INTERNETLAB

Leveling the Playing Field:

Achieving Fairness and Transparency in Content Moderation on Digital Platforms

Authors: Francisco Brito Cruz (coord.), Alice Lana e Iná Jost

Collaborator: Heloisa Massaro

InternetLab, 2023

About InternetLab

InternetLab is an independent research center that aims to foster academic debate around issues involving law and technology, especially internet policy. Our goal is to conduct interdisciplinary impactful research and promote dialogue among academics, professionals and policymakers. We follow an entrepreneurial nonprofit model, which embraces our pursuit of producing scholarly research in the manner and spirit of an academic think tank. As a nexus of expertise in technology, public policy and social sciences, our research agenda covers a wide range of topics, including privacy, freedom of speech, gender and technology.

Objectives of this document

This research project by InternetLab aims at contributing to the public conversation around content moderation within digital platforms. We seek to untangle layered moderation systems, those that bring additional layers of qualified analysis to certain types of content or profiles when determining which pieces of content should remain or be removed from the platforms.

We based our study on some questions, such as:

- *Should platforms' content moderation contemplate additional layers for content moderation regarding different types of profiles or content?*
- *If certain people or content will be treated differently by platforms and their content moderation processes, what framework should be used to ensure the efficiency and legitimacy of these systems??*
- *How should these systems be designed in order to protect user's rights, especially concerning fairness and transparency?*

The aim of this document is to present the concept and nuances of layered moderation systems and to issue recommendations aligned with human rights, fairness and transparency, instead of creating privileged bubbles or VIP lists that enjoy different sets of standards when publishing content online due to exclusive business-oriented needs.

Note: this research was supported by the Global Network Initiative through its *Emerging Voices Fellowship Program*.

1. *Introduction*

In 2019, the Brazilian football player Neymar posted on his Facebook and Instagram accounts nude images of a woman in a private conversation without her consent. The posts were part of the strategy the athlete designed to publicly respond to a rape accusation. Although Meta's policies forbid the publication of nonconsensual intimate imagery, the content remained on the platform for over 24 hours, being viewed by around 56 million people.

Neymar's episode exemplifies a *modus operandi* that would be confirmed two years later. In September 2021, the Wall Street Journal published a story revealing the existence of a system developed by Meta that added an additional layer to the content moderation process on its platforms. The mechanism, called *Cross-check program* by the company, provides for a different scrutiny for specific users, such as elected politicians, significant business partners, number of followers, among others. In practice, when profiles that belong to the list submit content flagged as potentially infringing, their posts are directed to a different queue, overviewed by a specialized team, instead of the regular moderation one.

A helpful analogy is the boarding line at the airport. Everyone agrees that the elderly and people with babies should board first. But what if the line, in practice, mainly applied to "premium customers"?

The disclosure of Meta's *Cross-check* raised several questions regarding the justification and legitimacy of such systems. Implementing such mechanisms raises concerns about transparency, equal treatment, and risks to fundamental rights. Should layered moderation based on users' lists exist? Would they distort or promote fairness and transparency in the platforms' operation? If they produce any positive effects, what would be the best parameters for them to be deployed?

2. *"Layered" content moderation: concept and cases*

2.1. *Freedom of Expression and Content Moderation Systems*

Before deepening the features of a layered moderation system such as Meta's *Cross-check*, it is important to rewind to set out a common ground of definitions.

As an operating definition used by InternetLab on our approach to the topic, *content moderation* refers to a key activity for a digital platform: elaborate and apply rules, procedures, and systems to remove, limit reach, label content, and suspend or remove accounts[1], as well as "*platforms' systems and rules that determine how they treat user-generated content on their services*"[2]. This exercise is, at the same time, both the management of an individual user's expressions and a part of the product and value that platforms can offer to the other users.

The activity of moderating content poses a logistical challenge to platforms, since they deal with an immense amount of content and multifaceted contexts. This is well established in literature that approaches its key challenges, and argued by scholars from different perspectives. There are researchers that consider that artificial intelligence could present an effective response to the massive scale of data and the constant state of violations. There are others that defend the existence of a structure of systematic decision-making, one that goes beyond the logic of individual evaluations, seeking to avoid the incapacitation of the services' operation[3].

---

[1] Thiago Dias Oliva, Victor Pavarin Tavares e Mariana G. Valente, "Uma solução única para toda a internet? Riscos do debate regulatório brasileiro para a operação de plataformas de conhecimento", Diagnósticos & Recomendações (São Paulo: InternetLab, 2020). Pg. 11 Available: https://internetlab.org.br/wp-content/uploads/2020/09/policy_plataformas-conhecimento_20200910.pdf

[2] Doeuk, Evelyn. Content Moderation as Systems Thinking. (Harvard Law Review, 2022). Pg. 528. Available: https://harvardlawreview.org/print/vol-136/content-moderation-as-systems-thinking/

[3] Ibid, pg. 551.
Gillespie, Tarleton. Content moderation, AI, and the question of scale. (Big Data & Society, 2020). Pg. 2-4. Available: https://journals.sagepub.com/doi/pdf/10.1177/2053951720943234
Gillespie, Tarleton. Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. (Yale University Press, 2018). Available: https://yalebooks.yale.edu/book/9780300261431/custodians-of-the-internet/
Klonick, Kate. The new governors: the people, rules, and processes governing online speech. (Harvard Law Review, 2017). Available: https://harvardlawreview.org/wp-content/uploads/2018/04/1598-1670_Online.pdf
Suzor, Nicolas. Lawless. The secret rules that govern our digital lives. (Cambridge, University Press, 2019). Available: https://www.cambridge.org/core/books/lawless/8504E4EC8A74E539D701A04D3EE8D8DE

Still seeking alternative solutions for the mass speech administration, one could say that layered moderation systems[4] could be one strategy employed by the companies to mitigate risks to human rights, since it gives an analysis' priority to a few types of users or content that should be carefully reviewed for protecting specific kinds of speech. It makes sense, for example, that activists or journalists have their expression more carefully evaluated than regular users, as their words have a different audience reach and impact, and their accounts and discourse could be constantly under strategic targeting by antagonists.

For instance, the accounts and discourse published by human rights defenders and journalists' tend to be - potentially, more than other civil actors - the target of attacks and harassment, which could effectively translate into intimidation and a tentative silencing of their voices. Sometimes, these kinds of threats can even pose significant risks to their safety and well-being. Therefore, protecting their speech and accounts with analysis prioritization could be an interesting approach to promote their safety.

In other words, ideally, layered moderation can be a tool that creates fairness inside an large-scale speech management process, functioning as an attempt to mitigate distortions created by the regular and industrial moderation processes by platforms.

But what if the layered moderation serves only to preserve business partners and commercial interests? What if the rules of the system are unclear and its gearing ends up promoting more inequality, contrary to the protection of human rights?

2.2. Layered moderation in practice: The Cross-check example

The existence of systems that offer different treatment to some users is certainly not unique to Meta, but, as mentioned before, the scoop published by the Wall Street Journal in 2021 revealed important details of this program, as well as the gap around transparency about those systems among the industry[5].

The system implements privileged levels of analysis for specific accounts - which Meta determines as "*especially susceptible to the risk of experiencing actions resulting in false*

[4] The term "layered moderation" is employed to address a type of content moderation that provides for a difference in treatment by the platform depending on the user or the content. This difference contemplates other layers of content verification that can add, for example, a stage of human analysis for certain cases. What we discuss in this policy paper is whether the system should exist and how it should be designed in order to protect speech rather than protecting interests that are not committed to freedom of expression.

[5] Horwitz, Jeff. Wall Street Journal. "Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt". Available: https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353

positives[6]" - based on criteria such as the type of user or entity (politician, journalist, significant business partner, human rights organization), number of followers or topics addressed by the entity. To reduce discretion, only a select group of Meta employees can add new entities to the list, which is regularly audited.

When users that belong to the special list have a content flagged as potentially infringing, they are directed to the *Cross-check* queue instead of the regular moderation one[7]. The prioritization criteria for analyzing these pieces of content are "topic sensitivity (how trending/sensitive the topic is); enforcement severity (the severity of the potential enforcement action); false positive probability, predicted reach, and entity sensitivity[8].

Following the disclosure, in October 2021, Meta's Oversight Board (OSB) accepted a request from the company to review *Cross-check* and make recommendations for its improvement. One year later, the body released a policy advisory opinion bringing key findings and guidance to ameliorate the system[9]. In general terms, the OSB concluded that, by providing unequal treatment for some users, *Cross-check* (i) caused a delay when removing violating content posted by the ones on the list; (ii) failed to track and disclose the metrics employed by the system; (iii) lacked transparency around its functioning. According to the Board, "while there are clear criteria for including business partners and government leaders, users whose content is likely to be important from a human rights perspective, such as journalists and civil society organizations, have less clear paths to access the program."

Among other recommendations, the Board suggested that the company should prioritize expression that is fundamental to human rights, as well as increasing transparency around *Cross-check*'s operation and damage reduction measures by content left up during the layered moderation process - which tends to be delayed. A summary of the 32 recommendations the Meta's Oversight Board published about the program in their policy advisory opinion can be found below.

---

[6] Meta defines false positives as the mistaken removal of content that does not violate the content policies that establish what is allowed on Facebook and Instagram. Pg. 6. Available: https://www.oversightboard.com/news/501654971916288-oversight-board-publishes-policy-advisory-opinion-on-meta-s-cross-check-program/

[7] The whole *Cross-check* operation is detailed in the Policy Advisory Opinion issued by the Oversight Board. Pg. 9-21. Available: https://www.oversightboard.com/news/501654971916288-oversight-board-publishes-policy-advisory-opinion-on-meta-s-cross-check-program/

[8] Ibid. Pg. 19.

[9] The OSB received 87 public comments related to this policy advisory opinion: nine from Asia Pacific and Oceania, two from Central and South Asia, 12 from Europe, three from Latin America and the Caribbean, three from the Middle East and North Africa, three from Sub-Saharan Africa, and 55 from the United States and Canada. Available on: https://internetlab.org.br/wp-content/uploads/2023/05/Public-comments-appendix.pdf

What questions were posed by Meta to the Board?	"*1. Because of the complexities of content moderation at scale, how should Facebook balance its desire to fairly and objectively apply our Community Standards with our need for flexibility, nuance, and context-specific decisions within cross-check?* *2. What improvements should Facebook make to how we govern our Early Response ("ER") Secondary Review cross-check system to fairly enforce our Community Standards while minimizing the potential for over-enforcement, retaining business flexibility, and promoting transparency in the review process?* *3. What criteria should Facebook use to determine who is included in ER Secondary Review and prioritized as one of many factors by our cross-check ranker in order to help ensure fairness in access to this system and its implementation?*"

First axis: Human rights and public interest considerations (enforcement)		
Type	**Recommendation**	**Justification**
Prioritize human rights/public interest expression	Inclusion of users likely to produce expression important to human rights or special public interest to X-Check's prioritized list. Separation of these users from Meta's business partners (or business priorities) included in the list. Guarantee that the pathway and decision making structure for this content is devoid of business considerations.	Avoid direct competition for limited review resources from Meta.
Process of inclusion	Informing members they have been included in the list and providing opt-outs if they so desire. Require invitees to review Meta's content rules and commit to following them before being added to X-Check. Require acknowledgement of the program's particular rules. Develop a system to inform users proactively of changes to Meta's content policies to facilitate awareness and compliance.	X-Check is viewed as providing benefits to included users. Meta should operate based on principles of user consent, transparency and fairness.
Process of inclusion	Engage with civil society for the purposes of list creation.	Having a multi-stakeholder perspective on privileged moderation systems.
Content-based criteria	Develop content-based criteria to protect content with high risk of erroneous over-enforcement directly, without regard to who posted it.	The current entity-based approach is insufficient to guarantee that important public interest and human rights contents (which may come from any user) is not removed.
Human-rights based system	Develop a second protection system, focused on detecting false positives (content wrongly removed) caused by X-Check and based on a human rights perspective. Prioritize the review order of this content based on the severity of the possible violation, the likelihood of being a false positive, and the likelihood of virality.	An algorithmic ranker for a false positive prevention system could prioritize content based on the types of decisions that are hard for automation and human moderators at scale (e.g., historically over-enforced speech or speech by marginalized communities).

Team specialization	Create specialized teams for list creation to ensure criteria are being met, with the benefit of local input. Public policy teams may nominate candidates, not be final decision makers. Individuals with personal or business relationships with nominated entities should not be decision makers.	Reduce conflict of interests with other teams, such as Meta's public policy teams, who often interact with lobby government actors. Ensure objective application of inclusion criteria.
Auditing of X-Check and removal	Promote yearly review of all included entities in any mistake-prevention system that provides benefits to such entities.	Maintain a standard of eligibility for the X-Check system.
Second axis: transparency considerations		
Type	**Recommendation**	**Justification**
Transparency and application	Establish clear, public criteria for inclusion in X-Check. Allow users who meet these criteria to apply to X-Check.	Enable users to apply for over-enforcement X-Check protections should they meet the company's articulated criteria.
Radical transparency	Include in transparency reports: a. Overturn rates for false positive mistake-prevention systems, disaggregated according to different factors. Publish overturn rates for entity-based and content-based systems, and categories of entities or content included. b. The total number and percentage of escalation-only policies applied due to false positive X-Check relative to total enforcement decisions. c. Average and median time to final decision of X-Check, disaggregated by country and language. d. Aggregate data regarding any lists used for X-Check, including the type of entity and region. e. Rate of erroneous removals (false positives) versus all reviewed content, including the total amount of harm generated by these false positives measured as the predicted total views on the content (i.e., overenforcement) f. Rate of erroneous keep-up decisions (false negatives) on content, including the total amount of harm generated by these false positives, measured as the sum of views the content accrued (i.e. underenforcement)	Third parties may tell whether the program is working effectively.
Publicizing of users	Publicly mark accounts for some categories of entities protected by X-check (i.e. state actors, political candidates and business partners).	Allow third parties to hold privileged users accountable for upholding commitment to the rules.
Prioritize human rights/public interest expression	Never publicize beneficiaries who are human rights defenders. Provide them with opt-in for public identification. Use the data compiled by Meta to identify "historically over-enforced entities".	Avoid harm arising from historical over-enforcement.
Appeal rights	Ensure that X-checked content can be appealed to the Oversight Board, when applicable, regardless of whether the content	Provide an alternative route for appeals out of undue application of

	reached the highest levels of review within Meta.	X-check.
Enhancement of X-Check	Publishing reports on metrics on adverse effects of delayed enforcement (i.e. publicize views accrued on violating content that was preserved due to X-Check). Determine a baseline for these metrics and report on goals to reduce them.	Error indicators should help Meta and third parties come up with solutions to increase correct content removals in the future or question the expansion of the system.
Researcher information	Create a channel in which researchers obtain non-public anonymized data about X-Check for public-interest investigations and provide recommendations for improvement.	Specialized researchers may tell whether the program is working effectively and contribute to its improvement.
Third-party audits	Promote external audits, by the Oversight Board or third parties (e.g., researchers or civil society) with anonymized and aggregate data.	Assess whether a mistake-prevention system mitigates negative human rights impacts
Third axis: reduction of damages		
Type	**Recommendation**	**Justification**
Alternative penalties	Consider alternatives to removal such as downranking, slowing the virality, hiding, or temporarily removing posts.	Reduce damage from the prompt removal of potentially violating content.
Prioritize human rights/public interest expression	Enable reviewers to conduct a cultural and linguistic analysis of texts, considering national, regional or local contexts. Provide skilled reviewers with the ability to take further context into consideration, regardless of whether the review is entity-based or content-based.	The Early Response Team does not require its reviewers to have cultural or linguistic expertise (even in high-risk regions).
Overturn rates	Use the rate of decision overturns to inform whether to default to the original enforcement within a shorter time frame or what other enforcement action to apply pending review.	Review decisions based on rate of error (overturn rates). If errors are consistently low for certain policy violations or certain languages, Meta needs to calibrate how quickly and how intrusive an enforcement measure it should apply.

3. Research to build a framework for assessing layered content moderation

The exercise of content moderation is a fundamental one for the functioning of platforms, and has many aspects that open avenues for research, especially because of its impact in the circulation of speech. In early 2022, InternetLab started to carry out research looking at layered systems in content moderation, seeking to create frameworks to help assess whether such a system is necessary, and its limits, mechanisms, guarantees, and safeguards for human rights. If the tool is important to tackle moderation's logistical challenges and even other politically sensitive issues, how should it be designed to not pose significant risks to fundamental rights, and, actually further human rights?

Furthermore, our research had a particular interest. Besides understanding its necessity and discussing transparency parameters, we wanted to use a regional lens to deepen the advantages and disadvantages of its application in specific social, political, economical and cultural contexts, for example, in Latin American countries.

We then conducted a series of focus groups with Latin American stakeholders whose opinion on content moderation would be helpful. Our main goal was to identify the central issues posed by layered moderation systems from diverse perspectives, and to discuss policy alternatives to build healthy guidelines. The material was compiled and the main conclusions are exposed below. After deepening these findings, we then break down our findings into two perspectives: the optimist's view and the pessimist's view, or the glass-half-full approach and the glass-half-empty approach.

3.1. Research in focus groups: method and conclusions

Two focus groups were initially conducted with different types of stakeholders. Participants were selected across sectors with presence in the online environment, taking into account markers of class, gender, race and LGBTQIA+ and aiming for parity. Both meetings were held under Chatham House Rule, to assure that everyone would feel comfortable to speak freely.

In the first group, we invited seven people from Latin America who study or act in the fields of election integrity, disinformation and journalism. We also invited people who we identified as influencers in the online environment. The second group was also composed of seven people, also from the region, who study or act in the fields of digital rights, both from academia and civil society. The two sessions were divided into three stages: (i) individual experience sharing; (ii) questioning about layered moderation systems; (iii) proposals for the future - guided by the following questions:

Individual experience sharing	1) Experiences (lived or observed) about content moderation, especially false negatives and false positives.
	2) How was the response from the platform? Did it hinder or help? How could it have been better, considering the amount of moderation that must be done daily?
Questioning about the system	3) Is a Cross-check-like system needed? For what/whom?
	4) Does this increase or undermine the protection of freedom of expression and other human rights?
Proposals for the future	5) What criteria should define which type of content to be cross-checked (reach, subject, any other)?
	6) What criteria should define which accounts should be cross-checked (number of followers, subject matter, any other)?
	7) How and by whom should these criteria be defined and updated?

After the sessions, all the participants were invited to present written contributions about their perceptions around the risks and legitimacy of these systems. The following pages reflect the outcomes of these discussions. It is important to point out that we chose to bring only quotes in the first section because it relates directly to participants' individual experiences. In this particular portion, we wanted to preserve their first-hand perceptions about the matters discussed, since we believe in the importance and richness of their voices and contexts to this research.

The two chapters that follow expose arguments employed to justify the existence of a layered moderation system, as well as proposals to make it a tool that is transparent at the same time that promotes fairness and equity within platforms.

3.1.1. Individual experience sharing about content moderation

Perceptions about the importance of context

- *"In Latin America, there is no awareness that you can't post any content because it is a private platform. Users don't even know an appeal mechanism exists, in case of blocking. Especially in journalism, we need to understand the context of the language, which may even include words that are prohibited by the platform but used in other contexts."*
- *"I am an inhabitant of a small country, and our context is less valued and considered in the company's analysis because moderators and policies are not involved or aware of the context."*
- *"In 2016, we created an app that people of any color can buy from black producers from various places in Brazil. It was taken down because a law professor said he would open a representation at the Public Prosecutor's Office for "equity violations". This post went viral, so the platforms removed the app's publication. We also have a Facebook group for black people who discuss social and political issues. Within the group, some people have not reflected on political positions, but they manifest themselves in the group because they consider it a welcoming, safe space. But Facebook perceived many issues discussed as aggressive. Facebook has a very difficult time moderating diversity and especially in a community that is diverse from each other.*

Lack of transparency mechanisms

- *"Platforms also deleted hashtags used in the context of protests in Colombia and posts with that content. Transparency is also important. We don't receive factual information from the platforms about the reason for the removals, making it difficult to question the platform's decision."*

Responsiveness of platforms

- *Contacting the platform is difficult when you are a small creator; it takes weeks for a response. Sometimes there is not even a response from the platform, and the creator's work is hampered by being demonetized without justification."*

- *On Instagram, a Brazilian television host with 7 million followers, said that the LGBTQIA+ community is disgraced and that it must be horrible to have an LGBTQIA+ child and not be able to kill them. This content stayed on the platform for a long time.*

> *We demonstrated that advertisers were still supporting and helping to monetize that content. In an Instagram and Facebook post, we explained why the content was problematic and made a complaint, criticizing the hate content we were denouncing. Within minutes, the Facebook post was removed. We contacted Meta and were not successful in the dialogue. After the LGBTQIA+ National Alliance, a partner of Meta, got in touch with the company, Meta restored the post – but the campaign had already lost engagement. Importantly: the original hateful content reported remained on the platform. So this is the appeal: to be more careful in moderating the content of reports.*

3.1.2. Questioning about the system

When asked about the necessity of having a layered moderation system, participants stressed the fact that it may be of public interest to treat some actors differently based on specific criteria. However, for the structure to meet its purpose, its justification and standards have to be transparent and public. The problem highlighted is then the lack of transparency, since the mechanism is not publicized. These nuances have to be weighed, because there are cases in which such form of privileged treatment is effectively necessary to preserve certain expressions and public debate, as opposed to situations when it would harm users' freedom of speech.

Furthermore, participants of the focus groups mentioned that they are aware that the programs that provide special attention for specific users, frequently based on business interests, exist on several platforms, but informally. This is seen as problematic because the methods employed are not transparent, and, above all, it generates discrimination, meaning the existence of different responses to similar situations, depending on who are the users involved in the propagation of the discourse.

The sessions also brought concerns regarding the economic interests of the platforms in moderation practices when deciding to keep or withdraw pieces of content, since there are certain types of expression, specially from commercial partners, that can impact their reputation or markets, generating profits and losses. How much money does a platform earn when delaying content moderation? These amounts are important to understand if platforms are purposely delaying blockings of inappropriate content from influential public figures, given the high financial return on this kind of content.

With regards to a regional perspective, the sessions brought up considerations about the low availability of data and resources to some countries, as well the lack of regional diffusion and pervasiveness in transparency reports published by platforms. Some of the participants pointed out that there is not enough structured content moderation data per country or in other languages.

For example: how many users covered by a layered moderation system a determined platform have in Mexico? How many moderators per thousand users? What is the difference in investment in content moderation in Colombia and in Germany? It's fundamental to have information about the level of resources invested, in order to analyze the need for layered moderation systems and their extent. How would it be possible to evaluate the impacts of a technology if there's no transparency tools available?

Still related to the importance of context, participants brought reflections about different applications of rules depending on specific regions. Do rules apply for every country? Why do countries have different treatments by platforms when compared to others, for example, when tackling disinformation during electoral periods?

3.1.3. Proposals for the future

When thinking about the criteria employed to define which type of user and content should enjoy a layered moderation system, participants mentioned the need for the platform to commit to the same rigor in disclosing and applying its policies regardless of the region, considering that a global company should have a global capacity to enforce its rules.

It is also fundamental to apply this set of rules with respect for cultural and local contexts and characteristics. The definitions that guide what could or could not circulate in platforms are not universal. Rather, they are culturally biased, based on parameters that apply to certain regions but not others, meaning that the removal of content may end up being unwarranted within specific contexts. A well designed layered moderation system is useful when taking regional nuances into consideration.

Under a layered moderation system framework, the creation of tools such as consultation instances could challenge the difficulties of cultural relativity, bringing checks and balances, and refinement mechanisms. These spaces could gather people from minority communities, represent local audiences affected by the posts, and promote the study of the application of rules to specific contexts.

Furthermore, it is fundamental that platforms publicize the criteria that motivate the inclusion of determined pieces of content and users in layered moderation lists. The perception of participants is that the selection of profiles that participate in programs of layered moderation cannot be exclusively based on the amount of followers and business interests of the platforms. Rather, the lists should contemplate, for example, journalists, minority groups and criteria such as user's speech outreach.

In conclusion, it is fundamental to consider safety and privacy when designing transparency tools regarding a layered moderation system. Participants of the sessions called attention to the fact that the publication of the lists themselves could be harmful, since it would import an unwanted level of exposure, especially in case of people that deserve extra protection, for example, human rights defenders and activists. To this end, criteria and statistical data should be public - gender, race, categories of actors, regions, among others - but not the names that are considered by the system.

3.2. The glass half full

The research led us to consider the necessity of layered moderation systems based on users and/or content, in order to pursue fairness, as opposed to formal equality. It is important to treat unequal individuals in accordance with their inequalities. This is an alternative to in-scale and automated moderation - which has the potential for misinterpretation and mistakes in sensitive cases - especially when seeking to promote human rights by protecting political and minority discourses, public interest journalism and activism.

Furthermore, layered moderation systems give room for us to think about local perspectives. In automated content moderation systems, global rules apply regardless of cultural and local characteristics. In other words, the criteria used to keep or remove content are conceived as universal, ignoring social, cultural, and political realities from other contexts. Having different lists and rules for different users and content can be useful because they take differences into account, consider minorities rights, and represent local audiences that are affected in different ways. Every context has particularities, and we need rules that take them into account.

Supposing that a country has a specific context of violation of a certain right. Defenders and advocates of this right should enjoy greater protection in their speech, especially when they represent minority rights, as opposed to countries that do not have similar issues. The examples vary. There are multiple examples: considering the nudity ban, what does *nudity* mean for a western country, when compared to a Brazilian indigenous tribe perspective?

3.3. The glass half empty

In theory, layered moderation should not change the rules applied, only the enforcement procedures. However, in practice, as shown by the *Cross-check* case, the “special” enforcement can alter the nature of decisions around content since it ends up implementing different outcomes for some privileged individuals. Thus, it can distort a principled and consistent content moderation across the whole range of users and contexts.

Although the concept of implementing a mechanism such as the *Cross-Check* program to protect speech plurality on online platforms is welcomed, its application can pose risks to human rights and potentially shield unfair business practices. On one hand, such a tool can be essential in safeguarding diverse opinions and ideas, but on the other hand, it can also be abused by companies to avoid accountability and neglect their responsibility towards upholding human rights. Additionally, companies may use these mechanisms for public relations purposes, such as shielding their reputation from content moderation scandals.

Moreover, the research shows that there is little attention to the impact of layered content moderation at a regional level. In those contexts, we have noticed a lack of literature and awareness around the issue of the usage of layered content moderation systems in order to counter violence against historically marginalized groups across different protected categories and social markers, making it challenging to have constructive conversations with industry players, particularly in regions like Latin America. Due to the data scarcity, we lack studies that consider the effects of the system on political, cultural and social features from particular countries and in different languages, for example. There is insufficient data and transparency resources for some regions to the detriment of others, and the ones left aside are precisely the ones where marginalized groups struggle the most to access a basic set of rights and guarantees. To conclude, besides the lack of transparency, we must ask if platforms have a financial incentive to delay the removal of inappropriate content. Do they benefit financially from this kind of acting? These are all factors that should be taken into consideration when evaluating layered moderation systems.

4. Recommendations: from VIP lists to fair protection for speech

As mentioned, we believe that the verification system must exist. This is due to the need for greater protection of some speeches and figures, seeking equity, not mere formal equality. Considering that scale is a major challenge in content moderation, and that technology will invariably be used to deal with this volume, ensuring a level of layered moderation mechanism to contemplate journalists and activists and other actors also means ensuring greater protection of relevant speeches on platforms.

In this case, we should advocate for clearer rules and parameters, as well as a stricter application worldwide. Global companies should have the will and capability to enforce their policies globally. Thinking about how to reform and improve a layered moderation system, we propose the inclusion of settings such as:

1. Clear and public criteria for being or not on the lists of users that will be accepted in layered moderation programs

The operation of a layered moderation system has to be based on transparency precepts, and the first key information to be available to the public is the criteria employed to add or remove users from the "protected list". The development of these programs and lists cannot be a matter of an informal selection that reflects only commercial interests of platforms, for instance. Thoughtful criteria must consider protection of speech, user's profiles, market sizes and impact of posts, among others. Layered moderation programs cannot be designed as a permission for some people to have more rights than others.

2. Publicity of profiles' categories and the percentages of each group in the list composition – for example, number of business partners, politicians, journalists, human rights defenders, as well as their regions, gender and race

In addition to transparent criteria, it is crucial that the public be provided with access to aggregated data on the lists themselves, broken down by categories of profiles, safeguarding the identity of the members. This data is necessary for a more comprehensive understanding of why certain types of users enjoy other layers of examination. It also helps to ensure that these systems are not being employed as mere public relations tools or for commercial purposes. Further, the geographic distribution of such programs should be made known to the public, so as to promote greater accountability and prevent any unintended biases that may arise from localized implementation.

3. Transparency regarding the procedure and its rationale, especially if there is processes of vetting participants and a queue for new participants, how the process for entering and leaving works, and if it is possible to apply or withdraw

Is there a formal procedure that allows determined profiles to apply to have extra layers of review? Who decides on the inclusion of users to the list? It is common that minority and rights advocates do not have as many followers as celebrities, for example, but deserve higher standards of speech protection. Would these people have a chance to apply to this degree of safeguards even if their profiles are not as popular or commercially relevant as others to the digital platforms? Responses to those questions provide legitimacy and the user's right to be informed about fairness in content moderation.

4. Deployment of processes and criteria that take into account political, cultural and social particularities of each region when adding users to the lists

The regional factor is fundamental for layered moderation programs, as well as political, cultural and social contexts of users. This is because different backgrounds can demand different application of rules. For example, if a determined country has high rates of violence against human rights defenders, the criteria should take these numbers into account. Layered systems seek to improve the exercise of moderation, and for that, they must start from local realities to define their application rules.

5. Periodical disclosure of data about the systems operation, including the number of decisions that were reversed by the layered moderation, false positives, false negatives, and so on.

The obligation to disclose periodical data reports about the outcomes of layered moderation is necessary to understand its impacts and the need for its existence within the operation of digital platforms, as well as its changes and evolution over time. Having this information available would allow civil society organizations, governments and the academia to evaluate the automated moderation gaps and to design better tools to fix its flaws.

Part of those recommendations are in line with the ones issued by Meta's Oversight Board on the Policy Advisory Opinion published in December, 2022. On the other hand, we came to the conclusion that a broader framework is needed for dealing with platforms with other formats, as well as specific requirements of transparency that were not addressed by the Board.

***

In this policy paper, we sought to unravel layered content moderation mechanisms, addressing the nuances of systems that dictate the circulation of online discourse, as well as the complexity of treating users in different ways. As demonstrated, we believe that layered content moderation systems must exist to balance drawbacks of industrial-scale

moderation systems within the complex logistical exercise of determining what should remain and be removed from Internet platforms.

Although these systems may be perceived by society as problematic, as they may seem like VIP lists that protect the interests of large platforms' commercial partners, it is fundamental to understand that, on the contrary, when well operated, by treating different users differently, they are capable of generating more fairness and protection to the speech.

Based on these principles, we formulated initial policy recommendations, so that the additional review systems can contribute to promote access to information and fairness among platform's users, instead of causing distortions based on commercial criteria, which foment, on the contrary, inequality in the digital environment. Layered moderation systems should provide for clear criteria and transparent metrics, taking into account local contexts and realities, preventing its purposes from being distorted to favor opaque interests that could prevent equal participation and exercise of human rights online.