ARTICLE 19

unesco
In partnership with UNESCO

Funded by the
European Union

Content Moderation and Local Stakeholders in Colombia

April 2024

SOCIAL
MEDIA
4PEACE

First published by ARTICLE 19, April 2024

ARTICLE 19
72-82 Rosebery Ave
London EC1R 4RW
UK

T:	+44 20 7324 2500
F:	+44 20 7490 0566
E:	info@article19.org
W:	www.article19.org
Tw:	@article19org
Fb:	facebook.com/article19org

This publication was produced with the financial support of the **European Union** and **UNESCO**. The designations employed and the presentation of material throughout this publication do not imply the expression of any opinion whatsoever on the part of UNESCO or the European Union concerning the legal status of any country, territory, city or area or its authorities, or concerning the delimitation of its frontiers or boundaries.

The authors are responsible for the choice and the presentation of the facts contained in this publication and for the opinions expressed therein, which are not necessarily those of UNESCO or the European Union and do not commit the Organisations.

ARTICLE 19 would appreciate receiving a copy of any materials in which information from this report is used. ARTICLE 19 bears the sole responsibility for the content of the document.

Acknowledgements

Many thanks to Carolina Botero Cabrera, Catalina Moreno Arocha, and Darly Diaz, who have conducted this research and authored this report in close consultation with ARTICLE 19.

Carolina Botero Cabrera is the Director of Fundación Karisma and a columnist for El Espectador and La Silla Vacía. She is a lawyer with a Master's degree in International Law and Cooperation, and also in Trade and Contract Law. For more than a decade she has worked in the promotion and defence of human rights on the internet. She is a member of the Board of Directors of Creative Commons and the Civil Society Information Society Advisory Council (the civil society observer group at the Organisation for Economic Co-operation and Development).

Catalina Moreno Arocha is the Coordinator of the Social Inclusion Line of Fundación Karisma and is a lawyer with a Master's degree in Public Law. She is dedicated to human rights and social justice issues. She worked for more than a decade in the Constitutional Court of Colombia and has also worked as a lawyer in political and legal advocacy on gender issues. At Karisma, she is dedicated to encouraging technologies to serve and protect social groups that are exposed to violence and discrimination.

Darly Diaz is a Colombian sociologist and journalist with a Master's degree in Development Studies focused on human rights, social justice perspectives, and gender. She has more than five years of experience working with various communities, Afro-Colombians, journalists, and children under protection measures, for the defence and promotion of human rights in physical and digital environments.

Many thanks to all who have shared their expertise and views and contributed to the research process. Thanks also to our partners at the United Nations Educational, Scientific and Cultural Organization (UNESCO) who have provided input and feedback at various stages of the report.

Contents

Executive summary

This report examines the content moderation practices of social media companies in Colombia and their implications for freedom of expression. It identifies significant challenges within the information ecosystem facilitated by social media platforms. The content moderation practices of major platforms in Colombia were viewed critically by many of the local stakeholders who participated in this research, highlighting several key issues.

Lack of transparency

Content moderation practices lack transparency and although platforms provide information about their processes they are not sufficiently clear in Colombia – just as in other parts of the world. Appeals often go unanswered and stakeholders believe that there is no predictability or proportionality in content moderation decisions. While platforms often notify users when content is moderated, there is little information on appeal mechanisms, and these mechanisms are generally seen as ineffective. Some platforms offer certain information on content moderation in Colombia – primarily through their transparency reports – yet this data is not always useful for identifying trends or comprehending the scale and impact of content moderation in Colombia.

This lack of transparency impedes civil society's and users' understanding, hampers advocacy efforts, and ultimately weakens accountability for the platforms' actions in Colombia. Transparency regarding content moderation practices is key for civil society to understand how content moderation functions and what its impact is in a particular context. It is therefore important to engage with any regulatory framework that directly or indirectly affects content moderation, including transparency obligations, or is influenced by it. A growing concern is the opacity and perceived censorship effects of curation practices. There is an urgent need to better understand content moderation and curation practices such as downranking or shadow banning, which are less visible and less understood than content takedowns or account suspensions.

Lack of contextual understanding

Although social media is a key player in the information ecosystem, there is still a lack of contextual knowledge about the functioning of content moderation and curation in Colombia. Platforms often approach content moderation as a global issue, despite the importance of understanding the local context. While platforms acknowledge this to some extent and have implemented strategies to address contextual nuances in certain situations, interviewees find these efforts insufficient.

One of the consequences of this lack of contextual understanding is the neglect of groups of people who are particularly vulnerable in Colombia and are subjected to online attacks, such as human rights defenders and journalists, especially women journalists. This increases the likelihood of these individuals' exposure to offline violence. There is also the difficulty of circumventing content moderation in order to report various public interest issues, including human rights violations. This can involve measures such as prohibiting actors who are actively engaged in peace processes with the government or removing content that exposes police abuse during periods of unrest, as well as censoring slurs used in protest chants during crucial political moments for the country.

Content moderation practices

The challenges posed by content moderation practices also impact the relationship between media actors and social media platforms. While social media platforms are a primary distribution channel, relying on them can have a negative impact on media due to these content moderation challenges and may result in self-censorship by media actors when making editorial decisions.

The widespread use of automated content moderation tools, while understandably deployed, has the potential to exacerbate all of the described content moderation issues and has been raised as an issue by various stakeholders interviewed.

State involvement

The Colombian state's involvement in content moderation has also raised concerns. Colombian authorities, lacking a legal basis to do so, have been submitting requests to platforms based on community rules to request content removals. There is a lack of transparency regarding these requests, making it challenging to understand the grounds for the requests, how platforms evaluate them against human rights standards in Colombia, and which authorities are responsible for making these requests.

Analysis and conclusions

Against this background, the report analyses the feasibility of establishing a local coalition on content moderation and freedom of expression in Colombia to support the establishment of channels of communication and cooperation with social media companies and regulators, and to address those issues that threaten freedom of expression online, media freedom, and societal cohesion.

The report concludes that, within the Colombian context, leveraging pre-existing networks, coalitions, or organisations focused on digital rights and freedom of expression would be more successful than establishing a completely new initiative. The report concludes with recommendations supporting this proposal.

Introduction

This publication has been produced as part of the United Nations Educational, Scientific and Cultural Organization's (UNESCO) project **Social Media 4 Peace** funded by the European Union (EU).

About the project

This report is part of the **Social Media 4 Peace** project that UNESCO is implementing in Colombia, Bosnia and Herzegovina (BiH), Kenya, and Indonesia with support from the EU. The overall objective of the project is to strengthen the resilience of societies to potentially 'harmful' content spread online, in particular 'hate speech' and 'disinformation',[1] while protecting freedom of expression and contributing to the promotion of peace through digital technologies, notably social media. ARTICLE 19's contribution to the project focuses on concerns raised by the current practices of content moderation on the largest social media platforms in the four target countries.

In addition to the four country reports elaborated with external research consultants, ARTICLE 19 also published a summary report for Bosnia and Herzegovina, Indonesia, and Kenya which compared the learnings and recommendations.

ARTICLE 19 considers that social media companies are, in principle, free to restrict content on the basis of freedom of contract, but that they should nonetheless respect human rights, including the rights to freedom of expression, privacy, and due process. While social media platforms have provided opportunities for expression, a number of serious concerns have come to light. The application of community standards has led to the silencing of minority voices. The efforts of tech companies to deal with problematic

[1] The terms 'hate speech' and 'disinformation' are not defined in international human rights law. Similarly, while many policymakers and also a number of legislations refer to 'harmful' content, there is no international consensus on its meaning, and it lacks a legal definition in international human rights law. For these reasons, ARTICLE 19 uses these terms in inverted commas throughout this publication.

content are far from being evenly distributed: for instance, it was reported in 2021 that 87% of Facebook's spending on misinformation went to English-language content, despite the fact that only 9% of its users were English speaking. The leaked tier list of the company also revealed that most of the content moderation resources were being allocated to a limited number of countries. At the same time, the transparency and dispute resolutions over content removals have so far been inadequate to enable sufficient scrutiny of social media platforms' actions and provide meaningful redress for their users. Finally, there are concerns that a small number of dominant platforms hold so much power over what people are allowed to see without more direct public accountability.

This report specifically looks at the situation of local actors in Colombia. The research conducted under the project for this report reveals that although these actors are impacted by the circulation of potentially 'harmful' content on social media or the moderation thereof, they often find themselves unable to take effective action to improve their situation in that respect. In some instances, they feel frustrated by the inconsistencies of platforms' application of their own content rules, and in others, they feel that platforms ignore their requests or misunderstand the specific circumstances and contexts of the country or region. Some actors lack understanding of content rules or content moderation.

The research examines the views of local stakeholders and the role that a local coalition on content moderation and freedom of expression could play in improving conditions and enforcing rights in the digital world. It seeks to provide guidance on the best ways and strategies to build connections to fill the gap between the realities of local actors, the public sector, and private companies that operate on a global scale in content moderation.

The idea of national coalitions relies on the premise that it is essential for social media platforms to acquire an understanding of the local context in which they operate and engage with local stakeholders. Gathering local knowledge and understanding of the local context (linguistic, historical, political, and societal) would allow social media platforms to improve their content moderation practices and make them contextually relevant. A local coalition on freedom of expression or alternative structure could therefore engage in a sustainable dialogue with social media platforms and contribute to addressing flaws in content moderation and improving the protection of fundamental rights online. It could further engage in capacity building through providing training and support on content moderation and freedom of expression to other local civil society actors that are impacted by content moderation.

Through this research, the idea of a local coalition on content moderation and freedom of expression in Colombia was submitted to local stakeholders. Their views enabled recommendations to be collated on how the coalition proposal could deal with content moderation issues in Colombia.

To that end, and while focusing on the local voices in Colombia, this report examines local-specific content moderation issues, including case studies, and the position, knowledge, and needs of various state and non-state actors. It highlights the diversity and complexity of Colombian society and history as a background to understanding the report. It also presents how deep conflicts within society have, at times, been exploited for political and economic profit.

The report begins by describing the social media landscape and exploring the dynamics and issues related to the use of social media and the practices of content moderation in the country. It then discusses how to form a coalition on content moderation and freedom of expression, and examines the needs, gaps, and strengths of a prospective coalition. Next, it analyses relevant stakeholder groups that deal with or are impacted by content moderation practices. The report concludes with recommendations on the feasibility of the formation of a civil society coalition on content moderation and freedom of expression in Colombia to bridge the dialogue between social media and local civil society.

For the purposes of this report, we rely on the following definitions:

- **Content moderation** includes the different sets of measures and tools that social media platforms use to deal with illegal content and enforce their community standards against user-generated content on their service. This generally involves flagging by users, trusted flaggers or 'filters', removal, labelling, downranking or demonetisation of content, or disabling certain features.

- **Content curation** is how social media platforms use automated systems to rank, promote, or demote content in newsfeeds, usually based on their users' profiles. Content can also be promoted on platforms in exchange for payment. Platforms can also curate content by using interstitials to warn users against sensitive content or applying certain labels to highlight, for instance, whether the content comes from a trusted source.

There may be some overlap between content moderation and content curation processes. For example, downranking a piece of content can be a content moderation measure but is also an inevitable part of the content curation process.

Methodology

A combination of methodologies was used in this report. First, the research relied on a thorough review of academic and non-academic sources to provide a comprehensive overview of the issues. Then, qualitative data collection enabled the gathering of perspectives from various societal sectors. Interviews were organised to understand the local experiences and challenges in dealing with platforms on content moderation issues.

Four different questionnaires were developed to guide interviews with social media platforms, civil society organisations, academia, and media outlets. Some of the questions were applied in a general manner, while others were personalised depending on the proximity to and role of the respondent in the content moderation processes. Some questions related to the respondent's opinions regarding potentially 'harmful' content and how they could imagine a coalition working together on these topics.

In total, 23 interviews were carried out and 5 written contributions received (see Annex B). The researchers reached out to a larger pool of stakeholders but did not receive a response from all those contacted. Most of the interviews were conducted on Zoom, and some contributions were collected through an online survey shared with actors who did not have the time to participate in an interview. The survey was also posted on Fundación Karisma's social media channels to include additional voices of interested stakeholders.

Each interview and survey explained the purpose of the project, the relevance of the participants, and how their responses were going to be used. Karisma distributed written consent forms that were signed by the participants. In some cases, researchers sought and received verbal consent at the start of the interview to use the information in this report.

Although some state institutions involved in the content moderation landscape were contacted, they were reluctant to be interviewed. The Ministry of Information and Communication Technologies of Colombia (MINTIC) only provided written comments. The researchers tried to engage the Prosecutor's Office and the Ombudsman's Office in the information gathering. At the request of UNESCO, the Ministry of Foreign Affairs contacted the Ministry of Defense and the Superintendence of Industry and Commerce.

The civil society organisations interviewed had a wide range of remits: protection and defence of children's rights, women organisations, trans organisations and individuals, Afro-Colombian rights collective, indigenous national network, freedom of expression, freedom of the press, tackling online disinformation, peacebuilding, and awareness of the use of psychoactive substances. Journalists, academics, and think tank centres were also interviewed.

Interviews with Meta, X (formerly known as Twitter), and Google involved the participation of various representatives from those platforms, therefore it was not a one-to-one meeting but rather a conversation with staff from different departments.

Colombia at a glance

To understand the dynamic of content moderation and freedom of expression in Colombia, it is important to first comprehend the country's complex political and governing system. The current governing structure of the country was established in the Political Constitution of 1991. Colombia's Truth Commission (Comisión de la Verdad)[2] stated that the proclamation of a new Constitution was a turning point in the history of the country.[3]

Before the new Constitution, under a bipartisan scheme, political movements were excluded between 1958 and 1977. Left-wing guerrillas emerged and armed resistance resurfaced as a result of social discontent. This discontent deepened between 1978 and 1991, the uprising was consolidated, and there was a repressive response from the Colombian state. The war on drugs also began. It was a time of a permanent state of siege, with an increase in human rights violations. This period ended with the establishment of the National Constitutional Assembly in 1991, an initiative promoted by social movements, especially students, within the peace dialogues with the M-19 guerrillas.

The new Constitution ended the former political system: it proposed a more pluralistic, inclusive, and democratic model of state. The Truth Commission, however, stated that the impact of the Constitution had been unequal for Colombia's different regions and population groups. The Commission also stated that there was a violent reaction to the democratic opening brought by the new Constitution because two of the country's most important armed groups (National Liberation Army [Ejército Nacional de Liberación] and

[2] Colombia's Truth Commission was established by the 2016 FARC peace accord to address the country's ongoing six-decade conflict that has affected more than nine million registered victims. The Commission released its final report on 28 June 2022.

[3] A report from Colombia's Truth Commission was used for the overview of the history of the armed conflict in the country.

Gaitanist Self-Defense Forces of Colombia [Autodefensas Gaitanistas de Colombia]) were not included in the new pact, and negotiations with the heads of the drug traffickers failed before they were even completed. At the same time, a peace movement gained force and managed to de-escalate the armed conflict, although, as of 2023, the conflict is still not over.

According to the 1991 Constitution, Colombia is a unitary republic, with an administratively decentralised organisation, distributing the administration between the national government and local governments. There are three branches of government: legislative, executive, and judicial (with a 'checks and balances' mechanism), and certain autonomous bodies with specific functions. Another judicial mechanism, the *tutela* action before the specialised Constitutional Court, is particularly relevant to fundamental rights.

Any citizen can file a *tutela* with minimum requirements on evidence and without being a lawyer or having legal knowledge. It must be resolved within ten days by a judge, and it serves for the protection of any fundamental right, including the right to freedom of expression, of press, and of political participation. Its use is increasingly popular: in 1992, when the mechanism was introduced, 10,000 *tutelas* were filed; in 2022, 633,463 were filed. *Tutelas* transformed the way law is understood, and they facilitated the sense of ownership of rights by citizens. Most of the case law of the Constitutional Court in relation to freedom of expression has come through judgments in *tutela* filings.

The Constitution protects freedom of expression and freedom of the press. Article 20 guarantees the right to express and disseminate thoughts and opinions, as well as to impart and receive truthful and impartial information. Article 20 also guarantees the right to establish media outlets and ensures the right to correct published information under fair conditions. It further states that 'there shall be no censorship'.

Despite this new constitutional framework, two concepts that had served to stigmatise social mobilisations and members of the political opposition – the 'national security' doctrine and the 'internal enemy' – were strengthened in the army's counter-guerrilla combat manuals and regulations. These concepts had a legitimising impact on violent actions by state actors against members of opposition parties and student, rural, and social leaders.[4]

The stigmatisation of human rights activists and political opponents continued to take place under the new Constitution. At the beginning of the 2000s, verbal attacks by high-ranking politicians against human rights defenders and journalists became more vigorous. Specifically, these intensified during the presidency of Álvaro Uribe Velez. For example, in 2003 and 2004, Colombian non-governmental organisations (NGOs) filed *tutela* actions against statements made by then President Álvaro Uribe Velez following his public demonisation of human rights defenders.[5]

This has also been highlighted by the Constitutional Court of Colombia, which has declared that those in positions of power hold a degree of responsibility towards the public and should ensure that their public speeches fall within the remit of the right to

[4] This has also been recognised by the Inter-American Court of Human Rights (IACtHR) in the cases *Manuel Isaza Uribe v Colombia* (2018) and *Unión Patriótica (UP) v Colombia* (2023). The Court declared the international responsibility of Colombia for systematic violations of human rights against the members and militants of the Unión Patriótica (UP) political party for more than 20 years. Regarding the context of the human rights violations, the Court found that state agents generated an atmosphere of stigmatisation against UP members in order to exclude them from the democratic game, presenting them as 'the armed wing of the FARC'. These kinds of statements had an influence on the public perception, which, in turn, influenced the violence against party militants.

[5] The facts disputed by the NGOs in the *tutela* action included: (1) a speech made during the promotion ceremony of a general of the Colombian army and broadcast on national television by presidential order in which the President targeted the organisations and described them as 'cowardly waving the flag of human rights, to try to return Colombia to terrorism', (2) a presidential speech after an attack in the department of Boyacá, in which the then President described human rights defenders as unsubstantiated talkers, and (3) when the then President, during his intervention before the Foreign Affairs Committee of the European Parliament, described the lawyer of a Colombian NGO present in the room as: 'he belongs to an NGO, El Colectivo de Abogados, which hides behind its quality as a human rights organization to defend the guerrillas'.

freedom of expression.[6] The Court recalled the criteria of truthfulness and impartiality for information, factual justification and reasonability of opinions, and respect for the fundamental rights of citizens as essential aspects in assessing speeches by public officials.[7]

Civil society organisations continued to litigate against public statements made by high-level public officials when they promoted negative ideas about vulnerable groups like migrants[8] or women journalists[9] or when they targeted the exercise of fundamental rights,

[6] Judgments T-1191 of 2004 and T-1062 of 2005, and Ruling T-627 of 2012.

[7] The claim was filed by a group of more than 1,000 women against the Attorney General and two Deputy Attorneys in matters of Childhood and Family and for the Public Function of Colombia. For three years, these officials made different public and institutional statements, with inaccurate and distorted information regarding women's reproductive rights and contraceptive methods. Among the facts that supported the claim, the Attorney General misrepresented an order of the Constitutional Court on sexual and reproductive rights and published a press announcement stating that 'the Court had ordered the implementation of massive campaigns to promote abortion'.

[8] In Ruling T-087 of 2021, the Court examined a *tutela* action filed by a Venezuelan citizen against Claudia López, mayor of Bogotá who, in 2020, addressed the security situation in the city and made reference to the participation of Venezuelan citizens in criminal actions, expressly pointing out their nationality. The Court indicated that the mayor exceeded her power-duty of communication with the citizens and freedom of expression because the statement was discriminatory. It also clarified that for the security situation – a matter of public interest – public officials must foresee the risks associated with their pronouncements, since they can create or aggravate the stigmatisation against certain groups of the population.

[9] In decision T-087 of 2023, the Court studied a *tutela* filed by a group of female journalists on the occurrence of online attacks of a misogynistic and sexual nature. These attacks aimed to invalidate their journalistic work, and several political parties and movements took advantage of this. The journalists stated that the National Electoral Council (CNE in its Spanish acronym) was called upon to adopt measures to cease this type of violence but failed to do so, which encouraged the occurrence of the aggressions. The Court concluded that since the journalists did not inform the CNE of the aggressions it could not have deployed the necessary measures. Nevertheless, the Court recognised the latent phenomenon of gender-based violence occurring on the internet and its multidimensional impacts. The decision highlighted that

such as sexual and reproductive rights. Stigmatisation of human rights defenders is still visible in Colombia in the digital world and jeopardises the work and safety of civil society organisations and activists.

It is noteworthy that, between 2019 and 2021, Colombia experienced the most important periods of protests in recent history. The protests were originally triggered by the passing of a tax reform and a higher education reform by the government before Congress. However, the protests increased in 2021 because of discontent related to the re-emergence of the Covid-19 pandemic. The government took a public order approach to these protests, and social demands were not processed through institutional channels.

After a working visit to investigate human rights violations during the 2021 social protests, the Inter-American Commission on Human Rights (IACHR) observed:

> *the existence of a climate of polarisation directly related to both racial, ethnic, and gendered structural discrimination, as well as political actors. This phenomenon is present in different social sectors and is manifested in stigmatising discourse that in turn leads to an accelerated deterioration of public debate. The Inter-American Commission finds this discourse especially worrisome when it comes from public authorities. (para 5)*

For the IACHR, the internet has enabled protesters in Colombia to report incidents and make open complaints about the use of excessive force by police, as well as to request protection of their rights, facilitating and enriching public deliberation, and denouncing human rights violations during demonstrations. This has highlighted the need to guarantee free access to the internet. The IACHR received complaints on alleged state measures that could curtail freedoms online, such as cyber-patrolling and profiling practices, the classification of internet content as true or false by law enforcement agencies, internet shutdowns, and IP address blocking. The IACHR stated that 'according to the information

public entities and political parties must undertake activities for the prevention of and timely response to such situations and indicated that a regulation that recognises and establishes specific mechanisms to respond to digital violence was needed.

provided by different actors, these measures were adopted based on subjective criteria instead of objective, legitimate, and transparent parameters in line with international human rights standards' (para 174).

The IACHR also noted that most of the stakeholders interviewed during their visit stated that although the internet is an important platform for public deliberation, 'they expressed fears that some discourse may encourage violence or be the basis for decisions about the internet that take away the voice of the public' (paras 173–175).

Between 2002 and 2016, the state responded to insurgent groups through military action. Such conflict caused serious human rights violations as the combatant groups directly and indirectly involved civilians. Transitional justice started in 2005, with a peace agreement between the government and right-wing paramilitary groups. Following dialogue between the government and the Revolutionary Armed Forces of Colombia – People's Army (Fuerzas Armadas Revolucionarias de Colombia – Ejército del Pueblo; [FARC-EP]), the Peace Agreement was finally signed in 2016.

Despite the Agreement, peace is far from being consolidated. Today, Colombia faces a series of fragmented regional confrontations that, 'although not entirely disconnected from each other, unlike in previous decades, do not have as their backbone the dispute for political power or control of the state'.[10] While the structural reforms proposed in the Peace Agreement are being implemented, the country continues to face the problem of drug trafficking and its illegal rents that feed the current violence. The assassination of social leaders and human rights defenders, as well as former FARC-EP combatants, has increased. According to the results of a 2021 study conducted by Movilizatorio, an expert laboratory for

[10] As indicated by Colombia's Truth Commission, https://www.comisiondelaverdad.co/no-mataras.

social transformation, the issues surrounding the Peace Agreement and the Special Justice for Peace[11] are among the most polarising topics in Colombia, at least on X.

Polarisation is not new in Colombia; it can be traced back to different moments of the country's violent political history. Movilizatorio's study found that social media contributes to the intensification of such polarisation because it allows for the rapid dissemination of certain content and the self-reaffirmation of a specific discourse, as a result of the platforms' recommendation systems. However, Movilizatorio's analysis also concluded that 'despite the wide perception of polarization, and the polarization generated around thematic agendas, there is in Colombia a unity of moral values on which it is possible to build great agreements for the country'.

On the other hand, the Freedom in the World Index, which assesses the condition of political rights and civil liberties around the world, ranked Colombia as a free country in 2023. The index mentioned that despite a polarised campaign, the election of 2022 was free and fair (the report's score was 4/4). The score on the rule of law indicator was 3/4 because 'the justice system remains compromised by corruption and extortion', including the fact that 'the Constitutional Court has repeatedly been asked to mediate polarizing political disputes, especially with respect to the Special Jurisdiction for Peace (JEP), a parallel judicial tribunal that lies at the heart of the 2016 peace accord's transitional justice system'.

The instrumentalisation of political polarisation seen in social media includes 'disinformation' campaigns, such as those linked to the peace referendum.[12] During the

[11] The Special Justice for Peace is the Colombian transitional justice mechanism through which FARC members, members of the Public Force, and third parties who have participated in the Colombian armed conflict are investigated and put on trial.

[12] See a Social Science Research Council report on 'disinformation' campaigns related to the country's 2016 peace deal referendum and the 2018 Colombian presidential election, which found that 'well-known politicians are primarily responsible for disseminating "disinformation" in Colombia'.

2021 protests, the government's response included a campaign by the Ministry of Defense to identify 'fake news' during the protests in Colombia. This led to undue limitations on freedom of expression and contributed to the spread of 'disinformation'.[13] For example, mass media have been co-opted by large economic and political conglomerates and they face a hostile environment as a result of public figures and politicians criticising journalists. In rural areas, people receive biased information due to lack of access to media[14] or have different lived experiences due to the presence of armed actors that dominate their territory.

[13] As part of this campaign, the government monitored the digital public space, and that information served as a basis to launch stigmatising and criminalising discourses against the protesters. The monitoring was also used to control the physical space using the information they collected for actions such as 'anticipating acts of vandalism' and prosecuting people.

[14] Fundación para la Libertad de Prensa, (n.d.) *Cartografías de la información*, Bogotá: FLIP.

The state of content moderation in Colombia

Social media landscape in Colombia

In 2021, Colombia had a population of 51.265 million people: 50.9% female and 49% male, with a median age of 31.2 years: 18% of the population lived in rural areas and 11.1% were part of minority ethnic communities.

Internet penetration has continued to increase steadily in Colombia since the Covid-19 pandemic in 2020. According to the Colombian Communications Regulatory Commission, the penetration rate for mobile internet service was 75.8 per 100 inhabitants in September 2022 – 5.8 percentage points higher than in September 2021 – while the residential access penetration was 49.3 per 100 households for fixed internet.

According to the 2022 third quarterly report of MINTIC, at the publication date, 8.52 million people had fixed access to the internet and 39.2 million had access to mobile connections (13.8% through 3G connection and 83.9% through 4G connection).

According to the annual Digital 2022: Colombia report, the number of users connected to the internet in January 2022 was 35.5 million, with a penetration rate of 69.1%. Kepios's analysis indicates that the number of internet users in Colombia increased by 770,000 (+2.2%) between 2021 and 2022. This means that 30.9% of the population remained offline. Data from GSMA Intelligence shows that there were 65.75 million mobile connections in Colombia at the start of 2022.

Information published in Digital 2022: Colombia shows that at the beginning of 2022, there were 36.25 million users aged 18 and over using social media in Colombia, representing 93.9% of the total population aged 18 and over at that time. More broadly, 97.7% of Colombia’s total internet user base (regardless of age) used at least one social media platform in January 2023. Kepios and We Are Social analysis reveals that the number of social media users in Colombia increased by 2.8 million (+7.2%) between 2021 and 2022. In January 2022, there were 41.8 million social media user accounts, which would represent 81% of the population. It should be noted that the user numbers are based on active user accounts and may not represent unique individuals.

Colombians spend 10 hours and 3 minutes daily using the internet across all devices, of which 3 hours and 46 minutes are spent using social media, ranking fourth in the world. The same report stated that the most used platforms are WhatsApp (94%), Facebook (91.7%), Instagram (84.4%), Facebook Messenger (73.8%), TikTok (69.5%), and X (50.8%) (para 54). The Meta group owns three of the five platforms with the most users in the country (Facebook, Facebook Messenger, and Instagram). Despite the emergence of others, such as Pinterest and TikTok, Meta continues to lead the sector in digital traffic.

According to the report, the largest increase in platform users is of people aged between 25 and 34 years, with women accounting for 14.8% and men for 14.9% (para 53), who mainly use Facebook and Instagram for 'keeping in touch with friends and family' (para 53). Colombians in this age group are using social networking platforms not only for entertainment but also for communication and educational purposes.

People in South Africa, the Philippines, and Brazil spend the longest amount of time on the internet, with Colombia in fourth position (para 27). The most common reason for using social networks is searching for information (para 29), such as news or political content, with X standing out for this use. This could be a response to several phenomena such as the lack of public services infrastructure in rural areas or the zero-rating policies and regulations whereby users are given free access to certain content or applications – mostly Facebook and WhatsApp – without that access counting towards data caps in an individual's plan. It could also be a reflection of a generational shift of people preferring digital platforms.

A 2022 report by the Reuters Institute found a very high rate of mobile phone use in its urban-based sample. The report states: 'Online samples will tend to under-represent the news consumption habits of people who are older and less affluent, meaning online use is typically over-represented and traditional offline use under-represented. In this sense, it's better to think of results as representative of the *online* population.' Also, the more urban-based population gets their news more frequently online (86%, including social media) than from TV (55%) or print (28%). TikTok in particular saw a rise, especially among younger people. According to the report, these numbers can be explained by the accelerating digitisation in society during the pandemic.

In relation to news searches, the Reuters Institute report highlights that 60% of people access news online through social networks, 35% type a keyword or name of a website into a search engine, and 27% access news through search engines using words that refer to particular news items. Only 27% of those surveyed reported directly browsing a web page or news application to access news. However, 61% of the respondents reported they were concerned about what is real or fake online.

The report also found that news consumption in general went into a slight decline post-pandemic. In the midst of the pandemic, there was a constant search for information on the internet for guidance, but after the worst peaks of contagion were over and the vaccination programmes advanced, people began to become uninterested or tired of the type of information offered about Covid-19. The news media then had the challenge of competing with relevant information in different formats and facing the consumption of entertainment platforms and social networks. The report found that the survey had been affected by electoral debates. As such, misinformation fears revolved around political issues and it was found that 'memes have become a popular form of political expression on social media'.

Besides this Reuters Institute report, there is no other general report assessing Colombian citizens' trust in news found on social media or the concrete impact of social media on political or social issues. However, there are some surveys referring to trust in the media in some populations such as youth. For example, the Sixth Youth Perception Study, conducted in 2023 with people between 18 and 32 years old, mentions that 36% of respondents trust social media networks, 28% the media, and 15% digital influencers. Conversely, 62% do not trust social media networks, 71% the media, and 82% digital influencers.

Overview: Impact of content moderation and curation on human rights and conflict

In Colombia, local issues of structural discrimination, political factors, and the growing lack of trust in the media worsen the public debate and can promote the appearance of online content that negatively impacts human rights. This is despite the constitutional order that contemplated greater constitutional guarantees and mechanisms to make fundamental rights a reality, as well as several peace efforts.

Structural patterns of discrimination persist on the basis of gender, sexual orientation or gender identity, race, disability, and national origin, among others. Polarisation also persists around the efforts to achieve peace. Discrimination and polarisation are common in the content circulating on social media in Colombia, sometimes violating the fundamental rights of people and having an impact on peace and stability. To better understand the impact of content moderation and curation on peace and stability in Colombia, it is necessary to explore instances of 'disinformation' and online gender-based violence circulating on social media within the country.

Disinformation

'Disinformation' is a real problem and the tactics used to spread such content have become more sophisticated. Electoral organisations, like the Misión de Observación Electoral (MOE), have stated that electoral information and political advertising in Colombia have changed enormously since the emergence of social media and the possibility to 'segment, profile, and measure the reaction of audiences to certain communication actions'. They have noted that content often goes viral without users necessarily evaluating the veracity of the data they consume on social media, especially when it aligns with their opinions and preferences. News reporting by France24 showed that in the 2022 presidential election campaign, the 'disinformation' strategies were more sophisticated. Fact-checkers in Colombia, like ColombiaCheck, warned that voters have been targeted by increasingly 'refined' montages, 'millimetrically' manipulated videos, and even users impersonating them. In these scenarios, 'false and misleading information has the potential to "radicalize positions" about the candidates and "disorient" voters by taking advantage of "a growing distrust" in electoral authorities.[15]

The Global Disinformation Index (GDI) reported that the dissemination of 'disinformation' has disruptive and impactful consequences for Colombia. GDI's assessment of the Colombian news media market found that most sites (44%) fall into the medium-risk category, 41% of sites into the low-risk category, and 12% have a high 'disinformation' risk. The overall ratings are generally brought down by operational shortcomings, especially regarding transparent information about a site's ownership and funding structure, and other operational and editorial policies, such as source attribution guidelines and fact-checking practices.

Online gender-based violence

In terms of online gender-based violence, during the 2022 elections, several acts of digital violence against women candidates for Congress and the Presidency of the Republic were recorded. For example, Francia Márquez, a candidate for the Presidency of the Republic, was a target of repeated sexist, racist, and classist comments on social media. According to electoral monitoring organisations, such as the MOE, which publicly denounced any racist, sexist, or classist statement that affects the fundamental rights of any person, this type of speech against female candidates is systematically replicated and has increased its reach in different media, as well as on social media, to hinder the political exercise of women and racialised people.

In a 2022 report, the MOE warned about a general increase in acts of violence against women social leaders. The report identified specific violations against women leaders due to their gender. It was noted that, unlike what happens with male leaders, where threats

[15] Quotes from a personal interview with ColombiaCheck.

are directed exclusively at them, in the case of women leaders, threats generally include references to their status as women and threats against the people close to them.

Online gender-based violence is embedded in broader patriarchal structures in society. It is well documented in Colombia that women suffer a high level of violence. The fact that it occurs on the internet does not diminish the seriousness of the violence. The impact of this violence can be heightened by the notoriety of the victim, because of the work they do, or because they belong to certain groups, like journalists or female candidates.

This type of violence can seek to reduce the participation and visibility of women and the issues they raise on the internet and prevent their participation in the public debate. Karisma has found that sexist violence in journalism is pervasive, targeting the bodies, appearance, tone of voice, professional skills, and capacity of women journalists and communicators. The multiple systematic patterns of patriarchy produce a continuum of male-dominant violence that is normalised. Psychological and sexual violence and sexual harassment occur where the journalist works, and aggressors can act with impunity in the closed physical and digital spaces.

In particular, there are calls to end abuse and stigmatisation directed against those trying to clarify the historical truth about the conflict and the experiences of violence and serious human rights violations. Indepaz, a civil society organisation devoted to peace, argued in the interview conducted for this research that there is a connection between stigmatisation and 'hate speech' by the political leadership and the persistence and reconfiguration of armed violence in the farthest corners of the country.

Regulation of 'hate speech', 'disinformation', and online gender-based violence

The Colombian legal system has only partially dealt with the concepts of 'hate speech', 'disinformation', and online gender-based violence. According to article 13 of the American

Convention on Human Rights[16] – which is part of the so-called ‘constitutional block’ and thus serves to interpret constitutional rights and duties – ‘hate speech’ shall be considered as an offence punishable by law. In this context, the Constitutional Court of Colombia has emphasised that ‘for the content of a message to be considered hate speech, it is not enough that the message criticises a conduct, or that it is offensive to the criticised subject. It is also necessary that the content of the message incites hatred or violence, or to commit an illegal act against the subject.’

Like international law, the Colombian legal system does not provide for a legal definition of ‘disinformation’. Even though the concept was alluded to in Ruling T-627 of 2012, the Constitutional Court did not define it, and only stated that public officials have a duty when making public pronouncements in relation to truthfulness and impartiality for information, factual justification and reasonability of opinions, and respect for the fundamental rights of citizens.

There is no specific regulation governing online gender-based violence in Colombia. However, the law on prevention of and attention to violence against women and some criminal offences that exist in the penal code could be used by authorities to combat it. Seeking protection of the rights to image and privacy through a *tutela* could also be used in the absence of more specific norms on online gender-based violence, as recently proposed by the Constitutional Court. These mechanisms enable women to request protection measures ranging from priority psychological care to the removal of content that may be restricted under international freedom of expression standards. It should be noted that the Constitutional Court has recognised the importance of the problem and has twice called on Congress to regulate the phenomenon (rulings T-280 of 2022 and T-087 of 2023).

[16] Article 13 defines ‘hate speech’ to be prohibited as ‘any advocacy of national, racial, or religious hatred that constitute incitements to lawless violence or to any other similar action against any person or group of persons on any grounds including those of race, color, religion, language, or national origin shall be considered as offenses punishable by law’.

Although state regulation may influence content moderation, the main ‘regulators’ governing content moderation are the community standards of social media companies. In simplified terms, these standards typically lay down the types of content they allow or prohibit. The limits of each platform may be stricter than the requirements under international, regional, and national human rights standards.[17] It may also be the case that the way in which community standards are applied may not adequately respond to the discourses that circulate on social media. The permanence of expression that affects any person or population on social media or the removal of content that is thought to be legitimate can have negative consequences on online public debate.

Despite the assurances of platforms during the interviews for this research that freedom of expression is a priority, the following section highlights that significant challenges and shortcomings in content moderation remain in Colombia.

Lack of transparency and procedural remedies

There is a well-reported lack of transparency around the rules that apply to content moderation and how these are implemented. This raises issues in the Colombian context and is reflected by the statements of the stakeholders interviewed.

Although the perception of the interviewees about social media platforms generally differs, the lack of transparency was a shared concern in terms of both understanding the platforms’ terms and conditions and their content moderation practices. The interviewees referred to the fact that users know that when they sign up on a social media platform, they sign an agreement that is long and not easy to read – most do not read it. There is a widespread belief that content moderation processes on platforms are unclear and biased. Despite some platforms making efforts to provide more information and transparency in this regard (community standards of the most used social media

[17] For an overview of the regulatory regimes and standards relevant to content moderation and how human rights apply to them, see the Content Moderation And Freedom Of Expression Handbook.

platforms in Colombia are available in Spanish), they still fall short of enabling users to fully comprehend content moderation policies and decisions.[18]

The interviewed stakeholders highlighted that it was unclear how the community standards apply to different categories of potentially 'harmful' content and how they are moderated by the platforms.[19] What complicates matters from a user perspective even more is that they have to consider a multiplicity of policies within each platform in order to understand what content is allowed on each of them.

In terms of how the policies are enforced, it appears that no platform is ready to provide the level of information on content moderation practices in Colombia that would help

[18] The reporting practices about policy enforcement vary among digital platforms. While the transparency reports of some platforms distinguish between content moderation under the platform's community rules, legal removal requests, and reports of intellectual property violations, others make clear that the total number of removals they provide may be the result of a combination of their enforcement of community guidelines and government requests. Some platforms only report total content removed on the grounds of community rules enforcement. This is the case for Facebook, YouTube, and TikTok. However, that indicator is not currently traceable over time. TikTok, X, and Google transparency reports only mention the total number of videos removed for violating their guidelines. Transparency reports for Meta, LinkedIn, Snapchat, and Pinterest do not contain disaggregated information for Colombia. Data on legal content moderation is interesting: according to Meta's content restrictions based on local laws, between July 2017 and June 2022, the platform restricted 1,078 pieces of content based on Colombian laws on Instagram and Facebook. The report also specifies that 90% of content restrictions due to legal requests concern Facebook posts, whereas on Instagram it is approximately 70% content and 30% accounts. Other details are provided but not enough to understand the scope of the situation. For YouTube, Google transparency reports have data on content moderation derived from community rules enforcement and requests by the Colombian authorities. YouTube also has data on requests for content moderation by local authorities.

[19] Despite the large number of policies, and the increased effort to make them more accessible, their scope is not completely understood by users and the number of different policies that prohibit different types of content complicates matters. To illustrate this point, Meta reports the enforcement of 14 content policies on Facebook, 12 on Instagram, and 22 policies overall. X lists 16 policies related to content. YouTube lists over 22 policies related to content. TikTok lists 26 policies related to content.

users obtain a meaningful understanding of the local application of community rules, nor are they willing to provide the data that would help to follow and serve as an oversight mechanism for government requests. The absence of country-level information, data classified by the volume of content removed, reasons for removal, type of moderation, origin of moderation, number of appeals received or their outcomes, and local rules in relation to moderation rules, among other indicators, is a major barrier for advocacy at the local level.

The representative from ColombiaCheck, one of the two Colombian fact-checking organisations that are signatories to the International Fact-Checking Network (IFCN) at Poynter, stated that the first step that platforms should take is to provide more information about their policies, beyond the initial agreement with users. The ColombiaCheck representative also pointed out that 'when platforms start investing in teaching methods, they need to start transferring knowledge about their own policies and which content for them is qualified as xenophobic or that can motivate further attacks'.

The interviewee from the feminist organisation Artemisas mentioned, 'The process should not get to the elimination of a tweet – for instance – without explaining why and how that decision was taken, without teaching. I think that many of those rules on the platforms are not understandable for people.'

Even if platforms have made reasonable progress in providing information and explanations, it is not enough to effectively explain the local impact of their moderation and curation practices. For example, while social media companies indicate that they comply with the laws of the countries they operate in, the extent to which national laws are enforced and/or impact their decisions is not clear, including in Colombia.

The lack of transparency can also impact the effective ability to resort to remedies against content moderation actions. Some interviewed stakeholders expressed their concerns about more structural problems and the absence of adequate remedies to challenge content moderation decisions. For example, a spokesperson from freedom of expression organisation El Veinte said, 'In content moderation, certain aspects mimic informal justice systems, but they lack the fundamental elements and due process guarantees that such systems provide to parties. I think it is necessary to establish clear, efficient, and concerted standards.'

A representative from the think tank Linterna Verde pointed out that, in practice, due to the amount of online content circulating on social media, digital platforms do not have the capacity to comply with their terms of service. He recalled a well-known moderation case that is currently with the Colombian Constitutional Court:

> *The case of Esperanza Gómez summarizes well the issue at stake: the content she posted on Instagram caused the deleting of her account without her receiving any explanation of the rule she broke. She also did not receive any response to her appeal even after sending several follow-up emails. That led her to create a new account and to start building her audience from scratch.*

An emerging concern is curation practices, which are perceived as particularly unclear, and interviewees believe they are being censored. Challenges in terms of transparency, and thus effective ability to resort to remedies, are especially pronounced when it comes to content curation or content moderation measures that are less evident than content removal or account suspension, for example restricted access to the service, demonetisation, or the placement of warning messages over certain content.

The lack of understanding by users about downranking and other curation and moderation practices was mentioned by the representative from the Argentinian and regional academic Center for Studies on Freedom of Expression and Access to Information (CELE; Centro de Estudios en Libertad de Expresión y Acceso a la Información). Their spokesperson said that 'these are very opaque practices that move outside the radar and on which civil society or academia can access very little information. Additionally, these practices are entirely discretionary. The reasons that trigger this type of practice are not reported and are not clear.'

The Uruguayan and regional freedom of expression organisation Observacom argued that downranking, in practice, has the same effects as content removal because it prevents users from viewing the content and is not subject to notification to the user by due process or remediation. He stated that 'content moderation has been studied deeply, but the same has not happened with content curation'.[20]

Similarly, civil society actors connect some of their problematic experiences on social media to consequences of content moderation that are difficult to detect.[21] The civil rights organisation Temblores, for instance, referred to the practice of shadow banning, where users have their content hidden or reduced in visibility without them being informed by the platform. This issue of feelings of censorship was also mentioned in the Guns versus Cellphones report by Karisma that studied the social protest in 2021.[22]

The interviews demonstrated the need to better explain how content that might violate the companies' rules is subject to different types of measures or curation because takedowns are no longer the only sanction. El Veinte stated that 'we also need more clarity about other forms through which platforms moderate the content because suppressing or amplifying are not the only ways algorithms enable the hiding of certain expressions; sometimes it is not that the platform removed the post but just put it away from your sight.'

[20] Within the definitions of content curation and content moderation used in this report, the downranking of certain content (beyond content removal) can, at times, be considered both a content moderation and a content curation measure simultaneously.

[21] When interviewees talked about content moderation, they did not specify which content moderation action may be taken (i.e. community rules violation, copyright infringement, or state requests). It is interesting that when dealing with content moderation during the social protest, as described in Case study 1, the feeling of censorship was often tied to requests by authorities.

[22] The report Guns versus Cellphones explores the feeling of censorship that citizens had during the protest, which was exacerbated by other problems in content moderation and lack of transparency at the platform level.

Stakeholders would like to see a more open debate, including representatives of social media platforms, about content moderation to enable accountability. According to the media development organisation Fundación Gabo:

> *I think this should become a matter of public debate and that we should find ways to not only have transparency, but also to have accountability and responsible management of the systems for the public, because they are working with public goods. That it is not only the information; it is also people's private life. It is peace. Then I do believe that beyond the fact that they have developed the technology and that they try to be responsible, it is definitely an issue that involves us all.*

The interviewee from the media outlet League Against Silence said:

> *I believe that moderation should be discussed, and it should be a more public issue that we have to understand better. What are the rules? That is something that neither an audience nor content creators have any idea of. Here we are guessing things, by trial and error, but it should be very public, moderation should be a very clear agreement in which freedom of expression foundations have something to say, not an absolutely unilateral decision.*

Similarly, La Silla Vacía indicates that 'From journalism, I think it would be very useful to be part of the discussion to intervene in the opacity that exists around the moderation of content. From there we can also fight against disinformation.'

Finally, it is worth mentioning that the interviews with civil society and media stakeholders showed that a relevant issue for actors in Colombia is the transparency of advertised content versus organic content. In other words, how can paid content – that users have paid the platforms to post – be recognised compared with content that has been posted by the users.

There is obviously a need for increased transparency obligations. This concerns not only the platforms themselves, but also state actors that may use different channels to influence content moderation. As will be explained in the section related to the state legal powers to request content moderation, transparency obligations have not been developed in Colombia – even in the locally regulated space of telecommunications operators where the legal faculty to block content has existed for over 20 years.

Shortcomings of automated content moderation

It is only possible for platforms to moderate content at scale if they rely to some extent on automated content moderation, because human moderation would be unable to process the amount of information generated by users. While human review is essential to interpret specific content in the context of cultural sensitivities, beliefs, or value systems, monitoring online content in real time is a mammoth task that may not be entirely feasible without technological assistance and its new challenges, including concerns about content moderators' situation in Colombia. At the same time, these tools can pose a serious risk from a freedom of expression perspective.

From a user's perspective, the automation of content moderation is clearly an emerging problem and is linked to a number of challenges, including lack of transparency around content moderation, lack of linguistically and culturally nuanced decisions, and the role in hindering media actors and public interest reporting by removing content related to topics such as extremist groups or human rights violations.

For the interviewed stakeholders, a key element of understanding automated content moderation relies on the information that platforms provide about the automation process and how it is carried out in Colombia. However, there is still a significant lack of transparency on the extent to which automatic tools are employed in content moderation. According to Meta, 90% of what is identified as problematic content is moderated through automated systems. Meta informs on the proactive rate detection on content, that is, the content or accounts acted upon to apply content moderation decisions before users report them. However, similar data cannot be found for other platforms.

The spokesperson for La Silla Vacía, the other IFCN fact-checking organisation, mentioned:

> *I think the content moderation processes are not clear, it is not clear who makes the moderation decision, and to what extent these are driven only by algorithmic decision-making, or the impact that moderation has on the moderators. Besides, I think social media platforms should continue to put efforts into making their rules about allowed and prohibited content clearer.*

Content moderation and public interest reporting

Journalists in Colombia have a good understanding of the importance of digital platforms for the distribution of information, particularly to people outside the main cities. As a result, media outlets have increased their digital presence and started disseminating content on Facebook, Instagram, TikTok, X, and other platforms.

However, relying on social media platforms for distribution can have a negative impact. The shortcomings of automated content moderation systems in preventing the circulation of 'harmful' content have, at times, complicated matters for media actors. As explained by League Against Silence:

> *There is a very important journalistic value, which is to call things as they are, to say femicide, to say homicide, to say reinserted (ex-combatant). But in platforms we have had to silence those words in videos or write them with numbers so that the algorithm does not 'punish' the content. We have had to use euphemisms that also end up casting a cloak of doubt over some struggles. That has been difficult.*

This is a serious complaint that should be addressed. There are important issues of public debate that, in order to have a presence in the digital public space, end up going through mechanisms of disguise that allow them to avoid content moderation. A similar situation was reported by several people covering the 2021 protests (see Case study 1).

The strategies that the press must implement to ensure its content remains online, despite the public interest it generates, are reminiscent of the debates that the Ministry of Culture described as a digital black hole.[23] One of the examples that the Minister of Culture provided to illustrate this black hole was the disappearance from social media of the voices of the FARC guerrillas during the peace process.

The Ministry also pointed to another separate issue that can prevent reporting on and research into public interest matters: platforms' lists of banned extremist groups. The Ministry explained that for anyone interested in the 2016 Colombian peace process and agreement, the social media accounts of its protagonists would be an important and unique primary source of information. However, the Minister warned that because the FARC was on the list of international terrorist groups, social network platforms – such as X, Facebook, or YouTube – frequently blocked content and cancelled FARC's accounts, despite the fact that they were parties to the peace process taking place with the government.

[23] The response to a request that Karisma made to the Directorate of Copyright – asking for all comments received during the procedure of the bill to reform the copyright law in Colombia in 2018 – includes a letter from the Ministry of Culture that explained the 'black hole concept' and their petition to include the legal digital deposit for the National Library to tackle the issue. The request by the Ministry that referred to the digital black hole was meant to avoid losing historical Colombian records due to the ephemeral nature of the internet and the inability to preserve web pages and social media content containing such records. Copyright law was not providing necessary support to the institution.

Case study 1: Social protests in 2021[24]

During social crises, platforms face important challenges to their decision-making processes on content moderation, content curation, and appeal mechanisms. In such cases, the described lack of transparency and the effects of automation are prone to worsen the situation. The volume of information that is produced during moments of social unrest, the concentration of content production and publication by single users at certain times – and its nature (for instance, content that denounces police abuse can be classified as violent content that is prohibited by community rules) – place additional stress on the content moderation processes during events such as protests. Moreover, people linked to those events are particularly critical of content moderation practices which they perceive as unfair or ill-motivated.

During the 2021 social protest in Colombia, platforms' content moderation, curation, and appeals systems appeared to malfunction, affecting people who were linked to the protests or were providing information about them.

On 28 April 2021, after President Iván Duque submitted a tax reform bill before Congress, a massive citizen protest movement known as the 'National Strike' began in Colombia. From then until 15 June 2021, marches and activities were held in various cities throughout the country.

Amid this social and democratic unrest and confrontations with law enforcement, social media posts documenting the excessive use of force by law enforcement against citizens

[24] This summarised case is based on Guns versus Cellphones by Fundación Karisma and includes quotes from interviews made by FLIP and shared with Fundación Karisma. The consent of the interviewees used in this report was obtained.

or the attacks by citizens against law enforcement officials and local infrastructure went viral.

Reports also emerged about possible actions by the state to limit the rights to freedom of expression and assembly, access to information, and privacy of citizens, exercised on or through social media platforms.

The measures adopted by the state included cyber-patrolling social networks. This allowed the state to profile people and counteract speech that deviated from the government's narrative. Other measures included internet shutdowns in places of high concentration, searching and reviewing content on protesters' mobile phones, often without their consent, and ordering internet service providers to block pages containing information about members of the security forces.

Jahfrann, a Colombian freelance photographer living in Cali,[25] reported the tension of documenting this situation on social media during those days. Many people were denouncing unexplained specific internet shutdowns affecting their capacity to use social networks. He said:

> *Let's say that right now it has calmed down because there is already a national and international outlook, but when it just started, it seemed that the entire network was lost. Siloé – a neighbourhood in Cali – was off one day for five hours, five hours in which nothing entered or left the district. I mean, I was on a walkie-talkie with the human rights people, and we didn't know if they were alive or dead – there was simply no signal – I saw the mobile unit with big equipment and an antenna. I*

[25] Reported in a testimony collected by FLIP and published in Guns versus Cellphones.

> *don't have a photo. Much of what I shared was 'atemporal' because there was no signal there.*

The Foundation for Press Freedom (FLIP) reported that on 6 May, the X account of Noís Radio, an independent media in Cali (@noisradio), had been 'repeatedly restricted'.

Noís Radio appealed the blocking decision, but the restriction was imposed several times, leading them to declare that social media appeal proceedings were ineffective. The media outlet held that the labels imposed on its account ended up silencing its voice, so that the warning notices were visible to users.[26] Accounts restrictions and blocking led some journalists to resort to alphanumerical code to circumvent the algorithm.

Content that is posted during times of social upheaval is of the utmost importance because it serves as documentation of human rights violations. Alejandro Gómez from the League Against Silence was working at the time on the digital portal 070, a media outlet that reconstructed violent events during protests, including the murder of protester Dilan Cruz in 2019. He noted that they 'did not have the capacity to cover police violence', making it impossible to verify or challenge official accounts of the events.

Even though instances of restriction of expression on social media platforms during the 2021 protests were identified, Meta only explained its software problems, while X did not explain its failure at all. Meta's decision to publicly admit the software problem and its consequences is good practice that should be adopted by other platforms in similar cases. However, Meta's explanations were not related to content moderation challenges and mitigation mechanisms during the protests.

[26] Reported in a testimony collected by FLIP and published in Guns versus Cellphones.

Additionally, journalists and librarians believe content moderation affects the information landscape. Certain situations produce an over-removal effect that impacts society at large due to the reliance of digital media outlets and memory sources on social media information.

Context in the process of content moderation and curation

There are a number of research reports and media investigations about content moderation and curation in Colombia. Examples of such reports include one by Karisma regarding copyright content moderation, the analysis of cases made by Linterna Verde, and a report regarding moderation and freedom of the press by Observacom. These reports confirm that content moderation decisions need to be interpreted in the local context and to consider the various legal, political, cultural, and linguistic specificities, or the existing special protection of certain population groups locally.

When we addressed questions about local context to the interviewed representatives of social media platforms during this research, Meta mentioned that they incorporate different voices and points of view when drafting policies. X emphasised the global aspect of the conversation, saying that the teams working on detection mechanisms are trained to consider diversity and context. In the same vein, Google stated that diversity is considered in the development of their policies.

Meta and Google have local offices in Colombia that go beyond marketing purposes; they have policy officials and good relations with civil society. X used to have a team located in Mexico City that engaged on policy issues with Latin American civil society, but this was dismantled in 2023. There is no information about the teams that moderate local content or where they are located. Up to the present day, these companies have developed direct channels with some local civil society organisations in special circumstances (for example, during the 2021 protest or the 2022 elections). These channels are activated during such times, and the companies have mentioned that they do take special measures that include nuanced local content moderation, as described later in this document. A journalist revealed that there were TikTok moderators based in Colombia; however, it is unclear what role they play in moderating content locally.

Moreover, some platforms with a presence in Colombia review case studies when seeking to address the context challenge. This is done by the Meta's Oversight Board and also by a YouTube initiative that selects cases of interest for the national context. However, despite a recent pandemic and serious social conflicts in Colombia, these tools have only been used once. For example, Meta's Oversight Board selected and decided on the case of the use of the word 'marica' (see Case study 2), while YouTube described how they decided not to eliminate from their platform a documentary critical about former President Álvaro Uribe. These efforts by platforms are not enough in terms of transparency; they are also insufficient in providing information for platforms and users to act upon.

Interviewees also made a connection between automated systems and the lack of linguistically and culturally nuanced decisions. A representative from Plurales, a think tank at Rosario University, said, 'This is a task that cannot be done by a machine. For instance, the word "marica" has many meanings: it can be an endearment but also a derogatory word, but not necessarily homophobic but can have homophobic and transphobic uses. Machines cannot understand these differences.' The interviewee pointed to the aforementioned Colombian case that was selected by the Oversight Board (see Case study 2), a case in which the President at the time was called 'marica' during the protests of 2021.

Linterna Verde added on this topic: 'when moderation becomes a series of forbidden words…without sufficient contextual knowledge and without assessing the situation in detail, it leads to mistakes', thus pointing to the problem of the context.

For Meta, this case proves that it is possible to correct a mistake made in content moderation by considering the local context. Meta stated in the interview that 'the decision of the Oversight Board related to content moderation of posts that used the word "marica" in Colombia will be taken into account in future cases. Content moderation is not static, it evolves. If excesses can be found in the interpretation of a content moderation rule, the platform can then limit them.' Aligned with this, Meta stated that 'moderation activities are as dynamic as content itself. If societies explore new trends and social movements, the content moderation landscape continually changes. It is always changing.'

This is not the only way platforms provide context to content moderation. Regarding the contextual application of community guidelines, there is a caveat for social upheaval periods. Platforms shared that they have special procedures during elections, and similar approaches were adopted during other times of social conflict. X informed the researchers that the approach to taking measures during normal times is not the same as during atypical times. During humanitarian crises, pandemics, or election periods, X adopts a crisis misinformation policy.[27]

Meta pointed out that much of the work they do is anticipating political events or periods of political uncertainty. They can establish an Integrity Product Operations Center temporarily, 'which is a working group composed of subject matter experts from our product, policy, and operations teams, [that] allows these experts to more quickly surface, triage, investigate, and mitigate risks on the platform', according to the Quarterly Update on the Oversight Board. During these times, Meta also takes advice from local partners (through their Trusted Partner Programme) who understand the specificities of each context, and establishes conversations with electoral commissions.[28] Finally, Meta also mentioned the existence of a third-party fact-checking programme during elections.

Another issue related to context in content moderation is its disproportionate impact on certain groups over others, as Linterna Verde described:

[27] X (2022) *Crisis Misinformation Policy*, accessed 29 October 2023.

[28] For example, the Colombian National Electoral Council (CNE) signed a Memorandum of Understanding with Facebook for the October 2019 elections. A similar agreement was signed in Mexico by the National Electoral Institute.

> *The platforms base their rules on hate speech against internationally protected categories such as nationality, sexual orientation or race. They, however, neglect to protect some other categories which are protected in the Colombian context of an armed conflict: notably human rights defenders, ex-combatants or journalists who, among others, talk about the conflict and are therefore much more exposed to targeted attacks.*

In a report on violence experienced by women journalists in Colombia, journalists mentioned that they did not use the response mechanisms available on the platforms due to a lack of knowledge about their existence and functioning or because they believe that such mechanisms are ineffective. The report recommended that companies 'communicate in a more effective and accessible way, and in local languages, the response mechanisms available to address the gender-based violence that occurs on their platforms'. It also recommended that companies should conduct regular consultations to improve their policies and practices.

Content moderation when it involves figures of local public recognition is also a matter of concern for the stakeholders interviewed. Both the case selected by the Oversight Board (see Case study 2) and the case selected by YouTube about the request to take down a documentary on former President Álvaro Uribe speak to how public figures (the Colombian President in the Facebook case or a political party and former President in the YouTube case) must have greater tolerance for public scrutiny and questioning by audiences, and also greater responsibility when they create content for distribution.

This nuance is connected to the reflections of some of the interviewees. How content moderation deals with the responsibility of public figures is a concern for local stakeholders. *Sentiido*, a lesbian, gay, bisexual, transgender, queer, and intersex (LGBTQI+) oriented media outlet, stated: 'It seems to [us] that a discussion on the responsibility of visibility is missing. In other words, a person whose visibility in the media is so important has a responsibility for the promotion of non-violent speeches against communities that have been historically marginalized.'

Case study 2: 'Marica' slur during social protests

The United Nations Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression published a report on user-generated online content moderation describing a paradox that platform content moderation faces: even if 'companies emphasize the importance of context when assessing the applicability of general restrictions…meaningful examination of context may be thwarted by time and resource constraints on human moderators, overdependence on automation or insufficient understanding of linguistic and cultural nuance'. Content moderation is a complex ecosystem to deal with, and the Colombian content moderation case selected by Meta's Oversight Board and decided in October 2021 describes this situation:

> *In May 2021, the Facebook page of a regional news outlet in Colombia shared a post by another Facebook page without adding any additional caption. This shared post is the content at issue in this case. The original root post contains a short video showing a protest in Colombia with people marching behind a banner that says 'SOS COLOMBIA'.*
>
> *The protesters are singing in Spanish and address the Colombian President, mentioning the tax reform recently proposed by the Colombian government. As part of their chant, the protesters call the President 'hijo de puta' once and say 'deja de hacerte el marica en la tv' once. Facebook translated these phrases as 'son of a bitch' and 'stop being the fag on tv'. The video is accompanied by a text in Spanish expressing admiration for the protesters. The shared post was viewed around 19,000 times, with fewer than five users reporting it to Facebook.*
>
> *Facebook removed this content as it contained the word 'marica'. This violated Facebook's Hate Speech Community Standard, which does not allow content that*

> *'describes or negatively targets people with slurs' based on protected characteristics such as sexual orientation. Facebook noted that while, in theory, the newsworthiness allowance could apply to such content, the allowance can only be applied if the content moderators who initially review the content decide to escalate it for additional review by Facebook's Content Policy team. This did not happen in this case. (case summary by Meta's Oversight Board)*

The Oversight Board overturned Facebook's decision to remove the post. The Oversight Board concluded that even though Facebook's removal of the content appeared to follow its Hate Speech Community Standard, the newsworthiness allowance should have been used to allow the content to stay online. According to the public comments and expert advice, the word 'marica' has several connotations and could be used without having discriminatory intent. Experts explained that the term had attained widespread usage in Colombia to refer to a person as 'friend' or 'dude', and also as an insult like 'stupid', 'dumb', or 'idiot'. However, there was consensus that its origins were homophobic and that it was used particularly against gay males. The Oversight Board pointed out:

> *The newsworthiness allowance requires Facebook to assess the public interest of allowing certain expression against the risk of harm from allowing violating content. As part of this, Facebook considers the nature of the speech as well as country-specific context, such as the political structure of the country and whether it has a free press.*
>
> *Assessing the public interest value of this content, the Board notes that it was posted during widespread protests against the Colombian government at a significant moment in the country's political history. While participants appear to*

use the slur term deliberately, it is used once among numerous other utterances and the chant primarily focuses on criticism towards the country's President.

The Board also notes that, in an environment where outlets for political expression are limited, social media has provided a platform for all people, including journalists, to share information about the protests. Applying the newsworthiness allowance in this case means that only exceptional and limited 'harmful' content would be permitted.

State legal powers to request content moderation

There is no specific regulation governing content moderation comprehensively in Colombia; however, there are scattered regulations that require the implementation of blocking orders at the telecommunications operator level. These regulations are not just for Child Sexual Abuse Material (CSAM);[29] they also exist to combat gambling[30] and include all other legal orders coming from administrative authorities.[31] In addition, there can be judicial and administrative blocking orders – such as precautionary measures in a *tutela* action – that take place during states of emergency and exception.[32] All such orders issued to telecommunication operators are channelled via MINTIC.

According to international and constitutional standards, whatever their legal basis or the authorities in charge, any blocking or restrictive orders have to meet the three-part test of legality, legitimate aim, and necessity and proportionality. Nonetheless, while there exists no detailed legal analysis of the above-mentioned scattered regulations from a freedom of expression perspective in Colombia, it is worrisome that some blocking orders have been as broad as to almost block access to entire webpages. This was the case of RapidShare for users of Telefonica in 2010, and it was almost the case for all users in Colombia with InternetArchive in 2021 during the social protest. The InternetArchive website ended up

[29] According to articles 7 and 8 of Law 679 of 2001 and articles 5 and 6 of Decree 1524 of 2002.

[30] According to article 38 of Law 643 of 2001.

[31] Some administrative authorities in Colombia have judicial powers to block content by giving orders directly to the telecommunication operators. This happens in cases of data protection, industrial property infractions, and consumer protection infringement by the Superintendence of Industry and Commerce (SIC) according to article 54 of Law 1480 of 2011, or in cases of copyright infringements by the National Directorate of Copyright.

[32] According to Law 1341 of 2009. During such states of emergency and exception, the government will need to deliver a specific decree indicating the scope, which will be reviewed by the Constitutional Court to determine if the derogations are in line with human rights standards.

only being blocked by Avantel and Emcali because the other companies did not comply with the order after their own analysis showed it was disproportionate.

On the other hand, social media platforms report requests to take down content on legal grounds in their transparency reports. This is something Google, X, and Facebook representatives confirmed in the interviews. The problem is that those reports do not have enough information to analyse which regulatory frameworks they are applying, or whether their content moderation rules contradict applicable local standards. This becomes more complex when the transparency report of X aggregates the figures of their community rules enforcement with government requests.

Existing regulations do not oblige social media platforms to be fully transparent on content moderation practices or require the government in Colombia to provide information on requests (such as removal requests) raised to social media platforms. When the researchers asked about content moderation and curation regulation, MINTIC informed them:

> *The telecommunications network and service providers regulation must block the websites with proscribed content from the lists with URLs containing child pornography*[33] *issued by the Criminal Investigation and Interpol Directorate (DIJIN), as published on the MINTIC website.*

MINTIC added 'the legislator did not attribute to this entity the competence to regulate, monitor and control the provision of contents and applications or technological platforms'. Even if this is true, as already described, blocking content orders in current telecommunication regulation is not just for CSAM, but MINTIC does not mention its broader role regarding its various content blocking abilities. The written statement also says that MINTIC understands that any regulation of information or content on social

[33] This is an old regulation (in terms of internet standards) from 2001, and it uses the expression 'pornography' rather than CSAM.

media must be approached with the greatest caution, as it could imply limitations on fundamental rights.

Although no platform regulation is yet in place, this may change soon. It is notable that the vast majority of the recent state regulatory proposals on social media refer to the control of lawful content – content that is, *prima facie,* protected by freedom of expression.

Recent bills have sought to impose filtering and blocking mechanisms to prevent potentially 'harmful' content for minors, or to prohibit any speech about prostitution or promotion of sexual activities on social networks that deviates from the regulators' view of the phenomenon. Some bills are intended to impose functions on administrative authorities to promptly and quickly block paid content that may affect women's political rights. These cases align with regulatory pressures in Latin America in countries such as Argentina or Peru, potentially indicating a broader trend in the region.

CELE mentioned that there is a global trend where 'many of the projects and laws enacted today, including those that regulate processes, involve some re-negotiation on the legitimate limits to freedom of expression on platforms'. The bills that have been introduced to the Congress in Colombia share this characteristic – the definition of what is undesirable content is generally ambiguous and it often involves active monitoring functions of social networks by the platforms and the state. For example, the law on violence against women in politics is under debate and the bill is, at the time of writing, under assessment at the Constitutional Court because there are different viewpoints on whether its content moderation dispositions comply with the Colombian Constitutional framework.

A representative from Red Papaz, a child protection organisation that takes part in the process of blocking CSAM with telecommunication operators in Colombia, insisted that they would like to see more action at the platform level. The interviewee said that 'it is necessary to understand that the abuse and exploitation of images of a child is used for trafficking or to make money – this is a very complex crime. In these cases, the rights of a child have been violated.'

State use of community rules to restrict content and accounts

The mapping of government regulations on platforms' content moderation in the 2018 report on user-generated online content moderation included government demands that were not based on national laws. The report explained that these demands often included pressure on companies to accelerate content removals through non-binding efforts and had evolved into coordination agreements between companies and states that can harm privacy and freedom of expression.[34]

The Rapporteur observed that states increasingly rely on social media platforms' terms of service to request the removal of content they find objectionable. State practices in requesting the removal of lawful content that can be regarded as extremism 'raise the prospect that states may rely on private terms of service to bypass human rights or domestic law norms against content restrictions' (para 53).

There is no evidence that the Colombian government has formal agreements with the platforms to coordinate content monitoring or removals. Nonetheless, platforms do have ways to address local contexts in a more tailored manner, which may include cooperation with public authorities. For example, during elections, digital platforms may formally coordinate with electoral authorities on voters' information on elections. There are also special projects such as the YouTube Priority Flagger. This programme includes government agencies and NGOs as partners because 'these agencies and NGOs are particularly effective at telling YouTube about content that violates our Community Guidelines'. In Karisma's experience – during the social protest and in more recent election periods – platforms also offer special and more expeditious reporting channels for partners to provide specific information.

[34] The report mentioned agreements to combat content that is 'offensive' (Pakistan) or 'incites violence' (Israel). It also listed the EU Code of Conduct on countering illegal 'hate speech' online, signed by four major companies to remove content, committing them to collaborate with 'trusted flaggers' and promote 'independent counter-narratives'.

Some platforms' transparency reports do recognise how the Colombian state uses community rules to ask for content removals;[35] however, the reports only provide a glimpse into those actions. Again, the absence of information makes it difficult to understand the grounds for the requests, how they analyse the request under human rights standards in Colombia, and the list of authorities responsible for such requests.

Without legal powers, Colombian authorities had been sending requests to the platforms using the community rules. From the data provided by some of the platforms on Colombia,[36] during the Covid-19 pandemic, the National Institute of Surveillance of Medicines and Food (INVIMA) used community guidelines to control the circulation of information. During the 2021 protest, the framework of violence and terrorism was used for the same purpose.[37] Case study 1 gives more data on how people felt during the 2021 protest and how this impacted them, as they speak of a sense of censorship and identify the state as an actor.

When interviewed, the representative from CELE warned that the role of the state had been underestimated in this regard. The interviewee said that the state is not only an instigator of content moderation asking the platforms to take down specific content or accounts; it also seeks to solve societal problems that it has failed to tackle through content

[35] This is the case for YouTube, where Google transparency reports have data on content moderation derived from requests by local authorities for content moderation on YouTube.

[36] For the period from 2017 to 2019, Meta indicates that these requests mainly respond to two themes: (1) items alleged to violate laws related to the sale of regulated goods and (2) private reports of defamation. From 2020 to June 2022, during Covid-19 restrictions, the majority of the requests came from the Colombia National Food and Drug Surveillance Institute (INVIMA) pertaining to unlawful public announcements regarding unregistered health products.

[37] Between 2011 and June 2022, YouTube was requested by the Colombian authorities to remove 73 videos. The most common reason was defamation, followed by privacy and security. National security was the next most reported reason and was mainly used between January and June 2021 during the social protests. Lastly were copyright and trademark grounds.

moderation. States are not resolving society’s structural problems, such as discrimination or violence, resulting in unsatisfied demands for protection being transferred to requests for content or account blocking by the platforms. CELE stated that the failures of public institutions to address complex problems and the way in which officials have behaved within public debates have led to the fact that people’s ‘hope [about the unmet demand on speech] is placed on the messenger, that is, on the intermediary’.

Although the numbers reported by the platforms on the state use of community rules to control content are low in Colombia, the lack of reasoning and publicity behind the requests is problematic. El Veinte also fears that government authorities may increasingly try to tighten control over the ‘digital public square’ and demand content moderation actions from the platforms, with negative implications for freedom of expression.

A local coalition on content moderation and freedom of expression

This research investigated the feasibility of forming a local coalition for content moderation and freedom of expression in Colombia. The research has found that:

1. In Colombia there are a series of initiatives that bring together organisations and movements that, whether directly linked to digital rights or formed in other areas of action, can be combined to work together on concrete issues related to content moderation.
2. Colombia has organisations that work in coalitions and maintain dialogues, whether at a national or regional level.
3. There are concrete points where research participants and other multistakeholder actors converge, and which can be worked on to form common interests.
4. The government and social media platforms are the most absent actors in this debate. Advocacy work will be needed to demand more rights-respecting practices when it comes to content moderation and curation. The most effective approach to achieve this will involve local stakeholders finding ways to collectively interact with the platforms.

This section examines the proposal of forming a new coalition in Colombia and how it can be done, taking into account the specificities of Colombian civil society and existing coalitions dealing with content moderation issues. It also assesses the proposal of a network of existing organisations and coalitions as a key driver to establish critical points to be taken to relevant actors in the process.

Forming a potential coalition

During the course of this research, relevant stakeholders working at the intersection between online content moderation and freedom of expression in Colombia were mapped out. Interviews (see Annex B) were conducted with a variety of different stakeholders to assess how local actors understand content moderation on social media. Research has found that there is currently no agreement across local stakeholders regarding what constitutes potentially 'harmful' content on social media and on key issues affecting content moderation. A representative from Plurales mentioned:

> *It is very difficult to achieve a universal definition of the concept of harm, for instance, because conservative organisations and users who see available information on sexual orientations and diverse gender identities can think that this content is potentially 'harmful' for their kids. So it is important to understand not only how I would consider something to be harmful, but also how society is organised in terms of power balances.*

Échele Cabeza, an organisation focused on drug abuse awareness, stated:

> *We work with psychoactive substances, which is a super transgressive theme. For instance, we can publish a video about how a person can inject themselves taking fewer risks. For me that is health information but for others that is promotion of drug use, so the perception of risk and danger is something that changes for each person.*

While the interviews showed a general perception that harm may relate to expressions that may convey 'hate speech' and violence, some interviewees (see the two quotes above) highlighted that certain topics are more complex. What some may perceive to be essential information that deserves to be circulated, others may find to be harmful information.

It is also important to understand that currently in Colombia discussions over content moderation mainly arise in the context of debates on specific topics, such as gender violence or elections. These experiences should be considered in any initiative aiming to address content moderation in Colombia.

Against this backdrop, the building blocks to create an informed group of stakeholders dealing with and advocating on key issues related to content moderation in Colombia are:

1. identify concrete topics or areas to work on as a group;
2. identify key targets of advocacy calls; and

3. harmonise the understanding of local stakeholders on content moderation through capacity building and exchanges within the group.

The following sections will provide further analysis of these requirements, identify the needs of various stakeholders to reach a common understanding of content moderation, and present some solutions based on the local reality in Colombia.

Content moderation: An open debate across local actors

Currently, civil society organisations in Colombia generally recognise that users' experiences on social media can be different and that there are certain groups more affected than others by 'harmful' online content such as 'hate speech' or 'disinformation'. In line with the mission of their organisations, some interviewees mentioned the need to regulate speech that is explicitly racist, classist, homophobic, transphobic, etc., while others expressed concerns that such regulation has the potential to harm freedom of expression.

To be more specific, some organisations believe certain types of content that are discriminatory in nature should be swiftly removed from social media platforms. Plurales states:

> *Clearly, there are some words associated with' hate speech' and it is important to be reasonable with what is being said. So attending to common sense, if it is obvious that one content is deepening some inequality related to gender, class, disability, ethnic, and racial issues, I think it is important to remove that content immediately.*

In contrast, a representative from CELE said the definitions of what constitutes each type of speech that may be subject to content moderation are typically vague, do not provide for exceptions when it comes to protected speech such as political or public interest speech, and can lead to undue control over online speech. In addition to affecting freedom of expression, content moderation practices seriously affect public debate.

Luisa Isaza, a Colombian expert on freedom of expression and a researcher studying at Oxford University, UK, highlighted the violation of freedom of expression that arises when content that denounces outrageous human rights violations, such as false positive campaigns, is removed, despite not having violated community standards. Further research analysing cases of content that was moderated without proper justification can provide new entry points to the debate. A researcher from Externado University, Colombia, validated this statement by saying that ‘there are many doubtful or borderline cases whose removal is not properly justified and may slip through. In these cases, it is clear that the rules and procedures for the removal of content can unduly interfere with freedom of expression.’

All consulted stakeholders recognised that racism, xenophobia, or transphobia can be amplified through social media. They also recognised how content moderation impacts freedom of expression and the problems associated with delegating the moderation of content to private companies.

A more in-depth knowledge of content moderation practices may produce new insights into societal issues and enable reflections on where the boundaries of free speech should lie. A representative from the feminist organisation Artemisas explained this by saying that ‘Issues like racism require a pedagogical exercise, especially when a tweet is deleted, people should know that a tweet has been deleted because it was racist or homophobic for instance.’ Taking a different approach, Wiwas Colectivo, an Afro-Colombian rights group, mentioned that their strategy is not to report racist comments but instead to ignore them: ‘We leave the hate comments without giving a response, we do not report them. It is our way to evidence racism.’

While there is consensus that content moderation can affect freedom of expression, there is currently no consensus among different stakeholders on the specific types of expression on social media that should be moderated. There is also no consensus on the best measures to address such content beyond content moderation, without jeopardising the rights to freedom of expression, privacy, and political participation. Even within the various organisations that make up civil society and the various bodies that monitor international human rights commitments, there are divergent positions on the content that should be allowed to circulate on social networks and the measures that should be adopted by states and platforms.

Organisations interviewed for this research are thus varied and reflect a broad spectrum of perspectives about the moderation and curation of content. Their interests are very diverse and their stakes for change are different; some of them may even be contradictory. However, there is a consensus that it is necessary to first understand in detail and discuss how the moderation and curation of content takes place, how automation in moderation works, what are the consequences for users and for free speech, and who is part of the process.

In the Colombian context, it would be beneficial for a coalition to initially focus on understanding the processes around the role of platforms and the state in the public debate and the risks associated with them. The functioning of content moderation processes, the impact and consequences of the current practice developed by companies, as well as the role played by the Colombian government are also among the topics to be worked on by a coalition.

Shared understanding of the role of the state and social media platforms

Any coalition or network on content moderation and freedom of expression in Colombia should not only build internal relations across members but also develop a dialogue with two key stakeholders: social media platforms and the state.

Interviewed stakeholders experience different levels of engagement with social media platforms. A representative from FLIP noted that direct channels of communication with platforms such as Facebook, through the Trusted Partner Programme, and X seem to depend on political will and staff engagement on a specific issue. FLIP also noted that while channels of communication with platforms were more efficient a few years ago, it is now harder to receive a reply from platforms to some of the reported content. There is an appetite from civil society to achieve a true conversation with platforms and expand the discussion from specific pieces of content to policies and opportunities for structural changes.

Observacom pointed out that the relationship among platforms, users, and researchers is asymmetric: 'There is still a lot of information relevant to users and researchers that is hidden from public scrutiny.' This suggests that platforms make unilateral decisions, without understanding the context, sometimes without notifications or responding to appeals.

If academics can mediate debates among users, as they are perceived as independent actors in the field, they can also develop research on the impact of content moderation and curation practices through accessing data from platforms.[38] Their capacity to do so, however, has been hindered by platforms limiting access to APIs (the tools that provide access to the actual platform data and, therefore, to what happens in that digital public space for researchers). Not only have platforms limited the types of researcher who can apply to access APIs, but the trend is to include a paywall, restricting access further for researchers.

There is a clear need for increased direct engagement with the platforms, both to be able to advocate for content moderation and curation practices that are better tailored to the Colombian context and to advocate for and gain more transparency on how these practices operate and influence online public debate in Colombia.

The state can significantly influence the way content is moderated or curated on social media platforms. Under current legal powers in Colombia, these functions are diffused across different authorities with no internal clarity about the relationship with the platforms. MINTIC is aware of the relationship between online content regulation and freedom of expression. The problem is that although MINTIC is the government entity

[38] Interview with Plurales.

responsible for internet governance issues, it did not provide comprehensive insights when interviewed on the implementation of current legal powers and how they impact the digital ecosystem. This is particularly relevant if local regulations are intended to target global platforms.

This research highlighted that the Colombian entities request content and account restrictions on legal grounds and through the community rules. It is reasonable to conclude that officials do not measure the human rights impact of these decisions. These are not public and tend not to be reviewed by other authorities.

Interested stakeholders should ideally develop strategies for engagement with the state to advocate for rights-based decision-making, encouraging the participation of stakeholders who are distant from dialogue with decision-makers. Any coalition dealing with content moderation in Colombia should balance different levels of knowledge, participation, relevance, and interaction with key actors. The coalition should also promote spaces for the exchange of information and knowledge.

Needs, gaps, and existing strengths for a prospective coalition

Latin American civil society has developed a number of coalitions and partnerships, and joined projects and productions that act upon or are related to the theme of content moderation and platform regulation. Besides the work developed by IFEX-ALC, Al Sur, Voces del Sur, Alianza Regional por la Libre Expresión y Acceso a la Información, La Alianza por el Cifrado en América Latina y el Caribe (AC-LAC), and the Brazilian Coalition for Network Rights, there has been a lot of cooperation within countries and in the region aiming to achieve some form of regulation of social media platforms.

This research shows that coalitions that already exist on topics relevant to content moderation (i.e. digital rights, terrorism, disinformation, accountability, or gender violence) are more suitable than creating a new coalition. Existing coalitions have already developed a common understanding of content moderation and specific calls for platforms and state institutions.

Multistakeholder coalitions with narrower common advocacy purposes are a good model for Colombia. Currently, many societal sectors have an interest in content moderation, and the model of engaging existing coalitions is perceived to be more successful in the local context.

While the idea of a coalition that focuses solely on content moderation and freedom of expression is commendable, the researchers recommend initially mapping and engaging existing content moderation coalitions to develop an alternative model of an 'extended network'. Engaging existing coalitions would require:

1. Identifying concrete topics for advocacy that connect the work of the organisations with content moderation. This can take place by focusing on a discussion on a bill, an executive order, a judicial decision, or a particular campaign. The bill on fighting gender-based violence in politics is an example (see Case study 3).

2. Building knowledge and awareness among prospective members of the extended network on content moderation and curation and determining the impact of social media platforms' business models. Research and capacity building must be stimulated to understand the intricacies of the social media landscape and to clarify how content curation practices or content moderation practices, beyond content removal or account blocking, can affect freedom of expression online and the public debate in Colombia.

3. Developing common demands that resonate with existing groups for advocacy actions towards the state and the platforms.

In order to expand the membership to a variety of different voices, coalitions can start agreements (such as partnership agreements) to collaborate with other organisations or networks working on topics that should be included in the content moderation debate, but that currently do not specifically work on content moderation issues.

This format ensures sustainability, given that existing coalitions have already developed trust to work together. It is also structured, with strategic plans and a commitment to dedicate time to the coalition.

After mapping and engaging coalitions, the researchers then recommend designing an agreement for this expanded network, as well as developing steps towards joint work on common grounds.

The proposed strategy is therefore two-fold. First, existing coalitions are in a better position to continue working on content moderation as they already have agreements and decision-making processes in place. They have also reached a certain level of common understanding and trust among members and a maturity to be able to articulate needs and gaps to expand the knowledge and reach of their advocacy. Second, they have enough leverage to create an expanded network that includes organisations not currently working on content moderation, such as those working on the rights of Afro-descent communities, indigenous people, or vulnerable groups (such as LGBTQI+ or children's rights). This could mitigate the risk that the work is too focused on one specific topic.

Case study 3: The Observatory of Violence against Women in Politics

Civil society is made up of organisations with different interests, agendas, and experiences, each of them with its own advocacy initiatives and priorities. While this is a synonym for richness of perspectives, the diversity of points of view means that certain topics can produce contradictory positions within a coalition. The broader the coalition, the more contradictory those positions can be. This situation should be considered in the development of any initiative aiming at tackling existing shortcomings with content moderation and curation in Colombia.

Diversity is not necessarily an impediment to coalitions, and this is the strength of multistakeholder initiatives. Indeed, diverse entities have organised themselves in coalitions for advocacy purposes, with different degrees of success. The key to sustainability lies in sharing a common understanding and similar goals, building trust, developing structure, and ensuring the availability of resources.

The experience of a coalition advocating the enactment of a bill to prevent and sanction violence against women in politics (including provisions on content moderation) provides a good case study. The Observatory of Violence against Women in Politics is a network of international and state actors working on the monitoring and analysis of violence against women in politics in Colombia.[39] The alliance is formed by state authorities and organisations working on violence in politics, elections, women's rights, digital rights, etc.

The Observatory was behind the successful advocacy campaign for the enactment in 2023 of a law that includes measures to counter online violence in politics. The text regulates the topic comprehensively, and includes sanctions, prevention, and capacity-building duties for public and private entities. This represents the first successful regulation of content moderation in Colombia after several attempts.

The Observatory first agreed on the understanding of the core issue it aimed to tackle: political violence against women. However, the Observatory's members had different views on how to address online violence. Some members had strong freedom of expression concerns and objected to provisions that gave legal power to authorities to request takedowns of content on social media without clear limits on the type of content that could be removed and the legitimacy of it. Organisations working directly with women victims of violence, on the other hand, called for stricter and urgent restrictions on the content, which can remain on social networks, due to the effects on real life. While this produced challenging discussions, and at times revealed friction between allies, focusing the debate on a specific issue helped members reach a compromise and intermediate solutions.

An expanded network working on issues not directly related to content moderation

Some important stakeholders in the debate on content moderation in Colombia are not yet working on content moderation and curation, for example organisations working on the rights of Afro-descent communities, indigenous people, or vulnerable groups (such as

[39] It was created by Conpes Document 4080, which contains the 'Public policy on gender equity for women: towards the sustainable development of the country'.

LGBTQI+ or children's rights). However, there are multiple coalitions focused on these issues in Colombia. As the research highlights, these communities are the target of structural violence in digital environments, especially on social media, and they are an essential part of the discussion on content moderation.

Interviewed representatives from organisations involved in these issues showed an interest in content moderation, even if their level of understanding was substantially lower compared to organisations already working directly on platform regulation and digital rights. They also shared similar concerns on the need for more transparency from platforms and on content moderation practices.

However, their likelihood of experiencing content moderation issues was less clear, leading to the risk of these organisations failing to follow, dedicate resources to, or contribute to an expanded network on content moderation. It is essential to include these voices and strengthen their advocacy through capacity building on questions of freedom of expression and how this applies in the online space, specifically on content moderation practices, the functioning of the systems developed by social media companies, and the impact of their business model. Increasing content moderation skills within these organisations will allow them to identify their interest in an agenda on content moderation regulation. It will also help them establish common interests with those already working on similar agendas.

Establishing an expanded network that can evolve in different phases and grow as membership increases is the best way to success. An expanded network will allow collaboration with organisations or coalitions by developing partnership agreements with clear goals, responsibilities, and activities to be carried out, such as capacity building, research, advocacy, campaigning, or raising awareness depending on common interests.

Analysis of stakeholders

This section lists existing multistakeholder coalitions that can be engaged on the topic of content moderation in Colombia.

El Índice de Derechos Digitales

This is a multistakeholder coalition composed of organisations with diverse backgrounds (academia, law and technology, journalism, or data analysis): El Veinte, FLIP, Fundación Karisma, ISUR, Linterna Verdem, and Dejusticia. One of the topics that this coalition examines is content control, including the analysis of content moderation by platforms and state requests during the Covid-19 pandemic.

Pros:

- has a multistakeholder approach;
- focuses on digital rights and has a mission tied to freedom of expression;
- holds open channels with social media companies and the state on these issues;
- has a small membership, which makes it easy to make decisions to start processes; and
- has a mission that can accommodate content moderation.

Possible topics of interest:

- content moderation and freedom of expression; and
- algorithmic transparency, platforms' transparency, and information sharing.

Red de Acción Cívica contra la Desinformación (ACD)

Managed by CIVIX, this network consists of Colombian organisations interested in 'disinformation', including public entities. 'Disinformation' is a topic with obvious links to content moderation. ACD is made up of media including La Silla Vacía, ColombiaCheck, El Mundo, Prensa Escuela, and El Universal; NGOs including Hablemos, Dividendo por Colombia, and Fundación Carvajal; and academia through Universidad Nacional de Colombia and several Education Municipal Secretariats.

Pros:

- has a multistakeholder approach;
- interested in 'disinformation' which is related to content moderation and freedom of expression; and
- has a large membership and a variety of stakeholders, which allows for multiple perspectives and broadly supported positions where consensus is reached.

Possible topics of interest:

- role of platforms in the information ecosystem; and
- algorithmic transparency, platforms' transparency, and information sharing.

Observatorio de Violencia contra las Mujeres en Política

The Observatory is a very diverse alliance in which there are civil society organisations working on digital rights, electoral rights, and women's rights, together with state authorities and international organisations. The Observatory consists of the Ministry of Justice Colombia, the Office of the President's Adviser for Women's Equity, UN Women, Transparencia por Colombia, the Secretariat for Women (Office of the Mayor of Bogotá), the Netherlands Institute for Multiparty Democracy, the National Democratic Institute, the National Electoral Council, and Fundación Karisma.

The bill on violence against women in politics was passed by Congress and the text is currently sitting in the Constitutional Court for review (expected completion date is early 2024). The work of the Observatory offers a clear opportunity to develop a regulation on content moderation in Colombia.

Pros:

- is the broadest coalition in Colombia and includes a variety of stakeholders such as state authorities and the private sector;
- was successful in delivering the first regulation in Colombia to include content moderation on social media platforms; and
- working on the topic of content moderation in upcoming months.

Possible topics of interest:

- enforcement challenges, good practices, and concerns (case studies, especially Mexico); and
- algorithmic transparency, platforms' transparency, and information sharing.

Alianza por la igualdad de las mujeres en los medios

This alliance includes journalists, academics, media outlets, and other civil society organisations interested in women's equality in the media. It also has strong ties with entities interested in gender violence. The alliance consists of the Red de periodistas con visión de género, FLIP, Sentiido, Colnodo, Fundación Karisma, Consejo de Redacción, and Línea del Medio.

Because the Constitutional Court asked the Congress to fill a legal void, the debating of a bill on the topic of gender equality in media will be part of the legislative agenda in the upcoming months. This draft law will consider the provisions of the bill on violence against women in politics and will probably include moderation of content on social media platforms.

Supporting this alliance offers another clear opportunity to develop a rights-based regulation on content moderation in Colombia – if this is considered beneficial by stakeholders.

Pros:

- has information on gender violence on social media, and is ready to provide this evidence and support the drafting of a law for women journalists;
- the Court's call for the Congress to regulate is a strong support for their advocacy;
- working in the upcoming months on the regulation of content moderation following their interest in digital violence against women in journalism; and
- has a small membership.

Possible topics of interest:

- enforcement challenges, good practices, and concerns; and
- algorithmic transparency, platforms' transparency, and information sharing.

Conclusion

During the course of the research, relevant stakeholders and issues at the intersection between online content moderation and freedom of expression in Colombia were mapped out. Stakeholders were interviewed to gather a realistic and full picture of how social media moderate and curate content in Colombia and the impact this has on freedom of expression in the country. The report also analysed the feasibility of establishing a local coalition on content moderation and freedom of expression. It concludes that the best strategy involves engaging existing coalitions working on issues related to content moderation and supporting them to expand their objectives to cover such issues. The aim of the coalition could be to ensure that content moderation in Colombia is informed by international freedom of expression standards and by the local context.

The background of the civil society organisations who participated in this study is very broad, their interests diverse, and their stakes for change different; some of them may even be contradictory. This results in a broad spectrum of perspectives about the moderation and curation of content. However, all interviewed stakeholders underlined the importance of reaching a common understanding of the functioning of content moderation, how automation works, what are the consequences for users and for freedom of expression, and who holds responsibility for these processes.

Currently, civil society, academia, and media do not agree on a number of issues. These include a common approach to ‘harmful’ content protected by freedom of expression, even among themselves; the measures to address it in the best way, without jeopardising the rights to freedom of expression, privacy, and political participation; and the measures that should be adopted by states and platforms regarding content moderation. The understanding of what may constitute ‘harmful’ content – inherent in the open nature of the meaning of the term ‘harmful’ – may also be subject to change due to the social differences between the stakeholders and users of the platforms in terms of religion, political beliefs, and specific context.

Nevertheless, there is a strong common position that Colombian civil society has a deeply rooted culture of freedom of expression, and all interviewed stakeholders recognised that content moderation may pose a risk to freedom of expression that should be addressed.

It was noted that civil society's work is often hindered by content moderation or curation practices, for example when denouncing abuses by state authorities. In these cases, stakeholders blame the platforms or the state and consider those practices as censorship tools. Stakeholders do not understand the different content policies and enforcement processes that are involved. The case studies and testimonies show how content moderation can lead to the silencing of voices, which can lead to a 'digital black hole'.

The interviewed stakeholders all shared the same complaint over the opacity of platforms' community rules and their content moderation process and decision-making. The stakeholders demand more transparency on both decision-making and processes, and that platforms provide more resources to inform content moderation practices through local context.

Another emerging common concern is curation practices. These are perceived as particularly unclear and distressing, and interviewees feel their effects as censorship. More research and capacity-building needs to be done to understand if the new forms of content curation produce this effect or sensation, or if it is confused with content moderation practices.

Considering that social media are key players in the information ecosystem, and their impact and influence, this report has shown that there is still a lack of contextual knowledge among stakeholders about the functioning and impact of moderation and content curation in Colombia. Even if platforms at times try to provide information and explanations, it is not enough to effectively explain how they account for local specificities and how their processes may impact the reality of online speech in Colombia. For example, while social media companies indicate that they comply with the laws of the countries in which they operate, the extent to which national laws are enforced and/or impact their decisions is not clear; this void prevents stakeholders from having more informed regulatory positions.

From the written response received by MINTIC, it is clear that the Ministry is aware of the relationship between online content regulation and freedom of expression. Although MINTIC is the government entity responsible for internet governance issues, their written position shows that the lack of internal analysis on the current regulation is a missed opportunity to gather lessons learned and information on the state's role if the legal powers are to be expanded via local regulations covering platforms. It is possible to conclude that officials do not measure the human rights impact of these decisions – that there is a big void as decisions are neither public nor subject to subsequent review by another authority.

The research has also shown that content moderation and curation have an important role in the peacebuilding process in Colombia. For example, content that may contain violent material may be in breach of community standards of different platforms, but may be important for the public debate, for its ability to constitute evidence of state abuse or its role in memory building. Numerous calls have been raised to keep these types of content visible in the public digital space.

In order to establish a mechanism that would deal with these issues, the research concludes that a broad multistakeholder coalition composed of organisations from different backgrounds and uneven skills would not work in the Colombian context. Instead, it proposes to build on existing coalitions that are already working on content moderation related issues, expanding their goals and connecting them to the current regulatory processes. This structure can be defined as an 'expanded network', led by an existing coalition with established structure and governance and joined by other organisations or coalitions working on issues not directly related to content moderation, but that represent important stakeholders in the topic (for example, organisations working on the rights of minority or indigenous groups, women, LGBTQI+ people, etc.).

Recommendations

The following recommendations aim to provide some guidance on the next steps towards the facilitation of a network that would focus on content moderation and freedom of expression in Colombia.

Content moderation discussions

Discussions on content moderation among certain stakeholders are complex. These can be aligned to existing regulatory opportunities where stakeholders feel strongly impacted by content moderation. Understanding specific blind spots affecting freedom of expression online (especially in relation to content moderation) as part of larger regulatory discussions can be a driver to engage existing coalitions in the work foreseen by the **Social Media 4 Peace** project in Colombia.

To leverage the upcoming regulatory discussions in Colombia, the 'expanded network' could focus its work and strategies on:

1. advocating for increased transparency from social media platforms on content moderation practices;
2. conducting research on the impact of new forms of moderation and curation on freedom of expression; and
3. conducting research on the identified phenomenon of 'digital black holes'.

Common goal

Experiences of successful coalitions in Colombia have shown that despite – or perhaps because of – the diversity of their positions, their strength lies in focusing on a specific advocacy goal or on a specific topic, such as a specific legislative process.

To be successful, the ‘expanded network’ should design a common goal to align the positions of its members. This could include specific requests from key actors, such as the platforms or the state, over content moderation in key advocacy streams.

Capacity building and knowledge

Key stakeholders in Colombia agree that social media platforms do not provide sufficient information about their content moderation and curation processes. There are also complaints that the community standards are often vague. Capitalising on these common complaints and concerns is key to ensuring the involvement of organisations or coalitions that are currently not directly focusing their work on content moderation.

To this end, the capacity and knowledge of prospective stakeholders of the proposed ‘expanded network’ on content moderation practices and their impact on freedom of expression should be strengthened. Capacity building and knowledge sharing should also focus on the transparency obligations of social media platforms and a better understanding of the key shortcomings of the current transparency reports. This would empower stakeholders with more arguments to demand transparency over content moderation on social media platforms and to ask for more active involvement of the state.

Collaboration

The ‘expanded network’ should focus on creating avenues for collaboration with social media platforms to engage in a sustainable dialogue that contributes to addressing flaws in content moderation and curation and the protection of fundamental rights. Interaction with platforms should go beyond the resolution of specific cases to address platforms’ structures and processes, the impact of their business model, and their local operation in Colombia.

Awareness of the relevance and functioning of content moderation and content curation among the general population is very low. While the state holds a responsibility to provide the skills and the knowledge to citizens in terms of increased media, digital, and information literacy, platforms should ensure that their users have at least a clear understanding of community guidelines. The ‘extended network’ of stakeholders could play an active role in working towards these objectives and could advocate for – and be involved in – a digital literacy programme conducted by the state or platforms.

Research

Civil society and academia need to access platforms’ data through free APIs to produce research that goes beyond case studies and includes key aspects of content moderation. The prospective ‘expanded network’ could embrace such a goal in its advocacy with social media platforms.

Annex A: Risk analysis

The **Coalition for Freedom of Expression Online and Content Moderation** emerges as a unique opportunity for participation and contribution by all the actors and as a mechanism for meaningful change. The coalition offers a path to consensus on key content moderation issues – and opportunities to address them. The following table provides an overview of the potential risks related to the formation and functionality of the coalition, identified by the respondents, including potential ways to overcome and mitigate them.

Risk type*	Description of risk	Likelihood**	Impact***	Monitoring and mitigation
Institutional	Organisations' time and human resources to dedicate to the coalition's work on content moderation.	Likely	Minor	• While organisations have scarce resources to dedicate to new endeavours or initiatives, content moderation is part of the agenda of some. • To be successful, any initiative has to capitalise on existing organisational work and build upon it.
Institutional	Reaching agreement among coalition members and creating trust.	Likely	Major	• If capitalising on existing coalitions, agreements, decision-making processes, and trust are already in place. Integrating any new member must follow a process of ensuring that the terms of the coalition are known and agreeable to all members. • Agreements on advocacy calls and positions on content moderation need to be established. Capacity building and network consolidation activities must be envisaged while an expanded network is put in place.

Financial	Sustainability and longevity of the coalition depend on the availability of funds.	Likely	Major	• Sustainable funding for the coordination of the coalition and for the establishment of its activities. • Joint funding application and participation of coalition members in donor meetings and the agenda-setting process.
Political	Effective participation in policy decisions on content moderation.	Unlikely	Minor	• The coalition can effectively assist and be consulted in regard to public policy initiatives, thus advocating for the values and objectives that potential future regulation in this field needs to ensure.
Institutional	Agreements between organisations for participation in the coalition.	Likely	Major	• The nature of organisations is diverse; there might be opposite perspectives regarding certain topics. The meetings need to have methodologies that allow for different opinions while guiding the conversation.

Notes:

* The risk type is pre-classified in the following categories: Political, Safeguarding, Stakeholder, Finance, Compliance, Reputation, Other, and Covid-19.

** The risk likelihood is presented on the scale: Unlikely, Possible, Likely, and Almost certain.

*** The risk impact is presented on the scale: Minor, Moderate, Major, and Severe.

Annex B: Interview sheet

The researchers held interviews with representatives from the following organisations:

Organisation	Name	Category	Theme
–	Luisa Isaza	Researcher from Oxford University	Specialises in freedom of expression online
Artemisas	Juliana Herrera	Civil society	Women's rights
Caracol TV	Anonymous	Media and communications	Colombia news channel
Centro de Estudios en Libertad de Expresión y Acceso a la Información (CELE)	Agustina Del Campo	Academia/think tank	Research on freedom of expression and information access
Centro Plurales UR	Flora Rodriguez	Academia/think tank	Center for Diversity, Equity and Inclusion from Universidad del Rosario
Colectivo Wiwas	Cindy Pérez	Afro collective	Transmission of Afro cultural knowledge of the Wiwas community
ColombiaCheck	Ana Saavedra	Media and communications	Fact-checking online content
Échele Cabeza	Julián Quintero	Civil society	Reduction of risks and damages of psychoactive substances
El Veinte	Ana Bejarano	Civil society	Freedom of expression and strategic litigation

Externado University researcher	Anonymous	Academia	Requested anonymity
Facebook	Group	Digital platform	Platform
Federación Colombiana de Periodistas (FECOLPER)	Jorge Velásquez	Civil society	Colombian network of journalists
FLIP	Jonathan Bock	Civil society	Freedom of expression and freedom of the press
Fundación Gabo	Ricardo Corredor	Civil society	Journalism
Fundación interpreta	–	Civil society	Research on complex social problems
Fundación Santamaria	Kika Ruiz	Civil society	Trans rights
Google	Group	Digital platform	Platform
Indepaz	Juana Cabezas	Civil society	Peacebuilding in Colombia
La Liga Contra el Silencio (League Against Silence)	Alejandro Gómez	Media network	Censored stories in Colombia
La Silla Vacía	Daniel Pacheco	Media and communications	News, stories, and debates about power in Colombia
Linterna Verde	Carlos Cortes	Civil society	Public opinion in digital spaces
Ministerio de las Tecnologías de la Información y las Comunicaciones	Written document signed by Aylin Torregroza Villarreal	Institution	In charge of information and communication technologies in Colombia

	(Viceministerio de Transformación Digital)		
Observacom	Gustavo Gómez	Think tank	Regulation and public policies related to the media, telecommunications, the internet, and freedom of expression
ONIC	Wilson Herrera	Indigenous network	National indigenous organisation
RedPapaz	Carolina Piñeros	Non-profit corporation	Child protection in digital spheres
Sentiido	Lina Cuellar	Digital media	Gender, diversity, and social change
Temblores ONG	Alejandro Lanz	Civil society	Social transformation
Wikimedia Colombia	Mónica Bonilla	Civil society	Education and knowledge access through digital tools
X	Group	Digital platform	Platform

Annex C: Content policies of main platforms

Meta	X	YouTube	TikTok
Violence and criminal behaviour • Violence and incitement • Dangerous individuals and organisations • Coordinating harm and promoting crime • Restricted goods and services • Fraud and deception	**Safety** • Violent speech • Violent and hateful entities • Child sexual exploitation • Abuse/harassment • Hateful conduct • Perpetrators of violent attacks • Suicide • Sensitive media • Illegal or certain regulated goods or services	**Spam and deceptive practices** • Spam, deceptive practices, and scams policies • Impersonation policy • External links policy • Fake engagement policy • Playlists policy • Additional policies	**Safety and civility** • Violent behaviours and criminal activities • Hate speech and hateful behaviours • Violent and hateful organisations and individuals • Youth exploitation and abuse • Sexual exploitation and gender-based violence • Human exploitation • Harassment and bullying
Safety • Suicide and self-injury • Child sexual exploitation, abuse, and nudity • Adult sexual exploitation • Bullying and harassment • Human exploitation • Privacy violations	**Privacy** • Private information • Non-consensual nudity • Account compromise	**Sensitive content** • Nudity and sexual content policies • Thumbnails policy • Child safety policy • Suicide, self-harm, and eating disorders policy • Vulgar language policy	**Mental and behavioural health** • Suicide and self-harm • Disordered eating and body image • Dangerous activities and challenges
Objectionable content • Hate speech • Violent and graphic content	**Authenticity** • Platform manipulation and spam • Civic integrity	**Violent or dangerous content** • Harmful or dangerous content policies	**Sensitive and mature themes** • Sexual activity and services

• Adult nudity and sexual activity • Sexual solicitation	• Misleading and deceptive identities • Synthetic and manipulated media	• Violent or graphic content policies • Violent criminal organisations policy • Hate speech policy • Harassment and cyberbullying policies	• Nudity and body exposure • Sexually suggestive content • Shocking and graphic content • Animal abuse
Integrity and authenticity • Account integrity and authentic identity • Spam • Cybersecurity • Inauthentic behaviour • Misinformation • Memorialisation		**Regulated goods** • Sale of illegal or regulated goods or services policies • Firearms policy	**Integrity and authenticity** • Misinformation • Civic and election integrity • Synthetic and manipulated media • Fake engagement • Unoriginal content and QR codes • Spam and deceptive account behaviours
Respecting intellectual property • Intellectual property		**Misinformation** • Misinformation policies • Election misinformation policies • Covid-19 medical misinformation policies • Vaccine misinformation policy	**Regulated goods and commercial activities** • Gambling • Alcohol, tobacco, and drugs • Firearms and dangerous weapons • Trade of regulated goods and services • Commercial disclosures and paid promotion • Frauds and scams

Bibliography

Access Now (2021) What You Need to Know about the Facebook Papers.

Adam, M. (2021, 7 May) [@mosseri]. 'Yesterday We Experienced a Technical Bug' [X post].

ARTICLE 19 (2018) Side-stepping Rights: Regulating Speech by Contract, Policy Brief.

ARTICLE 19 (2021) Social Media Councils: One Piece in the Puzzle of Content Moderation.

ARTICLE 19 (2021) Watching the Watchmen: Content Moderation, Governance, and Freedom Of Expression.

ARTICLE 19 (2022) Content Moderation and Local Stakeholders in Bosnia and Herzegovina.

ARTICLE 19 (n.d.) #MissingVoices.

Asamblea Nacional Constituyente (1991) Constitución Política de Colombia.

Banco Mundial (2020) Población en Colombia.

Botero, C. (2021) Public Memory and the Digital Black Hole.

Botero, C. (2021) Represión en la calle, sensación de censura en redes.

CELE (2022) Penar la intolerancia 'male sal'. Críticas a la Convención Interamericana contra toda forma de Discriminación e Intolerancia.

Cepeda, M.J. (1995) El derecho a la constitución en Colombia. entre la rebelión pacífica y la esperanza.

Charry, C. (2021) Los en vivo: Estar vivos y ser vistos, Punto y Coma.

Cifras y Conceptos (2023) Sexto estudio de percepción de jóvenes.

Cinep (2008) Comunicación y conflicto armado: El fin no justifica a los medios.

Comisión de la Verdad (2022) Hallazgos y Recomendaciones: Hallazgos y recomendaciones de la Comisión de la Verdad de Colombia.

Comisión de la Verdad (2022) Hay un futuro si hay verdad.

Comisión de la Verdad (2022) No matarás: Relato histórico del conflicto armado interno en Colombia.

Comisión de Regulación de las Comunicaciones (2022) Data Flash 2022-026: Internet Fijo.

Comisión Interamericana de Derechos Humanos (2021) Observaciones y recomendaciones: Visita de trabajo a Colombia.

Comisión Internacional de Juristas (2023) Colombia: Defensores de derechos humanos continuaron bajo presión y ataques.

Committee on the Elimination of Discrimination against Women (2020) General Recommendation No. 38 On Trafficking In Women And Girls In The Context Of Global Migration.

Congreso de la República (2008) Ley 1257 de 2008.

Facebook Oversight Board (2021) Colombia Protests.

Constitución Política de Colombia (1991) Art. 2, 3 de julio de 1991.

Corte Constitucional (2023) Estadísticas de tutelas radicadas en la Corte Constitucional.

Corte Constitucional de Colombia (2004) Sentencia T 1191 de 2004.

Corte Constitucional de Colombia (2005) Sentencia T-1062 de 2005.

Corte Constitucional de Colombia (2012) Sentencia T-627 de 2012.

Corte Constitucional de Colombia (2015) Sentencia T-066 de 2015.

Corte Constitucional de Colombia (2020) Sentencia T-031 de 2020.

Corte Constitucional de Colombia (2021) Sentencia T-146 de 2021.

Corte Constitucional de Colombia (2022) Sentencia T-280 de 2022.

Corte Constitucional de Colombia (2023) Sentencia T-087 de 2023.

Corte Interamericana de Derechos Humanos (2018) Caso Isaza Uribe vs Colombia.

DANE (2018) Grupos etnicos: Información técnica. Accessed 1 December 2023.

Dejusticia (2021) Homenaje a la tutela: El mecanismo que democratizó la Constitución de 1991.

Del Campo, A. (2022) Contenido legal pero dañino y poca previsión en la supervisión. CELE.

El Economista (2022) El Gobierno anunció un proyecto para regular las redes sociales para 'que dejen de intoxicar' a la democracia.

Fiorella, G. (2019) El segundo a segundo del disparo que mató a Dilan Cruz.

Fitzgerald, M., (2022) No es solo contra Francia: En política, los insultos son contra todas, Revista Cambio.

France24 (2022) Evolucionar: El último giro de la desinformación electoral en Colombia.

Freedom in the World Index (2023) Freedom in the World 2023: Colombia.

Frithjof, S.-M., Britta, H. and Melanie, V. (2012) How Stressful is Online Victimization? Effects of Victim's Personality and Properties of the Incident, *European Journal of Developmental Psychology*, 9, 2: 260–274.

Fundación Karisma (2014) Violencia Contra Las Mujeres Y Tic (Vcm Y Tic).

Fundación Karisma (2021) El proyecto de Ley 600 sigue su curso en el Congreso a pesar de las críticas.

Fundación Karisma (2021) Fallas de internet, bloqueos de redes y censura de contenidos en protestas: Realidades y retos para el ejercicio de los derechos humanos en los contextos digitales.

Fundación Karisma (2021) Periodistas sin acoso.

Fundación Karisma (2021) Pistolas contra celulares.

Fundación Karisma (2021) Violencias machistas atacan la libertad de expresión de periodistas y comunicadoras en Colombia.

Fundación Karisma (2022) Automatic Copyright Detection: A Tool For Inequality.

Fundación Karisma (2022) Dónde están mis datos.

Fundación Karisma, Fundación Para la Libertad de Prensa (FLIP), El Veinte e ISUR (2022) Comentarios al Proyecto de Ley 318.

Fundación Karisma, Fundación Para la Libertad de Prensa (FLIP), El Veinte e ISUR (2023) Comentarios PL (Medidas para prevenir, atender, rechazar y sancionar la violencia contra las mujeres).

Gago, E. (2022) La polarización como estrategia política.

Global Disinformation Index (2022) Disinformation Risk Assessment: The Online News Market in Colombia.

Google (2022) YouTube Community Guidelines Enforcement.

Google (2023) About the YouTube Trusted Flagger Programme.

Google (2023) Government Requests to Remove Content.

GSMA Intelligence (2022) Get Started Now with GSMA Intelligence.

Human Rights Watch (2020) 'Video Unavailable': Social Media Platforms Remove Evidence of War Crimes.

Indepaz (2019) Los discursos del odio y la estigmatización fatal.

Infobae (2021) Ministerio de Defensa confirmó la militarización en Cali: llegan 450 soldados.

Instagram (n.d.) Normas comunitarias.

Instagram (n.d.) Qué hacer si crees que Instagram no debería haber retirado tu publicación.

Inter-American Commission on Human Rights, Special Rapporteur for Freedom of Expression (2017) Standards for a free, open, and inclusive internet.

Inter-American Commission on Human Rights, Special Rapporteur for Freedom of Expression (2021) Disinformation and Freedom of Opinion And Expression.

Inter-American Commission on Human Rights, Special Rapporteur for Freedom of Expression (2022) Disinformation and Freedom of Opinion and Expression During Armed Conflicts.

Kari, P. (2019) Climate Misinformation on Facebook 'Increasing Substantially', Study Says, *The Guardian*.

Kepios (2022) We are Social, Digital 2022 Colombia.

Kouzy, R. (2020) Coronavirus Goes Viral: Quantifying the Covid-19 Misinformation Epidemic on Twitter.

La República (2017) Audio entrevista Juan Carlos Vélez Uribe.

Lesher, M,. Pawelec, H. and Desai, A. (2022) Disentangling Untruths Online: Creators, Spreaders and How to Stop Them, OECD Going Digital Toolkit Notes.

LinkedIn (2023) Government Requests Report.

McIntyre, N., Bradbury, R. and Perrigo, B. (2022) Behind Tiktok's Boom: A Legion Of Traumatised $10-a-Day Content Moderators, Bureau of Investigative Journalism.

Meta (2022) Como aplica Meta sus políticas.

Meta (2022) Meta Q4 2021 Quarterly Update on the Oversight Board.

Meta (2022) Restricciones de contenido en virtud de la legislación local.

Meta (2023) Contexto local en nuestras normas globales.

Meta (2023) Eliminar contenido infractor.

Meta (2023) Normas comunitarias de Facebook.

Meta (n.d.) Creo que Facebook no tendría que haber eliminado mi publicación.

Ministerio de Tecnologías de la Información y las Comunicaciones (2022) Boletín trimestral del sector TIC: Cifras tercer trimestre de 2022.

Ministerio de Tecnologías de la Información y las Comunicaciones (2022) Índice de Brecha Digital 2021.

Misión de Observación Electoral (MOE) (2018) Impacto de las redes sociales en el proceso electoral colombiano (Elecciones de Congreso y Presidencia 2018).

Misión de Observación Electoral (MOE) (2022) Los discursos de odio racistas y sexistas son legitimadores de la violencia.

Misión de Observación Electoral (MOE) (2022) Sexto informe preelectoral de violencia.

Movilizatorio (2021) Estudio sobre polarización de audiencias en Colombia.

Newton, C. (2021) The Tier List: How Facebook Decides Which Countries Need Protection.

Observacom (2020) Estándares para una regulación democrática de las grandes plataformas que garantice la libertad de expresión en línea y una Internet libre y abierta.

Observacom (2021) Declaración latinoamericana sobre transparencia de las plataformas de internet.

Observacom (2022) Moderación privada de contenidos en Internet y su impacto en el periodismo.

Observatorio de Violencias Políticas a las Mujeres (2022) En sus marcas: La carrera de las mujeres en la política.

Osorio, F.E. (2001) Entre la supervivencia y la resistencia. Acciones colectivas de población rural en medio del conflicto armado colombiano.

Paul, K. (2019) Climate misinformation on Facebook 'increasing substantially', study says, *The Guardian*.

Pérez, A. and Martinez, C. (2022) Moderación privada de contenidos en Internet y su impacto en el periodismo.

Pinterest (2022) Transparency Report.

Piper, E. (2021) Palestinians Bear the Brunt of Big Tech Moderation.

Privacy International (2018) Privacy and Freedom of Expression in the Age of Artificial Intelligence.

Ranking Digital Rights (2022) Methods and Standards.

Report of the United Nations High Commissioner for Human Rights (2017) Promotion, Protection and Enjoyment of Human Rights on the Internet: Ways to Bridge the Gender Digital Divide from a Human Rights Perspective.

Reuters Institute (2022) Digital News Report 2022 Colombia: Consumo y confianza de la información en entornos digitales.

Rotta, S. (2021) Qué pasó con las publicaciones en Instagram durante el paro nacional.

Semana (2016) Consejo de Estado dice que hubo ‘engaño generalizado’ en campaña del No en el Plebiscito.

Semana (2016) Consejo de Estado podría suspender resultados del Plebiscito.

Silva, S. (2017) Polarización en Colombia: Superar mitos y aceptar realidades, El Eafitense.

Snapchat (2022) Informe de transparencia.

Soyun, A., Baik, S. and Soy, C. (2022) Splintering and Centralizing Platform Governance: How Facebook Adapted Its Content Moderation Practices to the Political and Legal Contexts in the United States, Germany, and South Korea, *Information, Communication & Society*, 26, 14: 2843–2862.

Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression of the United Nations (2022) Joint Declaration on Freedom of Expression and Gender Justice.

TikTok (2022) Community Guidelines Enforcement Report.

TikTok (2023) Public Interest Exceptions.

TikTok (n.d.) Content Violations and Bans.

United Nations Development Programme (2021) Internet, libertad de expresión y acceso a la información en Uruguay Aportes para el debate sobre la gobernabilidad democrática en línea.

United Nations Educational, Scientific and Cultural Organization (2021) Addressing Hate Speech on Social Media: Contemporary Challenges.

United Nations Educational, Scientific and Cultural Organization (2021) Windhoek+30 Declaration: Information as a Public Good, World Press Freedom Day International Conference.

United Nations Educational, Scientific and Cultural Organization (2022) Finding the Funds for Journalism to Thrive: Policy Options to Support Media Viability.

United Nations Educational, Scientific and Cultural Organization (2022) How to Address Online #HateSpeech with a Human-Rights Based Approach?

United Nations Educational, Scientific and Cultural Organization (2022) The Rabat Plan of Action on the Prohibition of Incitement to Hatred [YouTube video].

United Nations Educational, Scientific and Cultural Organization (2022) Transparencia de la moderación privada de contenidos: Una mirada de las propuestas de sociedad civil y legisladores de América Latina.

United Nations Educational, Scientific and Cultural Organization (2023) Safeguarding Freedom of Expression and Access to Information: Guidelines for a Multistakeholder Approach in the Context of Regulating Digital Platforms.

United Nations Educational, Scientific and Cultural Organization (2023) Towards Guidelines for Regulating Digital Platforms for Information as a Public Good [YouTube video].

United Nations Educational, Scientific and Cultural Organization (n.d.) Countering Hate Speech.

United Nations General Assembly (2016) Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression.

United Nations General Assembly (2018) Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression.

United Nations General Assembly (2021) Disinformation and Freedom of Opinion and Expression.

UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2010) Inter-American Legal Framework Regarding the Right to Freedom of Expression.

Villena, D. (2021) Proyecto de ley pretende regular las redes sociales sin entender cómo funciona Internet.

WFB (n.d.) Country Comparison: Median Age.

X (2022) Colombia.

X (n.d.) About Country Withheld Content.

Yann, B. (2017) Claves del rechazo del plebiscito para la paz en Colombia.