The human rights impact of public and private measures to counter disinformation

Derechos Digitales' contribution to the upcoming Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression focused on disinformation

About Derechos Digitales

Derechos Digitales is an independent non-profit organization, founded in 2005, with main offices in Santiago de Chile. Our aim is the defense and promotion of fundamental rights in the digital environment in Latin America using advocacy tools among policymakers, private companies and the general public, to promote social change around the respect and dignity of all people.

On the issue of disinformation, Derechos Digitales has previously contributed in writing to the preparation of the Guide against deliberate misinformation in electoral contexts from the Organization of American States (OAS).[1] Derechos Digitales has also participated in strategic meetings of the group of experts that were entrusted to work for the discussion of that document, and more recently, we directly collaborated with the Special Rapporteur on Freedom of Expression of the OAS in the discussion of the spread of disinformation in the region in the pandemic context.

1. Information disorders as useful category

In our view, the phenomenon of information disorders is much broader than the circulation of disinformation, and includes other forms of alteration of the circulation of information that do not concern the veracity of an information, but rather the context or form in which it is placed, and that may not have a political character, but has a clear impact in the shape of democracy and the exercise of fundamental rights.

It is our opinion that this discussion needs to start by recognizing that disinformation is only a part of a broader and more complex phenomenon, that which academics such as Wardle and Derakhshan[2] have called «information disorders» and which contains three different types of information:

- misinformation: when false information is shared, but no harm is meant,
- disinformation: when false information is knowingly shared with the intention to cause harm, and
- malinformation: when true information is shared with the intention to cause harm, for instance, when information designed to stay private is moved into the public sphere.

[1] Available http://www.oas.org/en/iachr/expression/publications/Guia_Desinformacion_VF%20ENG.pdf
[2] Claire Wardle and Hossein Derakhshan, «Information Disorder :Toward an Interdisciplinary Framework for Research and Policy Making» (Council of Europe, 2017), https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c.

We find useful referring to «information disorders» in order to encompass a broader spectrum of behaviors that have as result the manipulation of public opinion and the harm in the exercise of fundamental rights. Among them, the use of technology (either by humans or by automated means) in a coordinated way to divert attention from certain information (which may be false or not). Rather than focus on the nature of the information, we have seen that these informational tactics are often used in our region to intervene in the visibility and dissemination of information, in order to capture and divert the public's attention, which impacts public debate not only by giving visibility to harmful information but also by hiding in plain sight other relevant informations for forming opinion. Influence on public debate can happen not only through false information, but also through these other informational tactics.

The term «fake news», as well as the term «disinformation», can contribute to the oversimplification of the problem if we do not discuss the different nuances and forms that information manipulation can take, and that cannot be addressed under the same parameters. False information that spreads by ignorance, or greed for capturing economic benefits coming from platforms' business models, or false information that is not entirely false but incomplete or decontextualized, are phenomena completely different from false information created intentionally with the purpose of deceiving.

For the above reasons, the strict dichotomy between false and true information seems problematic to us, insofar as it invites us to choose a focal point for evaluating information (in its veracity and quality), placing an excessive emphasis on the role of platforms and information verifiers. Again, it seems to us that the phenomenon of disinformation must recognize that there are ways of altering the circulation of information that do not end with its falsity. In particular, we find it problematic to place an excessive emphasis on the role of governments and public actors in qualifying information as true or false, since it has been repeatedly proven that the intervention of certain public actors not only can enhance the dissemination of false information, but also–as it has been found in our region–that certain public actors utilise premeditated articulations through armies of human trolls or automated mechanisms to influence the circulation of information, whether it is false or not. Furthermore, institutional violence against journalists is a form of disinformation exercised by public authorities in electoral contexts (and in others of political unrest or repression) as part of a strategy to maintain power.

At the same time, it is important to consider that the concept of «truth» is not always a useful measure of the value of information. Opinions, beliefs and other forms of expression cannot be evaluated under the lens of «truth/falsity» and are nevertheless affected by restrictions imposed by both governments and private companies under the umbrella of disinformation, even when they should be protected by the standards of free speech that are particularly strong in the Americas, according to Article 13 of the American Convention on Human Rights, that directly forbids prior censorship as a regulatory tool for controlling expression.

For all these reasons, and having into consideration this legal tradition of our region, as well as the concrete experiences that are shared in more detail later in this contribution, we believe that while there remains a need to protect spaces for the public expression of opinions, ideas, and information, not just strict facts, there might be something to be gained from efforts to «label» the origin of

information and «provide additional context» to extreme opinions in order to allow audiences to form their own criteria regarding the usefulness of a certain piece of content and its relationship with specific emisors. Some of these tools have been explored in a limited way by private actors, but improvements on their use can be achieved as well as regulatory incentives for their implementation in a more transparent and consistent manner across jurisdictions and implicated actors.

2. Public measures to counter information disorders: regulation and public policies regarding disinformation and misinformation in Latin America

As requested in the call for inputs, we share in this section some of the most salient examples of the measures that have been implemented by the States in our region.

2.1. Disinformation observatories

In 2020, Argentina created NODIO, an observatory of disinformation and symbolic violence in media and digital platforms, which has the stated aim of «protecting citizens from false, malicious and fallacious news».[3] Similar initiatives, such as the Social Observatory for Disinformation and Social Media Analysis in Europe, have been strongly community-led and with a wide participation of media, while NODIO depends from the Public Defender for Audiovisual Communication Services, a government official. For this reason, while NODIO states that its aim is to «study the strategies of fake news and identify their dissemination operations», its creation has been rejected by entities such as the Interamerican Press Association, which stated that «The observatories created to monitor and discuss freedom of expression issues have ended up being the first step with the obscure purpose of a government to regulate the media and meddle in content».[4]

In 2017, the Brazilian «Consultative Council on Internet and Elections» was created by the Electoral Superior Tribunal, and tasked with monitoring and potentially ordering the blocking of false news on social media ahead of the 2018 presidential elections[5]. The council has met with representatives of Facebook, Twitter, WhatsApp and Google in order to discuss their tools and procedures to block the dissemination of fake news. The Council includes members of the Electoral Superior Tribunal, civil society, the Brazilian Intelligence Agency, and the army, among others. It has been the target of criticism, as its meetings are held in secret, and no report has been released yet by the Council.[6]

2.2. Censorship and filtering

Latin-American governments, particularly those from Venezuela, Nicaragua and Ecuador, have a long track record of blocking online content or specific services. Several countries have or have had bills

[3] Defensoría del Público, «Llegó NODIO, el Observatorio de la desinformación y violencia simbólica,» Defensoría del Público de servicios audiovisuales, October 9, 2020, https://defensadelpublico.gob.ar/llego-nodio-el-observatorio-de-la-desinformacion-y-la-violencia-simbolica/.
[4] La Vanguardia, «La SIP rechaza creación de 'oscuro' observatorio de medios en Argentina,» La Vanguardia, October 13, 2020, https://www.lavanguardia.com/politica/20201013/484036637706/la-sip-rechaza-creacion-de-oscuro-observatorio-de-medios-en-argentina.html.
[5] Taisa Sganzerla, «Brasil aprova resoluções mais duras contra 'fake news' visando eleições de 2018,» *Global Voices em Português* (blog), January 18, 2018, https://pt.globalvoices.org/2018/01/18/brasil-aprova-resolucoes-mais-duras-contra-fake-news-visando-eleicoes-de-2018/.
[6] Alves, M. & M. Maciel, "O fenômeno das fake news: definição, combate e contexto". *Internet & Sociedade*, vol. 1 n. 1, 2020.

regarding the blocking of content for reasons such as «hate speech», «fake news» or «defamation», particularly in the context of elections. Some of the censorship measures taken by governments are then implemented by their own telecommunication bodies.

In 2018, Honduras attempted to regulate speech online providing very broad definitions of the kinds of illegal speech that Internet intermediaries should monitor.[7] In 2019, a similar Bill was proposed in Ecuador, including provisions that would make Internet intermediaries directly responsible for taking down speech deemed illegal.[8]

In Venezuela, the main ISP, CANTV, which is a State-owned telecom company, has repeatedly[9] blocked access to news sites without any kind of judicial or administrative process.[10]

Other measures are aimed at private companies: in August 2020, the Supreme Court of Brazil ordered Facebook to block several accounts that, according to the court, spread "hatred" and "fake news", and imposed the company a fine of USD $235,294 per each day in which the company refused or delayed compliance.[11]

2.3. Criminal penalties and civil sanctions

Criminal penalties and civil sanctions are the traditional ways that have been established in the region to deal with liability for harmful expressions. Given the American Convention on Human Rights forbids prior censorship, the way in which Latin American countries have traditionally regulated expression has been the use of ex-post measures like these, which have been declared disproportionate by the Interamerican Court of Human Rights. The issues of legality, proportionality and necessity of the criminal penalties and civil sanctions continue to be a problem in the region, especially now that there are administrative decrees or new laws that include them as a way to regulate expression in the digital environment.

One of the most salient examples can be found in Venezuela, where dozens of people have been arrested and imprisoned for communicating information deemed by the government as «false» or «dangerous», particularly workers from the health sector in the context of information related to the COVID-19 pandemic. However this behavior is far from new, as Osvaldo Alvarez Paz was arrested in 2010 for «spreading false information» in an interview in which he cited the existence of drug trafficking and terrorism within Venezuela and asked the government to act, claiming that not acting

[7] See the note from Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression; and the Special Rapporteur for Freedom of Expression of the Inter-American Commission on Human Rights "sobre ciberseguridad y medidas de protección ante los actos de odio y discriminación en internet y redes sociales" available at: https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL.HND.07.06.18.pdf

[8] See "Proyecto de ley Orgánica del Uso Responsable de Redes Sociales", available at: https://es.slideshare.net/fabriciovela1973/proyecto-de-ley-para-uso-responsable-de-redes-sociales

[9] Espacio Público. «Censura digital: CANTV bloquea 13 sitios en dos semanas · Espacio Público», 4 de marzo de 2019. http://espaciopublico.ong/censura-digital-cantv-bloquea-13-sitios-en-dos-semanas/.

[10] TalCual. «Cantv bloquea portal de noticias e investigación Runrunes». *TalCual*, 17 de mayo de 2020. https://talcualdigital.com/regimen-bloquea-portal-de-noticias-e-investigacion-runrunes/.

[11] Cordero, «Facebook bloquea cuentas de seguidores de Bolsonaro, tras presión de la Corte Suprema». https://www.france24.com/es/20200802-brasil-bolsonaro-facebook-twitter-justicia-noticias-falsas

was complicity, and Guillermo Zuloaga was arrested for «causing panic in the community through false information disseminated by the press.»[12]

In 2018, Venezuelan union leader Elio Palacios was detained and accused of spreading «false information» that could cause «panic and anxiety» in the population, after releasing an audio recording on social networks claiming that the state electrical company was short on personnel and equipment, and warning of the system's imminent collapse. One week after Palacios was detained, a massive blackout hit nine states in the western part of the country, yet Palacios remained imprisoned and subject to torture. In 2020, Jesús Enrique Torres Ascanio and Jesús Manuel Castillo Pereira were arrested and charged for causing distress in the population by presumably authoring and disseminating a video claiming that there were two positive cases of COVID-19 in the city of Los Teques. Journalist Darvinson Rojas was arrested for spreading misinformation after he added up the numbers of positive cases for coronavirus that different authorities had given and realized that they yielded a higher number than the government of Nicolás Maduro was reporting.[13]

The concepts of «public distress» and «conspiracy» are often used in Latin America along with or instead of the accusation of spreading misinformation. In Mexico, in 2011, two people were jailed by the government of Veracruz, accused of terrorism and sabotage for allegedly spreading false information about armed attacks and kidnappings that caused alarm in the population. Even though they were eventually released, the Veracruz state congress approved a legal amendment to the state Criminal Code to create the crime of «disturbance of the public order», which establishes a penalty of between one and four years in prison for anyone who, by any means, disseminates false information about explosives, armed attacks, or attacks with toxic substances that alarm or cause health damage to the population.[14] The Supreme Court of Justice of the Nation declared this rule unconstitutional based on the fact that its lack of precision resulted in disproportionate effects in relation to freedom of expression, given that it did not distinguish erroneous statements from fraudulent ones, sanctioning any means of broadcasting, including social networks and the Internet.

In Costa Rica, the crime of spreading misinformation exists only in the financial realm, where it is considered a felony to propagate or disseminate false news or facts capable of distorting or causing damage to the security and stability of the financial system, and it is punished with a penalty of three to six years in prison (Law N° 9048, article 236).[15]

Brazil, Nicaragua, Colombia, Chile, Paraguay and Peru all have or have had different bills considering misinformation, some of them in the electoral context and some motivated by the COVID-19

[12] Espacio Público, «Espacio Público repudia arrestos de Oswaldo Álvarez Paz y Guillermo Zuloaga · Espacio Público,» *Espacio Público* (blog), March 26, 2010, http://espaciopublico.ong/espacio-pco-repudia-arrestos-de-oswaldo-lvarez-paz-y-guillermo-zuloaga-2/.
[13] AFP, «Al menos 10 personas han sido detenidas en Venezuela por informar sobre covid-19,» Semana, May 7, 2020, https://www.semana.com/mundo/articulo/coronavirus-varios-detenidos-en-venezuela-por-informar-sobre-la-pandemia/669699/; Carlos D'Hoy, «Dos detenidos por difundir información falsa sobre el COVID-19», El Universal, March 14, 2020, http://www.eluniversal.com/sucesos/64379/dos-detenidos-por-difundir-informacion-falsa-sobre-el-covid-19; Guillermo D. Olmo, «"Dijeron que habían recibido una llamada por coronavirus y me llevaron preso": los periodistas y médicos detenidos en Venezuela en medio de la pandemia», *BBC News Mundo*, April 28, 2020, https://www.bbc.com/mundo/noticias-america-latina-52450803.
[14] Alberto Nájar, «México: liberan a tuiteros, pero tipifican nuevo delito,» BBC News Mundo, September 22, 2011, https://www.bbc.com/mundo/noticias/2011/09/110922_tuiteros_twitter_veracruz_libres_an.
[15] Asamblea Legislativa de la República de Costa Rica, «Reforma de La Sección VIII, Delitos Informáticos y Conexos, Del Título VII Del Código Penal,» Pub. L. No. N° 9048 (2012), http://www.pgrweb.go.cr/scij/Busqueda/Normativa/Normas/nrm_texto_completo.aspx?nValor1=1&nValor2=73583.

pandemic, but all of them attempting to include jail time and financial penalties. One of the several legal initiatives introduced in Chile in the past few years, bill No. 13605-07, presented in June of 2020, proposes sanctioning with prison (up to five years) and a fine (equivalent to around USD $14,000 as of February 2021) whoever «publishes, reproduces or disseminates on social networks or other media, false news intended to hinder the work of the authority in periods of health crisis».

2.4. Sanctions on intermediaries

There is a clear tendency in the region to seek the imposition of penalties on digital intermediaries for content deemed illegal under disinformation, hate speech and defamation laws. The previous mandate-holder of this Special Rapporteurship, in a joint statement with the IACHR Special Rapporteur, have indicated that:

> «objective or 'strict' liability, which holds the intermediary responsible for any content considered illegal on its platform, is incompatible with the American Convention [on Human Rights] because it is disproportionate and unnecessary in a democratic society. This type of liability promotes the monitoring and censorship of intermediaries towards their own users, which leads to larger amounts of content being censored in these platforms. This translates into a need for safe harbors, where intermediaries are safe from legal liability as long as they meet certain conditions. These conditions, however, cannot translate into a disproportionate obligation to monitor or control the user's activities. In this respect, the Joint Declaration on Freedom of Expression and «Fake News», Disinformation and Propaganda provides that «intermediaries should not be legally responsible in any case for third party content related to those services, unless they specifically intervene in such content or refuse to comply with an order issued in accordance with guarantees of due process by an independent, impartial and authorized supervisory body (such as a court) that orders the removal of such content, and have sufficient technical capacity to do so».[16]

With this in mind, we need to mention two concerning bills that are being discussed recently in the subject of misinformation in the region. In Brazil, the bill named «Internet Freedom, Responsibility, and Transparency Law», popularly known as the «fake news» bill, aims to prevent the dissemination of misinformation, with an emphasis on social media platforms, by holding telecommunication providers responsible for combating disinformation.[17] Among other things, this bill intends to place limits over the «maximum number of members» that a group on a private messaging platform can have (256 users) and to limit to how many users a message can be sent to at the same time (5 users, but only one during elections or during public emergency situations), and penalizes intermediaries with measures ranging from a warning to the prohibition of operating in the country.

[16] Declaration by the United Nations Special Rapporteur on Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe Representative on Freedom of the Media, the Organization of American States (OAS) Special Rapporteur on Freedom of Expression and the African Commission on Human and Peoples' Rights Special Rapporteur on Freedom of Expression and Access to Information., «Joint Declaration on Freedom of Expression and 'Fake News', Disinformation and Propaganda,» March 3, 2017, https://www.osce.org/fom/302796.

[17] Diogo Tulio dos Santos, «Brazil, Democracy, and the 'Fake News' Bill,» *Global Americans* (blog), January 4, 2021, https://theglobalamericans.org/2021/01/brazil-democracy-and-the-fake-news-bill/.

In Mexico, Senator Ricardo Monreal recently proposed a bill that aims for intermediaries who administer social networking platforms to request and obtain an authorization from the Federal Telecommunications Institute, and which includes fines for up to USD $4.5 million to companies that incur in violations regarding content takedowns that do not meet the proposed legislation's parameters.[18] The proposal includes as its subject of regulation a vague definition of «social networks» that encompasses practically any internet site or service that allows its users to disseminate information. Even services such as Wikipedia, news portals that allow comments, or messaging applications, will fall under this poorly defined category. For the purpose of establishing duties, the regulation creates a totally arbitrary threshold to identify «relevant social networks», these being any site or service with more than one million users or subscribers. There is no clarity if that reference considers the users in the Mexican territory or global presence of the service, or if this should account for registered or active users, or in what period. Relevant social networks are required to have policies to actively police content under ambiguous concepts such as «eliminate the spread of hateful messages», «prevent the spread of false news» and «protect personal data.» It is also worth noting that the Federal Telecommunications Institute (IFT), while being an autonomous body, is integrated by members proposed by the President, and would be given authority to restrict platforms where free expression is exercised.

2.5. Public policies on education and digital alphabetization

Media literacy is recognized to be one of the main strategies to be deployed to tackle misinformation and disinformation, having been proved beyond reasonable doubt that regulation alone is not enough. Countries such as Canada, Finland and Australia have incorporated digital literacy into their national educational curricula.[19]

We observe with concern the almost absolute absence of digital literacy, and specifically, of digital media literacy programs, in Latin American local public policies. Media literacy is a tool for citizen empowerment that provides them with tools for critical thinking before the vast array of information received in the current economy of attention, enabling them to differentiate the origins of a specific piece of information, evaluate its validity and its sources, sift what is important from what is not and assess for themselves what and who they can believe. Shifting away from an approach that gives the authority to someone (government, social platforms, media companies) to draw a stark line between what's «true» and what is «false», and instead, choosing to empower citizens to be able to not be deceived by information that is actively seeking to mislead, is in our opinion the main factor that can and will make a difference regarding the amount of disinformation running through our screens, along with other educational tools and methodologies long studied in academia.

A number of civil society-led initiatives have spread throughout the region, without institutional support from the States. Argentinian project Chequeado, with support from UNESCO, has released systematized materials to help the general population verify information independently. Chequeado also leads an education program aimed at journalists and journalism students and focused on

[18] Ricardo Monreal, «Iniciativa Con Proyecto de Decreto Por El Que Se Reforman y Adicionan Diversas Disposiciones de La Ley Federal de Telecomunicaciones y Radiodifusión» (2021), https://ricardomonrealavila.com/wp-content/uploads/2021/02/REDES-SOCIALES-Propuesta-Iniciativa-29.01.21.pdf.
[19] Digital Future Society, «Cómo Combatir La Desinformación: Estrategias de Empoderamiento de La Ciudadanía Digital,» May 2020.

fact-checking and data journalism.[20] Nicaraguan initiative Chequialo has made a point of publishing their entire methodology so that anyone can follow along their fact-checking process and arrive at the same conclusions independently.

3. Private measures to counter information disorders

3.1. Flagging, content takedown and account takedown

The ability to flag, remove or block content for misinformation has been present for many years now, but it has been implemented haphazardly. Facebook decided to take down false information regarding the pandemic but not against vaccines[21] and then switched gears a few months after due to a ruling by the Facebook Oversight Board.[22] The ability to flag content as false is not available in general outside of the United States. We have been told by these platforms that allowing this type of flagging globally is unattainable since it would require for them to have the ability to fact-check these claims worldwide.

While flagging seems to be a more sensible effort against disinformation, since it relies on the community, its effectiveness remains unclear. According to Facebook itself, academic research has shown that certain instances of flagging (showing a visual alert, such as a red flag, next to an article) might act in the opposite way than intended, contributing to entrenching «deeply held beliefs» instead of actually reducing their impact.[23] Instead, they found that showing «Related Articles next to a false news story leads to fewer shares than when the Disputed Flag is shown». Furthermore, there seems to be enough evidence to suggest that the report feature is often used out of malice,[24] even to attack a specific user by repeatedly reporting their account in order to trigger a response that will silence it. This is particularly true in gaming platforms, where a user can get banned if reported repeatedly, but we have also seen it happen to grassroots activists, particularly those who perform their activism in LGBTIQ+ spaces, even though the main social platforms have insisted that the number of reports on a certain piece of content or account does not have any effect on the actual measures taken upon such content or account.

3.2. Deprioritizing or delisting

Both Instagram and Facebook started aggressively deprioritizing information during the onset of COVID-19. In this context, deprioritization means the algorithmic demotion of certain content from appearing or being pushed to the user's feed, for instance, accounts being recommended by Facebook to be followed, or in «top content» chosen by the platform instead of the user. For example, when Twitter decided to deprioritize tweets from political figures who broke the platform's rules, they said

[20] El boom del fact checking en América Latina: Aprendizajes y desafíos del caso de Chequeado, September 2014, https://www.kas.de/c/document_library/get_file?uuid=c6a21701-5f10-84ea-397d-dbc75f1a69fe&groupId=287460
[21] Siladitya Ray, «Unlike Covid-19 Misinformation, Facebook Won't Takedown Anti-Vaxxer Posts, Zuckerberg Says,» Forbes, September 9, 2020, https://www.forbes.com/sites/siladityaray/2020/09/09/unlike-covid-19-misinformation-facebook-wont-takedown-anti-vaxxer-posts-zuckerberg-says/.
[22] Mike Isaac, «Facebook Says It Plans to Remove Posts with False Vaccine Claims,» *The New York Times*, February 8, 2021, sec. Technology, https://www.nytimes.com/2021/02/08/technology/facebook-vaccine-misinformation.html.
[23] Lyons, «Replacing Disputed Flags With Related Articles,» *About Facebook* (blog), December 21, 2017, https://about.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation/.
[24] Jillian C. York, «Facebook's 'real Names' Policy Is Legal, but It's Also Problematic for Free Speech,» The Guardian, September 29, 2014, http://www.theguardian.com/commentisfree/2014/sep/29/facebooks-real-names-policy-is-legal-but-its-also-problematic-for-free-speech.

these tweets wouldn't appear in safe search results, in Top Tweets for accounts that were set to display these instead of chronological tweets, in push notifications for recommended tweets, among others.[25]

While this is considered to be a "softer" form of content restriction, since any user can find the information if they know to look for it specifically, it can be pervasively damaging for its lack of transparency, if the user does not know how or why certain information is being shown to them instead of another. This approach risks a certain paternalism that treats all users in the same way that traditional content filters treat underage users, which, in our opinion, can contribute to stilting critical thinking and the ability to distinguish reliable from unreliable information.

3.3. Real-name policies

One of the reasons quoted by platforms as Facebook to preserve and uphold their controversial «real name» policy is that it is designed to quell user misconduct in the platform, including harassment and disinformation. However, this policy is also reported to be responsible for influencing the phenomena of «digital redlining», meaning that it encourages racist, sexist decision-making in the form of policy enforcement: «reporting someone for profanity or racism on Facebook is less likely to elicit a corporate response than reporting them for not using their "real name"», McMillan writes: «it is more common that Facebook will ban non-white, non-male, non-Western users for violating ethical codes when they write against racism or sexism or inequality than they will ban those who post actual racist or sexist content».[26]

Back in 2015, several Latin American and global human rights organizations sent a letter to Facebook[27] claiming that their real-name policy disproportionally and unfairly affected transgender users, ethnic minorities, and users who employ a pseudonym to protect themselves from political violence and harassment. Even though Facebook has since made changes to improve their real-name policies to avoid damages to these affected groups, the policy remains very strict, not accepting pseudonyms or any name that is not directly linked to one's legal name.[28]

The criminalization of anonymity, both via cultural discourse, private terms of service (such as Facebook's), and by state law such as the direct prohibition of anonymity in countries like Venezuela and Brazil, is a dangerous trend that directly affects freedom of speech. The ability to track down authorship of a certain piece of information in order to curb potentially damaging disinformation is not proportional nor it justifies the irreparable damage to the free flow of ideas necessary in a democratic society. The privacy derived from the possibility of preserving one's own identity can sometimes be the only way for certain people to protect their opinions and beliefs, especially in hostile environments.

[25] Zaheer Merchant, «Twitter Will Now Label and Deprioritise Tweets from Political Figures That Break Its Rules | MediaNama,» June 28, 2019, https://webcache.googleusercontent.com/search?q=cache:0XvWPbrcctcJ:https://www.medianama.com/2019/06/223-twitter-will-now-label-and-deprioritise-tweets-from-political-figures-that-break-its-rules/+&cd=16&hl=es&ct=clnk&gl=cl.
[26] Tressie McMillan Cottom, «Digital Redlining After Trump: Real Names + Fake News on Facebook,» Medium, November 14, 2016, https://tressiemcphd.medium.com/digital-redlining-after-trump-real-names-fake-news-on-facebook-af63bf00bf9e.
[27] Carta a Facebook, October 5, 2015, https://www.eff.org/files/2015/10/08/carta_a_facebook_5_oct.pdf
[28] Kantrowitz, Alex. «Facebook Responds To Open Letter Criticizing "Real Names" Policy». BuzzFeed News, 30 de octubre de 2015. https://www.buzzfeednews.com/article/alexkantrowitz/facebook-is-making-enforcement-changes-to-its-real-names-pol

3.4. Blocking or interrupting content sharing

Several recent policies in social platforms and messaging services aim to block, interrupt or slow down the sharing of content deemed as undesirable. A recent feature rolled out by WhatsApp in 2019 labels messages that have been forwarded in a chain of five or more chats as «forwarded many times». These messages then hit a limit and can only be forwarded one chat at a time, which, WhatsApp claims, «helps slow down the spread of rumors, viral messages, and fake news».[29]

A recent Twitter experimental feature launched in June 2020, which prompted users to read the article before sharing tweets which contained links to news sites, was declared successful in September, with the company declaring that people opened the articles 40% more often after being prompted.

While certain tools or markers that allow the user to consider more information before sharing can be useful, some other tools can be considered forms of prior restraint. Facebook does not allow sharing of certain links that are flagged as harmful, with the user getting a message that «the action attempted has been deemed abusive or is otherwise disallowed». While the company has been vocal about their attempts to create channels for appealing their decisions, this kind of behavior is unappealable, since there is no piece of content where to click a link for a claim, since the content does not exist in the first place. Facebook users can appeal to the removal of posts that were taken down for nudity, sexual activity, hate speech, or graphic violence, but there is no procedure for appealing for content that was never allowed to be posted in the first place.

4. Human rights impact of measures against information disorders

Measures against information disorders are themselves subject to analysis of their impact over the exercise of fundamental rights. In particular, by targeting either individuals, discrete pieces of content, or intermediary platforms, actions presented as measures against disinformation will likely have an impact on freedom of expression and opinion, as well as other rights such as access to information and knowledge, right to privacy, freedom of peaceful assembly or to not be discriminated against. It is necessary to weigh the effects of measures aimed at limiting the spread of misinformation on these other rights.

4.1. Freedom of opinion and expression

Civil and political rights can be directly affected by measures against disinformation. In particular, **freedom of expression** (Article 19 International Covenant on Civil and Political Rights, Article 13 American Convention on Human Rights), is often impacted beyond permissible restrictions under international human rights law. As already explained above (sections 2 and 3), many initiatives to combat disinformation have had concrete impacts in the exercise of freedom of expression through the internet, especially those that implement measures that restrict expression for purposes different from

[29] WhatsApp, «WhatsApp Help Center - About Forwarding Limits,» WhatsApp.com, accessed February 15, 2021, https://faq.whatsapp.com/general/chats/about-forwarding-limits/?lang=en.

the war against disinformation, but applied similarly to alleged «fake news» by governments and private companies alike.

There is high risk that these measures create freedom of expression concerns even outside those particular examples. General prohibitions on the dissemination of content or acts of expression, based on vague and ambiguous ideas, including «false news», where usually the category itself is undefined or poorly delimited are incompatible with international human rights law.[30] The aforementioned public measures adopted in the region (sections 2.2, 2.3. and 2.4 above), to the extent that they constitute broad prohibitions over the problematic categories of disinformation or false news, are thus not permissible. The imposition of after-the-fact penalties on individuals, even if formally not a form of prior restraint, can still disproportionately affect freedom of expression by producing an effect of self-censorship. Selective enforcement of these laws and regulations, in charge of governmental bodies such as the police, prosecutorial entities or communication authorities, can be used as a tool for crushing political dissent, thus infringing freedom of opinion. That risk is present in the recently enacted Nicaraguan Special Law on Cybercrime,[31] as well as the Venezuelan Criminal Code since its 2005 reform,[32] both explicitly penalizing alleged false news.

However, even without such legislation, other measures can still generate a chilling effect. In Brazil before the parliamentary elections of 2018, the federal government announced the federal police would search and punish «fake news» ahead of the vote, amounting to an attempt to control expression at a large scale.[33] Intermediary liability rules that impose duties to either monitor content or remove allegedly false information without impartial adjudication (sections 2.3 and 2.4 above) have the effect of giving priority to the interests of intermediaries against those of the individuals, favoring actions akin to prior restraint by limiting the reach of expression before their truthfulness can be assessed. Private actions in the same direction (sections 3.1 and 3.4.), even in the absence of legal mandates, achieve a similar effect.

In Venezuela, a bill named “Constitutional Law of Cyberspace” was discussed by the Constitutional Assembly in 2019, creating liability for intermediaries and establishing sanctions if they did not take down illegal speech. It forced messaging service providers (which may include social networks and instant messaging services) to censor content without prior judicial order, respect for minimum guarantees of freedom of expression, or due process. They would be also compelled to have the arduous duty of “preventing, reporting, neutralizing, or eliminating the disclosure of data and information that threatens the honor, privacy, intimacy, own image, reputation of the people, deceptive and illicit advertising, hate promotion, intolerance, discrimination, harassment, sexual

[30] Declaration by the United Nations Special Rapporteur on Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe Representative on Freedom of the Media, the Organization of American States (OAS) Special Rapporteur on Freedom of Expression and the African Commission on Human and Peoples’ Rights Special Rapporteur on Freedom of Expression and Access to Information., «Joint Declaration on Freedom of Expression and ‘Fake News’, Disinformation and Propaganda,» March 3, 2017, https://www.osce.org/fom/302796.
[31] AP News (October 27, 2020), «Nicaragua approves «cybercrimes» law, alarming rights groups», https://apnews.com/article/legislature-legislation-crime-daniel-ortega-cybercrime-ce252ed4721a759ed329798a7e2e30db
[32] Código Penal de Venezuela http://www.annaobserva.org/observatorio/wp-content/uploads/2018/03/C%C3%93DIGO-PENAL-DE-VENEZUELA.pdf
[33] Gizmodo (January 8, 2018), «Brazil’s Federal Police Says it Will 'Punish' Creators of 'Fake News' Ahead of Elections», available at: https://gizmodo.com/brazil-s-federal-police-says-it-will-punish-creators-of-1821945912

exploitation, child pornography, or economic, political or social destabilization of the Nation".[34] The initiative has not become law.

4.2. Freedom of information

Notoriously, the **freedom to seek, receive, and impart information and ideas**, an integral part of freedom of expression, is affected by measures, especially those by state actors. This is especially notable with the adoption of measures such as blocking websites or banning service accounts, based on vague, spurious or unsubstantiated claims of disinformation or «fake news», or the spread of such content (as explained in sections 3.1, 3.2 and 3.4 above). When enforced against news organizations and journalists, measures to combat disinformation can affect not only the right to impart information and the freedom of the press, but also the capacity of citizens to access different sources of information. Shaping one's own opinion freely is thus profoundly affected, as is participation in public affairs and the involvement in political debate.

As recently highlighted with regards to health disinformation, bans and blocks are not necessarily the best way to provide accurate information,[35] especially when there is no evidence of imminent physical harm. Less intrusive means of combating information disorders are useful without infringing as strongly on the rights of individuals. This is especially true with regards to the measures adopted by private platforms without intervention from the authorities (as explained in sections 3.1, 3.2 and 3.4 above).

4.3. Right to privacy and informative self-determination

The **right to privacy**, including the protection against interference in private life and private communications (see Article 17 ICCPR, Article 11 ACHR among others), can also be affected, when information control measures require identification between individuals and acts of expression. Several HRC resolutions and previous mandate-holders have stated the importance of privacy as a reinforcing right to freedom of expression, as well as the risks that both face in the digital age.[36] Although not necessarily meant specifically as measures against disinformation, restrictions on anonymity can not only limit expression and information (as explained in the examples in section 3.3), but may also place an undue burden on persons to give up personal data to be part of a communication environment, if only for the eventual liability that may be placed upon them, and regardless of their communicative purpose. Mandates against anonymity must be understood as incompatible with international human rights law.

In Ecuador, Communications Law between 2013 and 2019 made media websites liable for user comments unless users were registered and identified. Around 2015, then-president Correa publicly decried the anonymity of an online satire group that in his view were spreading false information

[34] See the joint declaration from Civil Society Organisations from the region "Against the Constitutional Law of Cyberspace bill of the Bolivarian Republic of Venezuela", available at: https://www.accessnow.org/against-the-constitutional-law-of-cyberspace-bill-of-the-bolivarian-republic-of-venezuela/
[35] Oversight Board, Case decision 2020-006-FB-FBR, https://www.oversightboard.com/decision/FB-XWJQBU9A/
[36] See, for example, Human Rights Council resolutions A/HRC/34/7 and A/HRC/39/29.

about him through memes, a group that was also subject to frequent content takedowns and account bans from different online services,[37] largely because of the political nature of the content they shared.

In Colombia, a bill presented in 2017 criminalised the use of anonymous accounts to «spread false news that could create confusion» in the public,[38] thus shifting attention towards anonymous acts of expression.

In Brazil, the «fake news» bill under discussion requires users to register for social media and private messaging services using government-issued identifiers, and also attempts to mandate traceability between messages and authors, thus impeding true control over personal data and communications information. Provisions like these not only allow, but facilitate, mass data collection and processing.[39]

Additionally, the accusations against alleged disinformation or «fake news» can be used by state actors to illegally surveil persons and communications of those involved in acts of journalism, as it happened after an investigation on police misconduct in Chile following a large police operation based on false evidence called «Operación Huracán»,[40] where journalists were falsely accused of lying to the public as justification for these practices.

4.4. Other rights

The lack of access to different sources of information and opinion not only amounts to a limited communication environment. It can affect the ability to form one's opinion, but also to explore one's own liberty and self-discovery. Because measures against disinformation may disproportionately impact access to information, diminished access can have a negative impact on the exercise of **cultural, economic, social and environmental rights**, contingent to one's ability to obtain information from different sources and forming one's own capacity to realize those rights. It can prevent people from engaging with other citizens with similar opinions or concerns, thus negatively impacting **freedom of association,** as exemplified above (section 2.1., 2.2 and 2.3), by limiting the capacity to organize around common concerns and collaborating in political action.

Measures targeted at preventing health disinformation through blocking or banning contents can prevent people from adopting appropriate measures to care for one's health and well-being, if valuable information loses reach. Measures to revert disinformation can also backfire, reinforcing belief in false information,[41] making it key to maintain levels of transparency and openness with regards to information handling by governments and platforms alike.

[37] The New York Times (May 3, 2015), "What Happened When I Joked About the President of Ecuador", https://www.nytimes.com/2015/05/03/magazine/what-happened-when-i-joked-about-the-president-of-ecuador.html
[38] El Espectador (July 17, 2017), «Proyecto de ley buscará combatir cuentas falsas en redes sociales», https://www.elespectador.com/noticias/politica/proyecto-de-ley-buscara-combatir-cuentas-falsas-en-redes-sociales/
[39] GNI (2020), «GNI Expresses Concern About Proposed 'Fake News' Law in Brazil». Available at: https://globalnetworkinitiative.org/gni-concerns-brazil-fake-news-law/
[40] CIPER (2018), «Los periodistas que fueron objeto de espionaje electrónico de Carabineros». Available at: https://ciperchile.cl/2018/03/07/los--periodistas--que--fueron--objeto--de--espionaje--electronico--de--carabineros/
[41] Lewandowsky, S., et al. "Misinformation and Its Correction: Continued Influence and Successful Debiasing." *Psychological Science in the Public Interest*, vol. 13, no. 3, 2012, pp. 106–131.

Undue restrictions on freedom of expression, as carried out through the enforcement of rules or the enactment of proactive measures against sources of content or information, can be in and of themselves forms of infringement on the right to due process guarantees, including adjudication leading to such restriction by an independent, impartial, authoritative oversight body such as a court of law. Without due process, even limited measures can still be incompatible with international human rights law.

5. Grievances and remedies in the private sector

5.1. Oversight boards

Facebook's Oversight Board is a quasi-judicial body announced in 2018 and which started operating in May, 2020, with the purpose of making content moderation decisions on Facebook. Among its stated goals, the Board intended to «prevent the concentration of too much decision-making within [Facebook's] teams», to «create accountability and oversight», and to «provide assurance that these decisions are made in the best interests of [Facebook's] community and not for commercial reasons.[42]

Criticism of the Facebook Oversight Board points out that it is funded by Facebook, even though the funding is conducted through a trust, and awarded a limited mandate in scope, making it unaccountable in many ways and incapable of being entirely objective regarding its judgment of Facebook's behavior. Vaidhyanathan[43] points out:

> «[The board] will hear only individual appeals about specific content that the company has removed from the service—and only a fraction of those appeals. The board can't say anything about the toxic content that Facebook allows and promotes on the site. It will have no authority over advertising or the massive surveillance that makes Facebook ads so valuable. It won't curb disinformation campaigns or dangerous conspiracies. It has no influence on the sorts of harassment that regularly occur on Facebook or (Facebook-owned) WhatsApp. It won't dictate policy for Facebook Groups, where much of the most dangerous content thrives. And most importantly, the board will have no say over how the algorithms work and thus what gets amplified or muffled by the real power of Facebook.»

According to its mandate the Oversight Board provides «guide» to Facebook for the modification of its policies, without making a change of these policies mandatory for the company. Although it is understandable that an external body does not have direct influence on the platform's rules, the decision to formulate those rules will remain exclusively in the hands of the company. In this regard, the ambiguous language used with respect to the binding nature–or not–of the recommendations on policy modification, versus the explicit recognition of the binding nature of the decisions referring to specific cases that are submitted to them, is striking. This ambiguity threatens the recommendations made by Derechos Digitales and other organizations during the consultation period.[44] The minimum

[42] Zuckerberg, «A Blueprint for Content Governance and Enforcement». 2018.
[43] Siva Vaidhyanathan, «Facebook and the Folly of Self-Regulation,» *Wired*, September 5, 2020, https://www.wired.com/story/facebook-and-the-folly-of-self-regulation/.
[44] AI Sur contribution to Facebook public consultation, p. 126, available at: https://about.fb.com/wp-content/uploads/2019/06/oversight-board-consultation-report-appendix.pdf

expected in this case, if there is no binding effect of the Oversight Board's recommendations, is for Facebook to endeavor to provide sufficient and reasoned explanations to justify its refusal to follow the recommendations.

Even more relevant in terms of the current limitations of this mechanism is that the process of reaching a decision and its implementation by Facebook is centered on determining whether a decision to remove content is consistent with the content policies and «values» of the company. In other words, it is about monitoring the observance of the rules of a private company, and not of national or international legal rules, in situations that reach the entire planet. Although explicit mentions are included for the first time to the necessary attention to the «human rights norms that protect freedom of expression» in its constitution document and to international human rights law among the values that inspire Community Norms (since most recent updates), it is necessary that a real commitment to fundamental rights be more explicit and broader; as well as a higher hierarchy than the values established internally by the company. Just raising the values of freedom of expression and not other human rights involved in content moderation–in line with the US perspective–is insufficient to recognize the frameworks of international human rights law and the regulatory complexities that may vary around the world.[45]

After the Oversight Board has started to deliver its firsts decisions, some have started to wonder about the possibility to expand the mechanism to other platforms and expanded mandate closer to what it was proposed in 2019 by ARTICLE 19 through the Social Media Council model.[46] The proposal here was the creation of several regional or local councils rather than one operating at global level, and with a broader jurisdiction over a number of different platforms. This idea was confronted with some skepticism from the global south countries context, in which having these councils publicly financed or in any other form influenced by the participation of governments, or exposed to the risk of pressure from public powers over the regional or local members, make it difficult to ensure its required independence.

5.2. Appeal mechanisms

As we mentioned above, appeal mechanisms have been incorporated into platforms relatively late. In Twitter, users can appeal against permanent suspensions or content removal, but not against other measures such as placing an account on read-only mode (a measure popularly known as «Twitter jail» and that can range from 12 hours to 7 days) or against the placing of a tweet behind a «sensitive» notice, a measure initially intended for adult media and graphic content.

As for Facebook, the appeal mechanism was rolled out in 2018, by allowing appeals against content takedown.[47] Instagram put in place a system where a decision to remove content can be reviewed «in most circumstances» but not «for some types of content», while leaving it unclear what those types of

[45] See more at Derechos Digitales, Consejo Asesor de Facebook: ¿La bala de plata para los problemas de la moderación de contenidos?, September 27, 20219, available: https://www.derechosdigitales.org/13885/la-bala-de-plata-para-los-problemas-de-la-moderacion-de-contenidos/
[46] See The Social Media Councils: Consultation Paper, June 2019, available at: https://www.article19.org/wp-content/uploads/2019/06/A19-SMC-Consultation-paper-2019-v05.pdf
[47] Zuckerberg, «A Blueprint for Content Governance and Enforcement». 2018.

content might be.[48] This revision can then be appealed, but only in some cases, before the Oversight Board:

> "Not all content decisions are eligible for appeal to the Oversight Board. If an Instagram content decision is eligible for review by the Oversight Board, you'll see an Oversight Board Reference ID within your Support Requests".[49]

The lack of transparency behind where the limits are is problematic in many ways. For a user, not only not being able to appeal but not being able to know beforehand that decisions over a specific kind of content cannot be appealed makes the entire process an arbitrary maze which they can only navigate in very specific ways over which they have no control.

This has motivated from civil society the requirement of improvements in terms of due process of the appeal mechanisms (among others). It is the case of the Santa Clara Principles on transparency and accountability in content moderation, a civil society proposal that came from the United States in 2018. Here, recommendations specifically aimed at companies were developed, focused on minimum standards for a meaningful appeal which should include: Human review by a person or panel of persons that was not involved in the initial decision; An opportunity to present additional information that will be considered in the review; Notification of the results of the review, and a statement of the reasoning sufficient to allow the user to understand the decision. As a desirable element "in the long term", independent external review processes are also identified as an important component for users to be able to seek redress.[50]

5.3. Community-driven approaches

Twitter's community-driven initiative Birdwatch is meant to allow users «to identify information in Tweets they believe is misleading and write notes that provide informative context». This initiative is only available in the US, is currently in pilot phase, and therefore it is difficult for us to have an opinion regarding its effectiveness. However, we have seen that community-based approaches that rely on adding context instead of subtracting information can be valuable. For a long time, Reddit has relied on its volunteer moderators to keep the site acceptably reliable, an approach that has proved to be at the same time extremely powerful yet insufficient, as communities such as #Gamergate and conspiracy theories have found in the platform a place to live and thrive.[51]

5.4. Transparency efforts

Most major social media platforms and telecommunication companies have certain standard practices for issuing transparency reports. Transparency reports should, at least in theory, be regularly released publications that should serve to review activities that impact freedom of expression and privacy, containing both details about the enforcement of community guidelines and terms of service, and

[48] Instagram, «I don't think Instagram should have taken down my post». https://help.instagram.com/280908123309761
[49] Instagram, «How do I appeal Instagram's content decision to the Oversight Board?» https://help.instagram.com/675885993348720
[50] See https://santaclaraprinciples.org/
[51] Steven Melendez, «'I Have a Duty to Do This': Meet the Redditors Fighting 2020's Fake News War,» Fast Company, March 2, 2020, https://www.fastcompany.com/90466966/i-have-a-duty-to-do-this-meet-the-redditors-fighting-2020s-fake-news-war.

requests from third parties (public and private) regarding user data and restrictions to content and accounts.

According to Access' Transparency Reporting Index, Latin America has the lowest regional diversity in transparency reporting[52]. An overwhelming proportion of transparency reports come from companies based in North America, and even though our own research has found that local transparency reports from telecommunication companies have become more of a standard in higher-income countries like Chile[53], this is not the case of the rest of the region, where there is still a huge disconnect between transparency practices led by local companies and those that are subsidiaries of foreign companies.[54]

Even when those transparency reports exist for the region, usually the level of detail and disaggregated information they provide is limited which makes them less useful. In the same lines, the orientation of those transparency reports does not address the issue of how policies and content moderation rules are applied across jurisdictions, making it difficult to evaluate the response of the concerned platforms in similar cases that have arisen in different countries, particularly from the global south. The recent decision made by Twitter to suspend the US President Donal Trump in the wake of violent actions taken for a group of his followers to take on the Capitol -and a similar decision adopted by Facebook later- was the last chapter in the mumbling of the platform to articulate and enforce clear rules regarding their own standards. The claim from Twitter is that the banning was applied following its pre-existent "civic integrity rules" and that the decision was made considering the cumulative impact of several violations to the policy in the last weeks of Trump's presidency. Voices from human rights advocates from all around the world were raised to say that even though those factors have been present in many cases that have happened in other jurisdictions (notably in the global south) no similar measures were taken.[55]

In summary, it is not clear how the contextual interpretation of circumstances that was possible for a US company regarding US political impacts, it is an element possible to be replicated when comes to decision making for content moderation on other jurisdictions, such as Latin American countries, in which the platforms cannot have locally-based staff or enough nuance in its understanding of the languages and cultural norms implied in the understanding of the content shared. Transparency is required then to better understand the different elements that can lead to this decision making in other contexts.

[52] Access Now. «Transparency Reporting Index», octubre de 2019. https://www.accessnow.org/transparency-reporting-index/.
[53] Derechos Digitales. «¿Quién defiende tus datos? 2019 | Derechos Digitales». Accedido 15 de febrero de 2021. https://www.derechosdigitales.org/qdtd-2019/.
[54] Asociación por los Derechos Civiles. «¿Quién defiende tus datos? 2019». Asociación por los Derechos Civiles, 4 de marzo de 2020. https://adc.org.ar/2020/03/04/quien-defiende-tus-datos-2019/.
[55] See After Barring Trump, Facebook and Twitter Face Scrutiny About Inaction, Abroad, New York Times, available at: <https://www.nytimes.com/2021/01/14/technology/trump-facebook-twitter.html#click=https://t.co/oHG5oXgz0X>

6. Recommendations

Based on the information and comments provided in this submission we respectfully recommend the Special Rapporteur to focus on the following lines of work to protect and promote the right to freedom of opinion and expression while addressing disinformation:

6.1. Regarding the private sector

The development and enforcement of community standards and other forms of self-given rules for content platforms has been problematic from a human rights perspective since the very beginning, but the problems were exponentially increasing once the social media platforms reached a user base at a global level that transforms them into the new "public space" for the interchange of ideas and the influence of politics and public discourse.

The private rules of content moderation allow platforms to directly decide on how their users are able (or not) to exercise their freedom of expression, but by doing so, it is not only expression that is impacted, but rather the full range of possibilities of participation in the public debate that takes place in them–as explained–and its consequences extend well beyond the digital space, having direct consequences in the ability of groups to share their ideas and reach audiences with them, and become those ideas in actionable exercise of public power through elections or mobilization of groups to support political perspectives. In that sense, the impact of content moderation decisions from platforms extend to the whole range of exercise of fundamental rights.

Content platform companies have responsibilities to respect human rights according to the UN Guiding Principles on Business and Human Rights.[56] This mandate has expressed previously that *"human rights law gives companies the tools to articulate and develop policies and processes that respect democratic norms and counter authoritarian demands"*.[57] To be clear, the problem of the content moderation by platforms according to their own rules has fallen short in both sides according to human rights standards: there are cases in which public discourse that can be controversial or shocking is taken down or reduced in its visibility, silencing or chilling expressions from vulnerable or traditionally marginalized groups (e.g. cases related to women sexual reproductive rights, or LGBTQI+ expressions) or political dissenting voices; while in other cases the platforms have allowed disinformation content that lead to physical harm (e.g. content inciting genocide against the Rohingya in Myanmar).

The hesitant and inconsistent behavior from platforms in addressing their problematic use by different political leaders or groups that spread content that cannot be qualified as illegal, but arguably is barely allowed by international standards of freedom of expression, has put the platform content moderation on the public eye and questioning its inconsistency. The UN Guiding Principles on Business and

[56] See Human Rights Council resolution 17/4 of 16 June 2011. "Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework", available at: <https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf>

[57] UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. A/HRC/38/35 of 6 April 208. "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression", available at: <https://www.undocs.org/A/HRC/38/35>

Human Rights should provide a path forward for the request by this mandate and the Human Rights Council from the companies further accountability and transparency in their decision making at a global level and further stressing their obligation to respect and protect the human rights of their users by due process implementation in their internal procedures and by directly linking their private rules with international human rights standards.

These rules should be interpreted in a manner that encourage the platforms to perform a portfolio approach of measures directed to take responsibility for information disorders. From the onset platforms should privilege the provision of context for the circulation of information through tags, fact-checking, or prompted alternative informations, when a content is identified as intended to promote a public positioning in a way that impacts the exercise of fundamental rights. Reduced algorithmic visibility and limitation on further sharing can be an additional step to implement when dealing with content that prompt information disorders. Removal of content should be considered always in the most limited way possible when the previous approaches prove insufficient to ensure the protection of human rights. In those cases, content removal should be proportionate and geographically limited. Platform should provide transparent information in the way in which information disorders are identified, including what content has been human or automatically flagged and what additional steps of verifications have been taken.

6.2. Regarding State action

Regulations directly addressing disinformation issues have proven problematic in Latin America, more often than not used to quash dissent and political opposition. Disinformation laws that force platforms making direct decisions about content removal, without judicial intervention, seem incompatible with the prohibition of prior restraint in Article 13 of the American Convention on Human Rights.

Recommendations for States should be directed to the implementation of public policies that support digital literacy of their citizens to allow a responsible and critical engagement with the information provided by online platforms. Further efforts and public resources should be devoted from an education perspective to address the problem in a sustainable manner for supporting an active digital citizenship from generations that have been born under the influence of digital networks and media.

From a regulatory perspective, States should be encouraged to test regulatory approaches that impact on the business models of platforms and their internal rules that facilitate their abuse for the spread of information disorders. Rather than attempt to impose directly or indirectly obligations of policing content, State regulation through consumer protection or other specific regulation could advance obligations for digital platforms that require them implementing content moderation that is consistent with international human rights standards, that is enforced according due process, that ensures transparency in its application (in the way described in the previous section) and that provides mechanisms of appeal and redress.

Additional ways to confront information disorders in electionary context should be explored through the updating of elections legal frameworks in which the electoral authorities could be entrusted of prowers for the oversight of the use of digital platforms for political advertisement and the expending

is subject to transparency requirements from the political actors and from the platforms. Personalized profiling and political advertising that look for fragmenting and polarizing audiences in electoral context could be banned through electoral laws.

Information disorders could be further addressed through the limitation of collection of specific sensitive information of the users through consumer regulation or other regulatory framework that prevent the exploitation of certain characteristics from the users such as political inclination, sex, sexual orientation or ethnic origin for the purpose of profiling and provide suggestions of content.[58]

In summary, we see more benefit in approaching information disorders through the handling of the misplaced incentives in the functioning of the platforms, than in the direct intervention from the private actors or the State in the decision-making on the nature of specific content, that always will be problematic from the perspective of its impact in the freedom of expression exercise and its related consequences for the exercise of other human rights and the ensuring of a resilient and open digital public sphere.

In case you should want to expand in any of the aforementioned points, please reach Maria Paz Canales <mariapaz@derechosdigitales.org>, J. Carlos Lara <juancarlos@derechosdigitales.org> or Marianne Díaz <marianne@derechosdigitales.org>.

[58] See K. Sabeel Rahman & Zephyr Teachout, "*From Private Bads to Public Goods: Adapting Public Utility Regulation for Informational Infrastructure", available at:* https://knightcolumbia.org/content/from-private-bads-to-public-goods-adapting-public-utility-regulation-for-informational-infrastructure