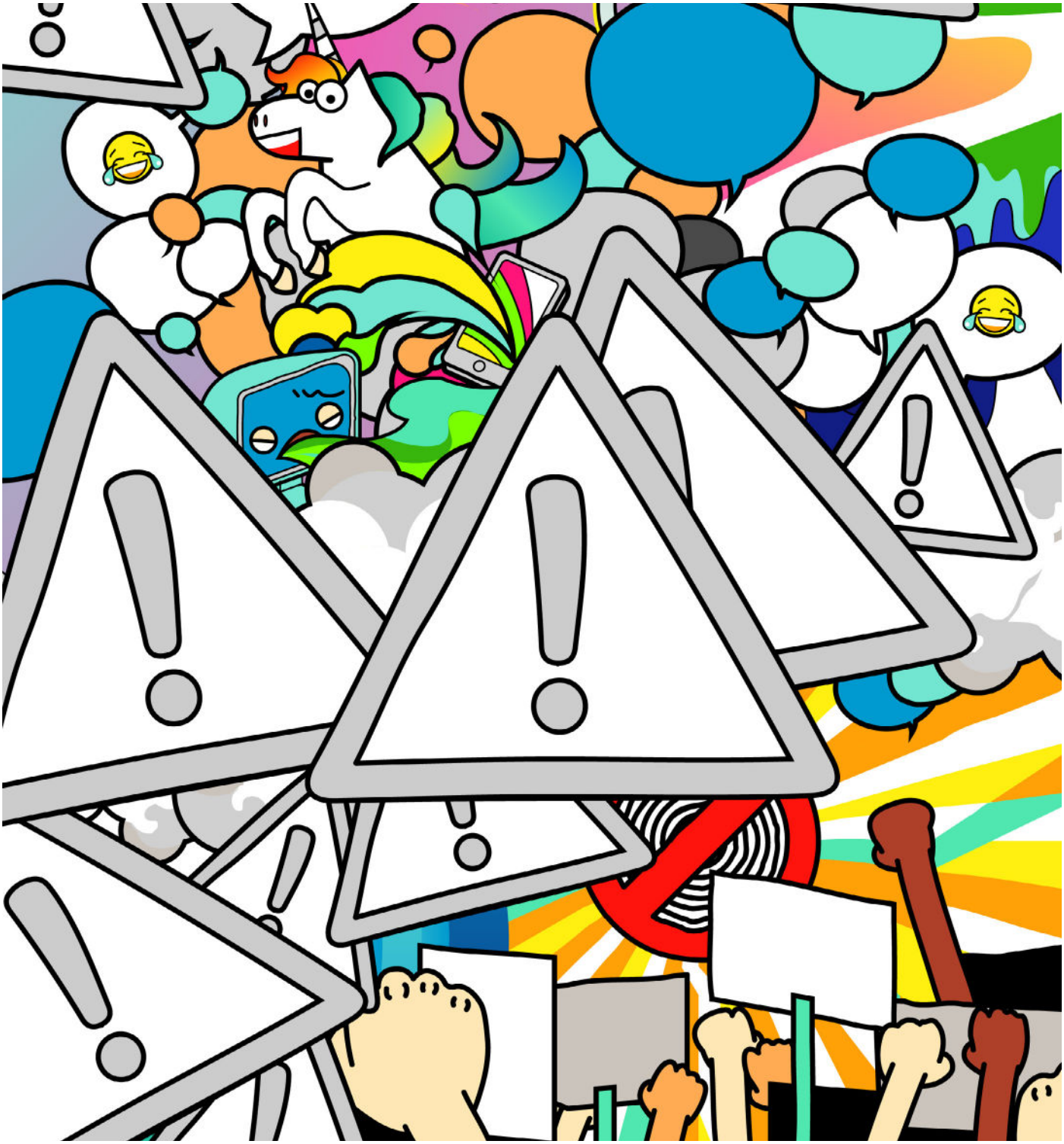
accessnow

accessnow.org



PROTECTING FREE EXPRESSION IN THE ERA OF ONLINE CONTENT MODERATION

ACCESS NOW'S PRELIMINARY RECOMMENDATIONS ON CONTENT MODERATION AND FACEBOOK'S PLANNED OVERSIGHT BOARD

Protecting free expression in the era of online content moderation

Access Now's preliminary recommendations on content moderation and Facebook's planned oversight board

MAY 2019

TABLE OF CONTENTS

I. EXECUTIVE SUMMARY

In January of this year, Facebook announced the launch of a process to create an independent "oversight board for content decisions" to review some of the company's decisions about what user speech to leave up or remove according to the terms of service rules,[1] a practice known as "content moderation."

Access Now welcomes Facebook's recognition of its key role in safeguarding fundamental rights and civil liberties in the digital age and its intention to explore new approaches to address the important issues raised by content moderation. This paper lays out a set of key principles for content moderation that will protect free expression. It provides our preliminary recommendations for Facebook's planned oversight board to govern its content moderation decisions, and our analysis of the implications of the project for human rights.

II. INTRODUCTION: WHAT IS CONTENT MODERATION?

Content moderation is the practice through which an online service that deals with user-generated speech,[2] such as a search engine or a social media platform, makes decisions about whether to host or continue hosting a specific piece of content, or to grant the content relative prominence or prioritization, under the "terms of service" rules.[3]

A decision about whether to host content could entail taking the content down permanently or temporarily, either on the platform as a whole or in relation to certain groups of users in a specific geographical area.

Decisions regarding the prominence of content determine how many and which groups of users are exposed to the content, and these decisions are carried out following different criteria and methods.[4] A decision could mean boosting the reach and exposure of some forms of speech, or demoting or limiting that exposure.

[1] Clegg, N. (2019). Charting a Course for an Oversight Board for Content Decisions. Facebook Newsroom. Retrieved from https://newsroom.fb.com/news/2019/01/oversight-board/
[2] The concept of speech, as used in this paper, includes user expression in all the forms that are technically possible on a platform, that is, via text, images, videos, etc.
[3] These rules might be called "community guidelines" instead of "terms of service," among other names.
[4] Companies can make decisions on exposure/reach on the basis of increasing "relevance," responding to user interests, policy decisions, etc. With regard to methods for content curation, companies could determine what to prioritize independently or implement an automated decision-making system. They could prioritize certain content for a whole category of users or tailor it to the preferences of individuals.

According to the draft charter for the implementation of Facebook's oversight board,[5] the new body would review only decisions about the removal of content, not prioritization. For the purposes of this paper, we narrow our discussion of content moderation to content removal, leaving decisions about the prominence or presentation of content for future analysis.

Content moderation vs. mandatory takedowns: an important distinction

The terms of service that internet companies establish often ban different kinds of content. Typically, this will include both content that is illegal and content that, despite its legality, an internet service considers undesirable. Examples might include nude images, conversations about sex,[6] and certain kinds of discriminatory or hateful speech.

What is considered legal content on the internet varies across countries and regions. When a competent legal authority orders a company to take down content, the company is not exercising a content moderation decision. It is subject to a mandatory takedown.

Conversely, an internet service can decide to take down illegal content without a request from authorities, based only on its terms of service. Taking action to enforce the terms of service, which as we note above can include a prohibition on legal or illegal content, is voluntary.

As we have previously argued,[7] the only time that taking down content should be mandatory for Facebook or other platforms is upon the order of an independent and impartial judicial authority. In addition, to protect human rights, laws that impact speech must respect principles of legality, necessity, and proportionality, must serve a legitimate aim, and must ensure that the people impacted are afforded due process.[8]

Content moderation, in contrast, is an activity that companies undertake to apply their own rules and procedures. This activity is not the same as the legal process we just described.

However, even though companies are not legally required to provide due process when they undertake content moderation, under the international human rights framework, they have a duty to respect human rights in elaborating and applying their terms of service rules. Companies also have the obligation to provide access to remedy to the extent their content moderation causes, contributes, or is linked to human rights harms.[9] Under the UN Guiding Principles on Business and

[5] Facebook (2019). Draft Charter: An Oversight Board for Content Decisions. Retrieved from https://fbnewsroomus.files.wordpress.com/2019/01/draft-charter-oversight-board-for-content-decisions-1.pdf

[6] Bridge, M. (2018, December 10). Facebook bans sex talk and 'solicitation' online. The Times. Retrieved from https://www.thetimes.co.uk/article/facebook-bans-sex-talk-and-solicitation-online-ckvwd6hbr

[7] Stepanovich, A. (2017). Saving our agnostic internet, part I: censorship and free expression. Access Now. Retrieved from https://www.accessnow.org/saving-agnostic-internet-part/

[8] Human Rights Council (2018). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. A/HRC/38/35. Retrieved from https://freedex.org/wp-content/blogs.dir/2015/files/2018/05/G1809672.pdf

[9] Micek, P. (2013). Access delivers Remedy Plan for telcos to redress human rights harms. Access Now. Retrieved from https://www.accessnow.org/access-delivers-remedy-plan-for-telcos-to-redress-human-rights-harms/

Human Rights,[10] companies must not only cooperate with legal inquiries and judicial processes, but also go beyond what is mandated by the courts and participate in non-judicial grievance mechanisms that serve communities whose rights may have been infringed. Unfortunately, many companies, including Facebook, fall short of meeting these obligations.[11]

It is necessary for all companies to apply these essential human rights principles when they make decisions that impact user speech, regardless of whether the decisions are made in-house by content moderators or externally by an oversight board. This is particularly important for dominant platforms like Facebook, which, despite being private entities, have become essential intermediaries in public discourse because of their reach and impact on speech.

III. REGULATIONS FOR CONTENT MODERATION: HUMAN RIGHTS CONSIDERATIONS

Facebook's plans to take action on content moderation do not exist in a vacuum. Under the international human rights framework, governments have the responsibility to prevent, investigate, punish, and provide redress for human rights abuses through "effective policies, legislation, regulations, and adjudication."[12] In the case of content moderation, that implies the responsibility to develop legal protections based on the human right to freedom of opinion and expression, which includes access to information.

A number of actors are voicing concern that self regulation is insufficient to address those challenges. Some civil society groups,[13] for instance, call for incorporating into law human rights-based principles for content moderation, such as transparency, proportionality, and remedy, to protect freedom of expression beyond the mere will of private companies (or their associated independent oversight bodies).

But such a task should be studied and implemented with extreme care. Some of the purported solutions for the free expression issues surrounding content moderation that governments and other internet actors are proposing risk human rights,[14] such as proposals that call to outright ban

[10] Office of the High Commissioner on Human Rights (2011). Guiding Principles on Business and Human Rights. Retrieved from https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf
[11] Ranking Digital Rights (2018). Indicators of human rights compliance. Retrieved from https://rankingdigitalrights.org/index2018/indicators/g6/
[12] Office of the High Commissioner on Human Rights (2011). Guiding Principles on Business and Human Rights. Op.cit.
[13] ADC, Observacom, Intervozes (2019). A Latin American perspective for a democratic regulation that limits the power of big internet platforms and guarantees freedom of expression on the Internet. Retrieved from https://i2.wp.com/www.observacom.org/wp-content/uploads/2019/04/BigPlatformsSmartRegulation.jpg
[14] See Index on Censorship (2019). Online harms proposals pose serious risks to freedom of expression. Retrieved from https://www.indexoncensorship.org/2019/04/online-harms-proposals-pose-serious-risks-to-freedom-of

certain kinds of speech or that entail making companies responsible for screening all content in their platforms to eliminate potential harm.

Legislation that delegates censorship decisions to platforms is not acceptable, especially in relation to speech that demands interpretation in order to determine its legality. It risks consolidating the power of dominant platforms to decide upon the contours of online speech.

Instead, a rights-respecting solution might be for governments to study and elaborate clear laws with potential to enable online companies to maximize the diversity, reach, and quality of public debate. Measures in such laws should include protection for information intermediaries, such as safe harbor rules, from liability for the content published by their users, as well as legal safeguards to ensure that any government requests to delete speech align with human rights principles.[15]

Such laws could also establish limitations to the discretionary power that companies exert when they take down content voluntarily; for example, by establishing requirements for transparency, due process, and compatibility with human rights standards. In any case, those limitations would need to be innovative, created through a participatory process, and take into account the differences among communications intermediaries and the degree of influence that specific platforms and services have on the public sphere, among other factors.[16]

IV. THE HUMAN RIGHTS RISKS LINKED TO CONTENT MODERATION

Decisions regarding content moderation affect the capacity of users to express their ideas and access information online. There is impact at both the individual and collective level, since individual takedown decisions have a cumulative impact, shaping the space for discussion and potentially silencing the voices of entire communities. This is a significant risk for vulnerable or marginalized communities in particular.

Consequently, when private companies set rules for content moderation, to establish minimal legitimacy, they must ensure that the rules are legal, comport with international human rights principles, and are freely accepted by users.

-expression-online/ and Solomon, Brett (2019). Jailing social media bosses won't make us safer from terrorists. Sydney Morning Herald. Retrieved from https://www.smh.com.au/national/jailing-social-media-bosses-won-t-make-us-safer-from-terrorists-20190403-p51afe.html

[15] Several authors (2015). Manila Principles on Intermediary Liability. Retrieved from https://www.eff.org/files/2015/10/31/manila_principles_1.0.pdf

[16] Feld, Harold. (2018). Platform Regulation Part I: Why Platform Regulation Is Both Necessary and Hard. Public Knowledge. Retrieved from https://www.publicknowledge.org/news-blog/blogs/why-platform-regulation-is-both-necessary-and-hard

It is not always clear how private companies make content moderation decisions. Recent research has shown that content takedowns are largely arbitrary and subject to the will of the companies.[17] This is particularly worrying in a context in which governments are exerting public pressure on companies to perform content moderation at increasing speed and precision, vastly outstripping the real-world technical capacity to do so. Inconsistency is to be expected, since the human beings that make content moderation decisions work under tight deadlines with little training or support, [18] and when moderation is machine-assisted, artificial intelligence can fail spectacularly to understand the contextual nuances of language.

Furthermore, users often have little or no chance to respond to content takedowns, assert the legitimacy of the content, or get remedy for improper removal. There are well-known cases in which content that has artistic or historical value has been taken down due to overly strict interpretation of a company's terms of service.[19] For these reasons, the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression has called for greater transparency and accountability in content moderation decisions,[20] as have a number of civil society organizations. Perhaps as a result, some companies, including dominant players such as Facebook, have begun to share more information about their internal procedures and are exploring ways to be more open about their moderation decision-making.[21]

V. CONTENT MODERATION: HUMAN RIGHTS-BASED PRINCIPLES AND RECOMMENDATIONS

Decisions about content moderation and the consequences, which could include permanent removal of content, account suspension, or even banning a user from a platform, can have ramifications not only for free expression but also other fundamental rights, such as the right to freedom of association, as well as for the enjoyment of economic, social, and cultural rights.

[17] Kretschmer, M. & Erickson, K. (2018). How much do we know about notice-and-takedown? New study tracks YouTube removals. Kluwer Copyright Blog. Retrieved from http://copyrightblog.kluweriplaw.com/2018/06/12/much-know-notice-takedown-new-study-tracks-youtube-removals/
[18] Newton, C. (2019). The Trauma Floor. The Verge. Retrieved from https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona
[19] See for example Roger Denson, G. (2017) Courbet's Origin Of The World Still Too Scandalous For Media-Savvy Facebook. Huffington Post. Retrieved from https://www.huffpost.com/entry/courbets-1866-the-origin_b_1087604 and Levin, S. et al. (2016). And Facebook backs down from 'napalm girl' censorship and reinstates photo. The Guardian. Retrieved from https://www.theguardian.com/technology/2016/sep/09/facebook-reinstates-napalm-girl-photo
[20] Human Rights Council (2018). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. Op.cit.
[21] Facebook, n.d. Community Standards. Retrieved in Aprli 2019 from https://www.facebook.com/communitystandards/introduction/

To prevent, address, or mitigate human rights violations on the one hand, or to promote the enjoyment of human rights on the other, Access Now has **developed basic human rights principles**[22] for content moderation, outlined below.

Any entity that makes decisions about the speech of third parties should follow these principles. As companies grow in size, geographical reach, and influence, serving as intermediaries of public discourse, straying from them will represent heightened risk for the human rights of users. For the dominant platforms such as Facebook, which can have a significant impact on public discourse,[23] it is critical to interpret and follow the principles strictly.

Regulating online speech is a particularly complex exercise, and one that needs careful fine tuning so that legitimate policy goals do not have unintended consequences for fundamental rights and freedoms. Any content moderation mechanisms that Facebook designs and deploys, such as the oversight board, should adhere to these principles:

Prevention of harm	Companies must consider human rights from the design of their products through the development and implementation of content moderation policies. This should include the design of better content moderation and curation policies, as well as other human rights-based practices, to achieve an online environment that furthers the free exchange of ideas, empowers users, and protects the rights of vulnerable communities.
Evaluation of impact	Companies should also perform participatory and periodic public evaluations to determine how content moderation decisions are impacting the fundamental rights of users and take the necessary steps to mitigate any harms.
Transparency	All content moderation rules, sanctions, and exceptions should be clear, specific, and properly communicated to users in advance. To be valid, users must be able to accept them freely. The information disclosed should include guidelines to explain the company's internal process for interpreting and applying content moderation rules, to ensure that decisions on content are as predictable and understandable as possible. All information should be available in the official language of the country where the service is provided and be written in simple terms, avoiding excessive technical terminology and references to other documents. Companies should notify users of any changes to these rules and they should be explicitly accepted by users before they can be applicable.

[22] These principles are largely consistent with those recommended by the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression and those contained in the Santa Clara Principles On Transparency and Accountability in Content Moderation (n.d.). Available at https://santaclaraprinciples.org/

[23] Mirani, Leo (2015). Millions of Facebook users have no idea they're using the internet. Quartz. Retrieved from https://qz.com/333313/milliions-of-facebook-users-have-no-idea-theyre-using-the-internet/

Proportionality	The sanctions companies impose on users for violating content moderation rules should be proportional. Severe penalties, such as the banning of a user from an online service, should be a measure of last resort and only take place if there is a serious infringement or after repeated offenses. Details of what constitutes a violation of the rules, what the corresponding sanctions are, and how the policy will be applied should be fully disclosed to users.
Context	Companies should not apply content moderation rules in a "one size fits all" fashion. In addition to using human rights principles as a universal baseline for making content moderation decisions, companies should take social, cultural, and linguistic nuance into account, as much as possible. To achieve that, companies should develop content moderation rules and any accompanying guidelines for their interpretation with the input of local civil society, academics, and users.
Non-discrimination	The application of context-based, nuanced content moderation decisions should be as coherent, systematic, and predictable as possible in order to avoid arbitrariness. Companies should pay special attention to how their content moderation rules are implemented, whether by the company moderators using internal processes or by users via reporting mechanisms, to ensure that they do not unfairly target marginalized communities.
Human decision	Companies should not rely on automated decision-making for content moderation. If the necessities of scale or the sheer volume of user-generated information make reliance on automated decision-making necessary, users need to be informed about use of that technology and have the right to request a human review of their case. Automated decision-making systems need to be as transparent as possible. Companies should publish information about how these systems are used and the procedures behind their application, and should make the systems available for independent auditing.
Notice	Users should get notice when a content moderation decision has been made about their content or speech. This notice should contain adequate information about what sparked the decision, the specific rule that was broken, how content moderation guidelines were interpreted, and the action that will be taken. It should also contain the necessary information to ask for a review of the decision.
Remedy	Companies should provide remediation to users affected by its policies, products, or practices. This includes content moderation decisions, in the cases in which they cause harm to users, such as when the rules have been applied erroneously or excessively. Additionally, company policies should not prevent users from pursuing legal recourse for content moderation decisions, nor force users to renounce such recourse.

VI. FACEBOOK'S OVERSIGHT BOARD FOR CONTENT DECISIONS: HUMAN RIGHTS RISKS AND CONSIDERATIONS

In January 2019 Facebook announced a plan to create an independent oversight board to consult in content moderation decisions.[24] The plan contemplates the creation of a board comprised of a group of independent experts tasked with overseeing content removal decisions on Facebook's social media platform. It would have the power to overturn decisions the company makes, thus acting in an independent capacity.

Facebook is engaging in a global process of consultation with civil society and academic actors focused on the creation and implementation of the board. Among the issues put forward and yet to determined are the composition of the board, the process for its selection, its remuneration, and how cases will be decided.

Facebook has decided that the external oversight board will act as a final decision maker in the most difficult and contested cases of content moderation within the platform.[25] It appears that Facebook may intend for the oversight board to function something like a public court, but without the independence, accountability, or oversight of a legitimate governmental body. Will this private body improve Facebook's content moderation? How might the creation of the board influence efforts across the technology sector to develop long-term solutions for the issues that content moderation raises?

Overall, it's not likely that the board will serve as a "silver bullet" to solve Facebook's content moderation challenges. It has the potential to help, but may not provide sufficient clarity or other human rights safeguards for content moderation decisions. For example, the board could not satisfy requirements for transparency, proportionality, grievance, or remedy on its own. Complying with human rights principles related to the freedom of expression and preventing avoidable damage will require evaluating the company's business incentives. This includes looking into revenue models, recommendation algorithms, advertising transparency, and other issues.

1. Potential human rights benefits of an oversight board

From the information that is available to date, the creation of an external oversight board shows promise for addressing the issues surrounding content moderation. It is a step in the right direction when companies provide more transparency and develop inclusive processes for the exercise of power over expression.

With regard to handling of user-generated speech, it is imperative to bring as much transparency as possible to the process of identifying, investigating, and making decisions regarding purported violations of terms of service rules. This group of external experts, which should comprise

[24] Clegg, N. (2019). Op.cit.
[25] Clegg, N. (2019). Op.cit.

members of civil society, academia, journalism, and other public interest groups, could shed light on how decisions are made within the company by discussing and clarifying interpretation criteria publicly. By acting consistently, the board could also help make future decisions more predictable, bringing badly needed clarity to how speech is policed within the company. Eventually, decisions by the board could contribute to a body of knowledge and shared experience that could be useful for other platforms and services, as well as public authorities, to critique and consider in crafting norms, rules, policy, and regulations for online speech.

In addition, a diverse, international board may help reconcile the need for clear global criteria for content moderation with cultural and technical nuances that are often overlooked. As it stands, despite the fact that Facebook runs a global service, its community guidelines are designed to be applied in a monolithic fashion across regions and cultures, with no room for users to negotiate and any changes decided unilaterally. This approach is highly problematic, so it is welcome that the board's review of content moderation decisions could potentially take into account local points of view.

With respect to the interpretation of global community guidelines, the board should follow existing and well-proven human rights standards as a baseline for content moderation decisions in a variety of cultural and legal contexts.[26] We show how those standards could be put into practice in the recommendations we outlined in Section V.

Consideration of national legislation, provided that it complies with human rights standards, can also help illuminate the cultural aspects that need to be taken into account.

Finally, it has been announced that most of the difficult questions about the workings of the board will be subject to public consultation.[27] The process of selection of the board members, the design and logistics of the board itself, including questions about independence and cultural diversity, and the selection of cases it will decide, are among the issues that Facebook has promised to submit to the consideration of the internet community in the months to follow. We expect Facebook to honor this commitment to make sure the process is truly transparent and inclusive.

2. Potential human rights risks of an oversight board

Building a content moderation mechanism is a highly complex exercise that could have unintended consequences for human rights. Making a dedicated oversight board an integral part of that mechanism, whether it is an internal or external board, could have a severely detrimental impact if it's not done correctly. Here are some key considerations to avoid human rights risks.

Facebook must be realistic about the significance of a self-regulating oversight board. Such a structure is not a democratic, public institution with the legitimacy to determine the right to receive and impart information, nor does it have the same accountability. Even with the participation of independent experts, academics, and civil society organizations, **this initiative cannot and should not aspire to replace democratic public institutions such as the judiciary.**

[26] Human Rights Council (2018). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. Op.cit.
[27] Clegg, N. (2019). Op. cit.

The functioning of the oversight board **should in no way undermine or negatively affect the implementation of grievance mechanisms for communities whose rights may have been infringed.** Human rights principles require the existence of those grievance mechanisms,[28] and companies like Facebook should use this opportunity to develop an integrated and comprehensive vision for rights-respecting policies that does not limit in any way access to subsequent judicial remedy.

The board could make mistakes. It is imperative that **board decisions are properly evaluated and substantiated.** As we note above, if a decision is at odds with fundamental rights, users should also have the opportunity to easily and appropriately seek subsequent legal redress.

The definition and implementation of the oversight board needs to be truly inclusive, including the public consultations for creating it. This initiative will succeed only if there is sufficient and significant community engagement and a clear commitment to listen and act upon the input and recommendations of the community.

Facebook should not expect stakeholders to dedicate time and resources to improving Facebook's business model and practices for free. The company should continue **committing the necessary resources** to consult stakeholders who cannot afford to dedicate time to this project or attend in-person meetings pro bono.

VII. FACEBOOK'S OVERSIGHT BOARD FOR CONTENT DECISIONS: PRELIMINARY RECOMMENDATIONS

The functioning and decision-making of Facebook's oversight board should follow the principles for rights-respecting content moderation outlined in Section V.

As the process of implementation moves forward, there is a unique opportunity to build these principles into the design and functioning of the board, including during the preparatory consultation stages. Following are specific recommendations for applying these principles.

[28] See Office of the High Commissioner on Human Rights (2011) United Nations Guiding Principles on Business and Human Rights (2011). Available at https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf
It is noteworthy that the guiding principles also establish criteria of effectiveness of grievance mechanisms, which constitutes a useful guide for their implementation (see Principle 13).

To enhance **transparency**, the operations of the board, as well as its decisions and the reasoning behind them, should be made public. This would allow users, civil society, academics, the judiciary, and other interested parties to scrutinize its functioning.

The process for setting up the board must be fully transparent. Participants taking part in consultations to create the board, or engaging in the process once it is set up and running, must be free to discuss and publish ideas and plans openly, without limitations other than those linked to protecting personal user information.

In order to have meaningful access to **remedy**, users should be able to appeal a content moderation decision before a judge if their fundamental rights are harmed. Facebook should abstain from limiting or conditioning the exercise of this right.

To be successful, the board must by design include groups and individuals that represent or are closely related to the users most at risk of harm due to content moderation decisions, integrating their perspectives in daily operations. Facebook must proactively seek to include representatives of minorities and disenfranchised communities from different regions of the world. This would allow the board to consider the context of user conduct and to avoid discrimination.

Facebook must be realistic and honest about the scope, reach, and impact of this initiative and communicate accordingly. It should not exaggerate or over-sell its legitimacy or usefulness. If things go wrong, Facebook should not shield itself behind the board, and should always take responsibility for whatever happens on its platform or any other of its products and services.

In addition to these specific recommendations for applying human rights principles, there are other considerations that Facebook should take into account in the consultation and creation of the oversight board to better **evaluate** and **prevent** adverse human rights impacts.

Authentic engagement and responsiveness. Facebook must genuinely listen to academic and civil society experts and incorporate their views as much as possible. They will provide Facebook with invaluable feedback about user interests at stake in content moderation decisions.

Open and inclusive planning. All activities for creation of the board should be planned in advance and publicized in due time to enable stakeholders – especially civil society – to provide input. This includes notifying the public of deadlines and providing effective tools to submit feedback.

Adequate resources for inclusion. Facebook should provide the resources necessary for the participation and work of those stakeholders who cannot afford to dedicate their time to the consultations related to the board at their own cost.

As the process to create the board progresses and new questions arise through consultations, we may offer additional recommendations, which would apply to any company that seeks to create a similar board to assist in content moderation decisions.

VIII. CONCLUSION: THE CONTENT MODERATION CRISIS IS AN OPPORTUNITY TO EMBED HUMAN RIGHTS

As the internet grows and some parts of it consolidate,[29] internet services often seek to establish common, shared guidelines to govern expression on their platforms. Meanwhile, in a diverse and global information ecosystem, actors of all kinds, including investors, governments, civil society organizations, and the users themselves, are pressing platforms to eliminate certain kinds of speech — including content that may be perfectly legal, though perhaps unsavory — increasingly rapidly.

The demands of content moderation at scale, the current incentives for sharing incendiary material, and governments' push for control of expression, are putting human rights at risk. In order to attain a healthy environment for meaningful and empowering online expression, our priority ought to be finding ways in which users can express themselves freely and limitations to expression are decided and implemented through democratic means.

[29] Internet Society (2019). Consolidation in the Internet Economy. Retrieved from https://future.internetsociety.org/2019/wp-content/uploads/sites/2/2019/04/InternetSociety-GlobalInternetReport-ConsolidationintheInternetEconomy.pdf

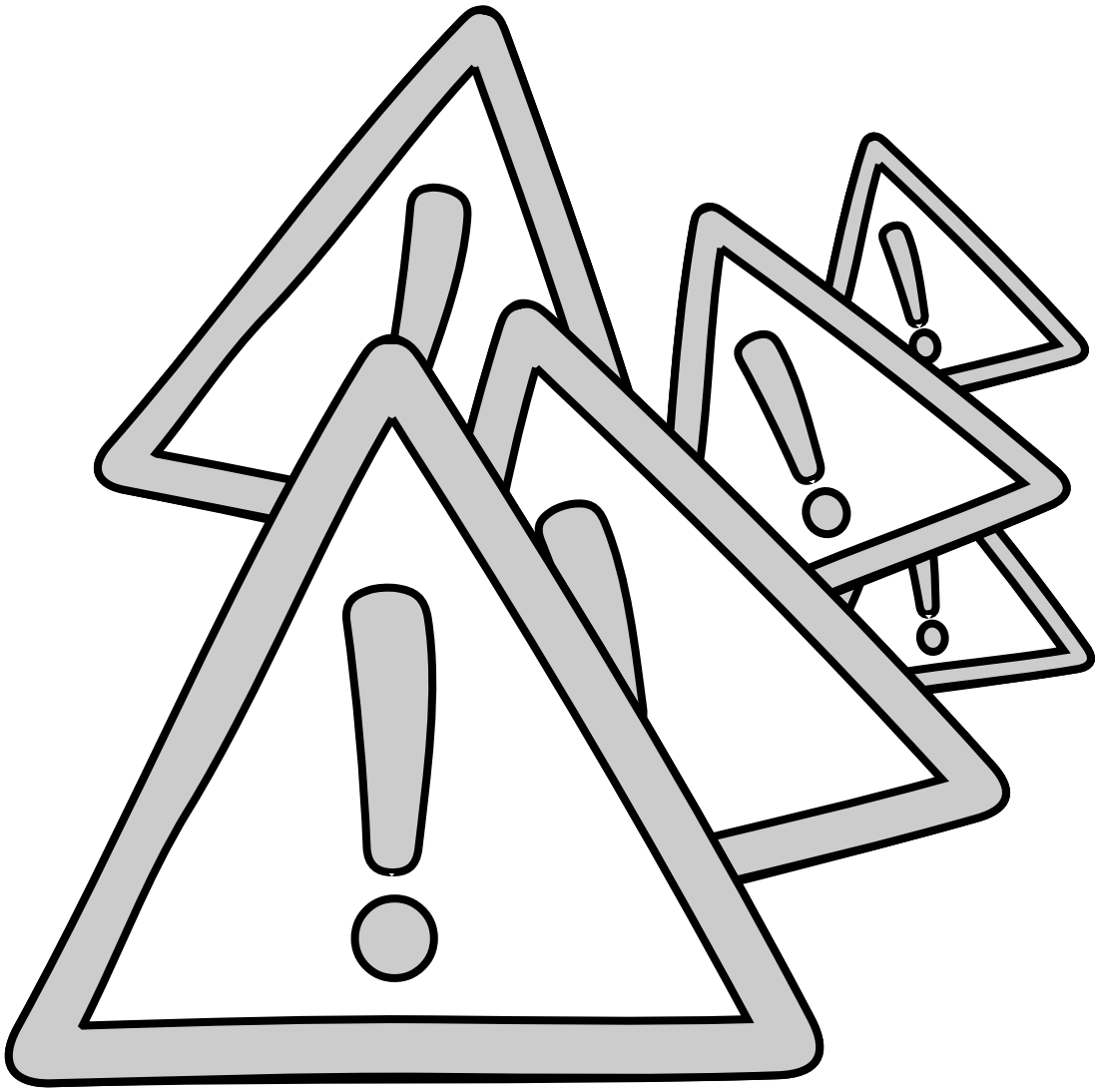
We welcome the voluntary measures undertaken by companies that aim for greater transparency and participation in their content moderation decisions. Facebook's oversight board is potentially a good example of a self-imposed limitation on a company's discretionary power over user speech. It will of course take time to evaluate the impact of the oversight board. Regardless of the outcome, however, achieving a healthy information ecosystem is a challenge that goes beyond the case-by-case decisions that a company makes on specific kinds of content.

This is the wider cultural, economic, and societal challenge that has to do with how we create, publish, engage with, and act upon information in the digital era. There is no single actor responsible for harmful information online, and no single solution will address all the issues of concern. We believe that long-term, holistic solutions can emerge from evidence-based, participatory, and committed public policy that is based in human rights principles such as those we have outlined in this paper. That policy needs to be properly evaluated and discussed by all stakeholders. If legislative solutions for content regulation are rushed, it could prove even more dangerous for human rights than the current approaches.

This discussion paper is a publication of Access Now and was written by Javier Pallero with the collaboration of Access Now's policy team. The author would like to specially thank Amie Stepanovich, Fanny Hidvégi, Guillermo Beltrà, Peter Micek, Raman Chima, Melody Patry, Donna Wentworth, and Sage Cheng for their contributions.

For more information about this report, please contact **Javier Pallero** (javier@accessnow.org)

Access Now (https://www.accessnow.org) defends and extends the digital rights of users at risk around the world. By combining direct technical support, comprehensive policy engagement, global advocacy, grassroots grantmaking, and convenings such as RightsCon, we fight for human rights in the digital age.



PROTECTING FREE EXPRESSION IN THE ERA OF ONLINE CONTENT MODERATION

ACCESS NOW'S PRELIMINARY RECOMMENDATIONS ON CONTENT MODERATION AND FACEBOOK'S PLANNED OVERSIGHT BOARD

accessnow

Access Now (https://www.accessnow.org) defends and extends the digital rights of users at risk around the world. By combining direct technical support, comprehensive policy engagement, global advocacy, grassroots grantmaking, and convenings such as RightsCon, we fight for human rights in the digital age.