation that companies must follow, in real time, everything that public authorities post. The latter rule, for example, would simply do away with the internet as we know it today and would by default end up in the removal of any controversial content.

You are from Colombia and have been the Special Rapporteur for Freedom of Expression for the Inter-American Commission on Human Rights (IACHR). Would you say there are problems in the field of content regulation which particularly affect people in the Global South?

Certainly! The large platforms have a clear orientation towards the Global North. The rules are mainly in English and the translation into other languages does, in many cases, leave much to be desired; the rules governing the community are based on US legal culture; training is conducted by people immersed in this culture; and until very recently the political, cultural, or social context was practically neglected when it came to moderating content. This is particularly serious for the Global South. Unlike in the Global North, the population in many countries of the Global South is informed and communicates exclusively through one particular platform, and the moderation of out-of-context content can simply generate unacceptable barriers to information from a human rights point of view. Let me give you an example: the platform's community standards ban adult nudity and especially female nipples, with some exceptions. In many indigenous communities in Africa or Latin America (just to mention two regions) women live with naked torsos and, consequently, all content with images of these communities was removed. Yet, some of their content denounced, for example, serious environmental damage or gross human rights violations. In that sense, marginalized communities who need to communicate in order to denounce violations of their rights were deprived of all social power, completely silenced, made invisible and ultimately erased from the digital sphere. As if they did not exist. This has

10 https://www.oversightboard.com/decision/FB-691QAMHJ, last accessed on 12 December 2021.

11 https://www.oversightboard.com/decision/FB-B6NGYREK, last accessed on 12 December 2021.

been improving but there is still a long, long way to go. The Board plays an important role in this process – we carefully examine the local context of content.

The Oversight Board is the first of its kind. There is no similar mechanism for YouTube or Twitter. Do you think the OB serves as a role model – and that other platforms are going to implement such private adjudication bodies as well? Are there any efforts to have a common body overseeing all these networks?

The Board is one model of independent and impartial self-regulation. We are focused on something which other efforts are not – steering Meta to act in accordance with human rights standards, and improving the company's policies in response to problems we are seeing right now on their platforms. However, this model is not the only one possible. It may even prove to be very difficult to adopt for smaller platforms. I believe that the existence of other models and independent oversight bodies would enrich the landscape and allow for increased dialogue on best practices in the difficult field of global content moderation. By no means do I consider that this should be a task solely for the platforms or a single overarching body. However, I do think it is good practice for other models to include minimum guarantees of independence and to operate on the basis of the rules of the international human rights system.

Zusammenfassung: Die Verbreitung von Desinformation und Hassrede hat zu Aufforderungen an die sozialen Netwerke geführt, eine aktivere Rolle bei der Moderation von Inhalten einzunehmen. Andere befürchten dagegen Zensur und Willkür. Inmitten dieser heftigen Debatte hat Meta sich dazu entschieden, die Verantwortung für die Durchsetzung seiner Gemeinschaftsstandards zukünftig mit dem Oversight Board (OB) zu teilen, einer unabhängigen Institution, welche Entscheidungen von Meta über das Entfernen von Inhalten oder das Sperren von Nutzerinnen und Nutzern überprüfen kann. Erik Tuchtfeld bespricht mit Catalina Botero Marino, der ehemaligen Sonderberichterstatterin für Meinungsfreiheit der Interamerikanischen Menschenrechtskommission (IAMRK) und Ko-Vorsitzenden des Oversight Boards, wie sich das Board zusammensetzt und inwieweit es die globale Natur der Meta-Plattformen Facebook und Instagram widerspiegelt. Das Interview thematisiert die Beziehung zu und die Parallelen zwischen nationalen Gerichten und dem OB sowie die Bedeutung, die internationale Menschenrechtsinstrumente, der Grundsatz der Verhältnismäßigkeit und öffentliche Äußerungen von Amtsträgerinnen und Amtsträgern für die Arbeit des Boards haben.

Summary: The spread of misinformation and hate speech has led to calls for a more active role of social networks when it comes to curating content. Others fear censorship and arbitrariness. In this stormy debate, Meta decided to share the responsibility for the enforcement of its community guidelines by setting up the Oversight Board (OB), an independent body which can review decisions taken by Meta on the removal of posts and blocking of users. Erik Tuchtfeld interviews Catalina Botero Marino, the former Special Rapporteur for Freedom of Expression for the Inter-American Commission on Human Rights (IACHR) and current co-chair of the Oversight Board, on how the Board is composed and in how far its composition reflects the global nature of Meta's platforms such as Facebook and Instagram. The interview deals with the relationship and parallels to national courts. Also, the importance of international human rights instruments, the principle of proportionality and particular difficulties with public statements by government officials are discussed.