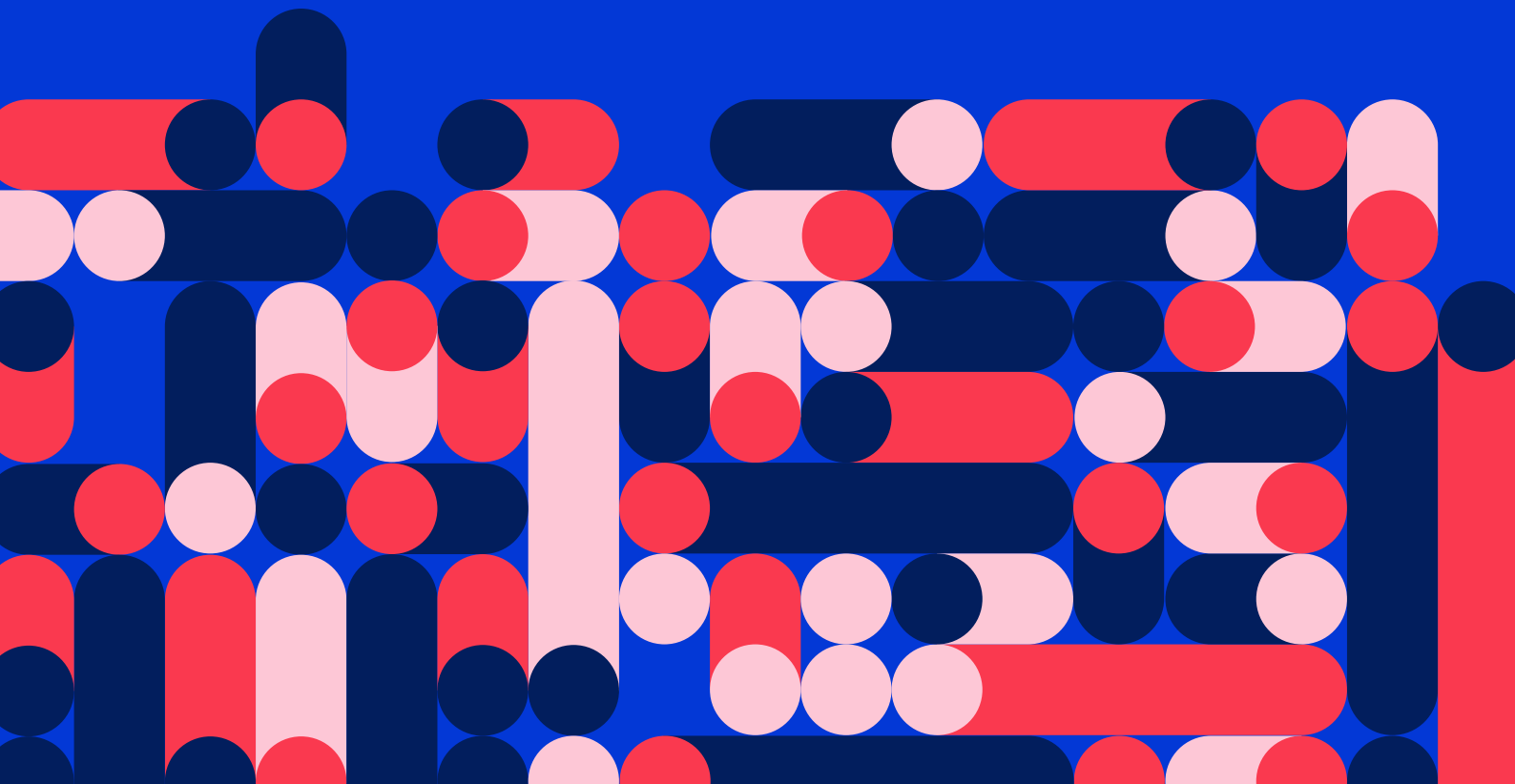
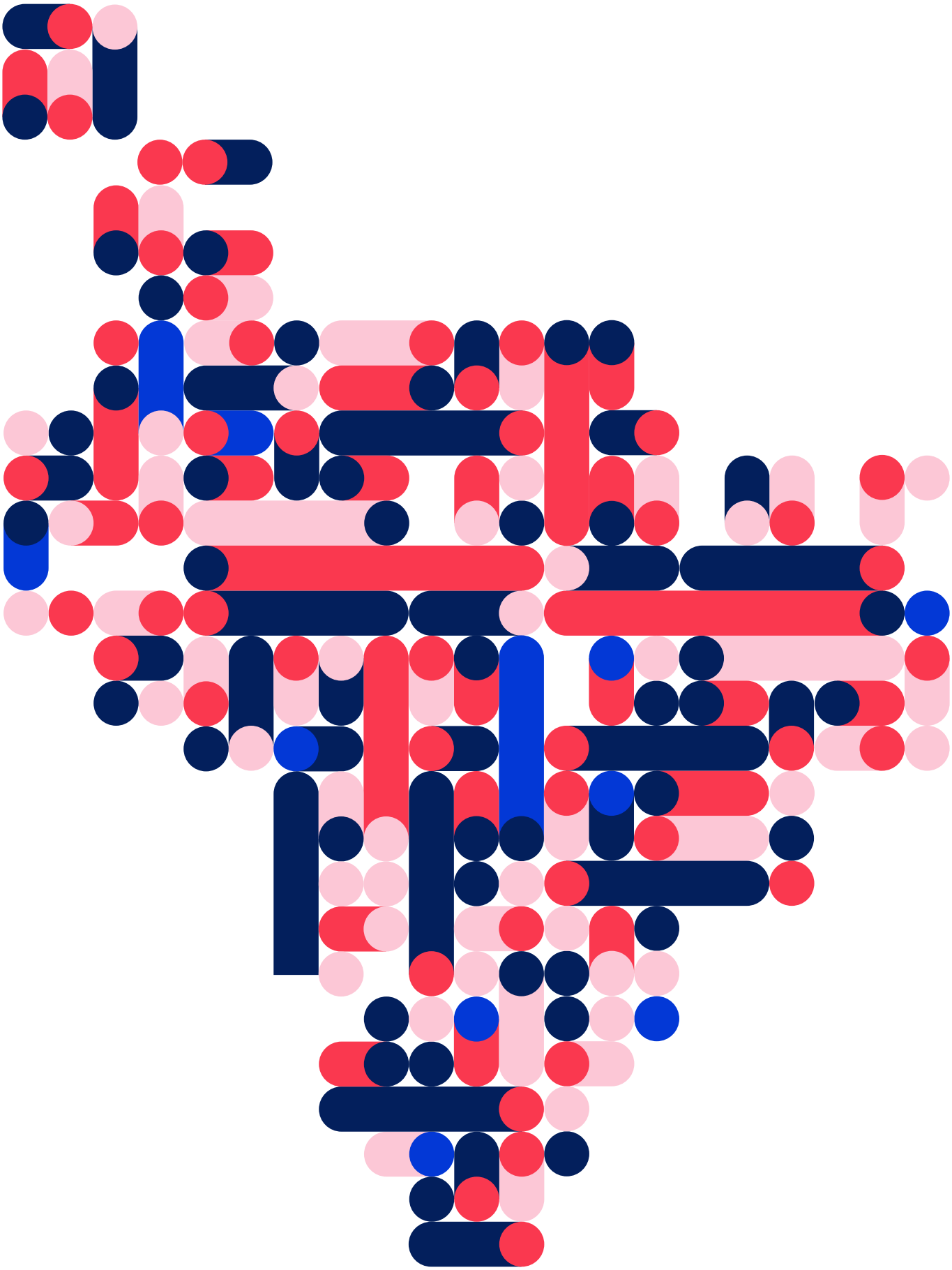
Content moderation from an Inter-American perspective

AlSur





Content moderation from an Inter-American perspective

AlSur

Content moderation from an Inter-American perspective
By: Vladimir Alexei Chorny Elizalde, Luis Fernando García Muñoz,
and Grecia Elizabeth Macias Llanas

Al Sur's contribution to the Dialogue of the Americas on Freedom of Expression on the Internet to gather input for the development of standards on the subject, launched by the Office of the Special Rapporteur for Freedom of Expression (RELE) of the Inter-American Commission on Human Rights (IACHR).

AlSur is a consortium of organizations working in civil society and academia in Latin America that seek to strengthen human rights in the region's digital environment by working together.

Mexico City. Mexico, March 2022

For more information about Al Sur and its members, visit https://www.alsur.lat/en.

Table of contents

I. Right to freedom of expression in the Inter-American system 6
- a. General principles 6
- b. Protected speeches and special protected speeches 9
  - *i. Political speech and speech on matters of public interest* 10
  - *ii. Speeches on public officials and candidates for public office and other public figures* 10
  - *iii. Speeches expressing essential elements of personal identity or dignity* 11
- c. Non-protected speeches 12
  - *i. Incitement to violence* 12
  - *ii. Direct and public incitement to genocide* 12
  - *iii. Sexual abuse of minors (child pornography)* 13
- d. Limitations on freedom of expression 13
- e. Prohibition of prior censorship and indirect restrictions 18
- f. Public officials and freedom of expression 19
- g. Freedom of expression and the Internet 20

II. The liability of Internet intermediaries for expressions of third parties 22
- a. The role of intermediaries on the Internet 22
- b. The principle of non-liability of intermediaries 23
- c. Section 230 of the United States Communications Decency Act of the United States of America 25
  - *i. The Origin of Section 230* 25
  - *ii. The principle of non-liability of intermediaries in section 230* 27
  - *iii. The non-liability for unilateral measures with respect to content moderation* 28
- d. The principle of non-responsibility of intermediaries in commercial treaties 29

III. Content moderation 30
- a. Objectives and justifications for content moderation 30
- b. Moderation rules 31
- c. Moderation Procedures 32
- d. Effects of moderation on freedom of expression and other rights 34
  - *i. Practical considerations and limits of different moderation methods: What happens if there is no moderation?* 34
  - *ii. The effects of vagueness or ambiguity in moderation criteria* 36
  - *iii. The link between concentration and the impact on human rights* 38
- e. The difficulty of moderation at scale 39
- f. Jurisdictional Aspects of Content Management 42

IV. Transparency and accountability 46
- a. The Santa Clara principles 47
  - *i. Transparency* 48
  - *ii. Notification* 49
  - *iii. Appeal* 50

V. Recommendations on the regulation of content moderation by dominant internet intermediaries 51

I. Right to freedom of expression in the Inter-American system

a. General principles

The right to freedom of expression, acknowledged in Article 13 of the American Convention on Human Rights (ACHR), has special treatment within the Inter-American Human Rights System (hereinafter referred to as the IAHRS). For this reason, the Inter-American Court of Human Rights (hereinafter "IACHR Court") has highlighted the importance of freedom of expression and has reiterated that this represents the cornerstone of democratic societies and that it is also an essential condition for them to be sufficiently informed.[1]

Based on the above, the IACHR Court has emphasized that "[...] the guarantees of freedom of expression contained in the [ACHR] were designed to be the most general and to minimize restrictions on the free circulation of ideas".[2] As a consequence, the restrictions of other systems—such as the European system—cannot be directly applied in the Inter-American framework.

The Inter-American Commission on Human Rights (IACHR) and the Inter-American Court of Human Rights have played a key role in providing freedom of expression with content. One of the particularly significant aspects emerging from the Inter-American doctrine is the acknowledgment of the dual dimension, individual and collective, of the right to freedom of expression.[3]

The double dimension "requires, on the one hand, that no one be arbitrarily undermined or prevented from expressing their own thoughts and therefore represents the right of each individual; but it also implies, on the other hand, a collective right to receive any information and to know the expression of the thoughts of others."[4] This means that people have the right to express their point of view and to hear and know other people's points of view,[5] thus, if an act

1 I/A Court H.R., Series C No. 73. Series C No. 73. Case of "The Last Temptation of Christ" (Olmedo Bustos et al.) v. Chile. Merits and Reparations and Costs. Judgment of February 5, 2001, para. 68.

2 I/A Court H.R., Series A No. 5. Series A No. 5. Compulsory Membership in an Association of Journalists (Arts. 13 and 29 of the American Convention on Human Rights). Advisory Opinion OC-5/85 of November 13, 1985, para. 52.

3 IACHR, Office of the Special Rapporteur for Freedom of Expression. Inter-American Legal Framework on the Right to Freedom of Expression. OEA/Ser.L/V/II IACHR/RELE/INF.2/09, December 30, 2009, para. 2.

4 IACHR Court. Series C No. 107. Case of Herrera Ulloa v. Costa Rica. Preliminary Objections, Merits, Reparations and Costs. Judgment of July 2, 2004, para 108; Series C No. 111. Case of Ricardo Canese v. Paraguay. Merits, Reparations and Costs. Judgment of August 31, 2004, para. 77; and I Series C No. 74. Case of Ivcher Bronstein v. Peru. Reparations and Costs. Judgment of February 6, 2001, para. 146.

5 I/A Court H.R., Advisory Opinion OC-5/85. Compulsory Membership in an Association of Journalists (arts. 13 and 29 American Convention on Human Rights). November 13, 1985, Series A No. 5, para. 33; IACHR, Office of the Special Rapporteur for Freedom of Expression. Inter-American Legal Framework on the Right to Freedom of Expression. OEA/Ser.L/V/II IACHR/RELE/INF.2/09, December 30, 2009, para. 15; I/A Court H.R., Case of Palamara Iribarne v. Chile. Merits, Reparations and Costs. Judgment of November 22, 2005, Series C No. 135, para. 107; I/A Court H.R., Case of Ricardo Canese v. Paraguay. Merits, Reparations and Costs. Judgment of August 31, 2004, Series C No. 111, para. 81; IACHR, arguments before the I/A Court H.R. in the case of Herrera Ulloa v. Costa Rica, transcribed in: I/A Court H.R., Case of Herrera Ulloa v. Costa Rica. Preliminary Objections, Merits, Reparations and Costs.

of a State affects or restricts the individual dimension of the right of the issuer, it affects in the same way and to the same extent the social dimension of the recipient.[6]

When the IACHR Court, for instance, has discussed the issue of controlling false reporting, it has upheld this relationship between the two dimensions of freedom of expression:

> "...it would not be lawful to use societies' right to be truthfully informed as a basis for the prior censorship and elimination of information which is false in the opinion of the censor. Nor would it be admissible, on the basis of the right to disseminate information and ideas, to set up public or private monopolies over the media in an attempt to shape public opinion according to a single point of view."[7]

The Inter-American human rights system sets out a system of duties and responsibilities that has different scopes and requires different actions from the different subjects that may be involved with the rights contained in the ACHR. To ensure the fulfillment of freedom of expression, the Inter-American framework requires the State to take negative measures or abstention from the rights, for example by not legislating against freedom of expression, but also to take positive measures to make the right truly effective, for example by taking action against particular or private actors to prevent their actions from violating some dimension of the right[8] (as in the cases of anti-monopoly or anti-concentration of the media).

The relation of state obligations to the actions of non-state actors has been recognized on various occasions by the Inter-American Court. Two examples are the Juan Humberto Sánchez case and the Maritza Urrutia case, where the Court explicitly stated that the framework of the ACHR "recognizes positive duties that impose specific demands on both State agents and non-State actors. recognizes positive duties "that impose specific requirements on both State agents and third parties acting with the State. and third parties acting with their tolerance or acquiescence and who are responsible for the detention" (paragraphs 81 and 71, respectively). In this regard, see: I/A Court H.R., Case of Juan Humberto Sánchez v. Colombia. Case of Juan Humberto Sánchez v. Honduras, Judgment of June 7, 2003, Preliminary Objections, Merits, Reparations and Costs; I/A Court H.R., Case of Maritza Urrutia v. Honduras, Judgment of

---

6 Reparations and Costs. Judgment of July 2, 2004, Series C No. 107, para. 101-1-a; IACHR, Merits Report No. 90/05, Case No. 12.142, Alejandra Marcela Matus Acuña, Chile, October 24, 2005, para. 39.

7 I/A Court H.R., Advisory Opinion OC-5/85. Compulsory Membership in an Association of Journalists (arts. 13 and 29 American Convention on Human Rights). November 13, 1985, Series A No. 5, para. 33; IACHR, Office of the Special Rapporteur for Freedom of Expression. Inter-American Legal Framework on the Right to Freedom of Expression. OEA/Ser.L/V/II CIDH/RELE/INF.2/09, 30 December 2009, para. 1. Fake news is, at its core, nothing more than disinformation or promotion of inaccurate information. I/A Court H.R., Advisory Opinion OC-5/85. Compulsory Membership in an Association of Journalists (arts. 13 and 29 American Convention on Human Rights). November 13, 1985, Series A No. 5, para. 33; IACHR, Office of the Special Rapporteur for Freedom of Expression. Inter-American Legal Framework on the Right to Freedom of Expression. OEA/Ser.L/V/II IACHR/RELE/INF.2/09, December 30, 2009, para. 1.

8 The relation of state obligations to the actions of non-state actors has been recognized on various occasions by the Inter-American Court. Two examples are the Juan Humberto Sánchez case and the Maritza Urrutia case, where the Court explicitly stated that the framework of the ACHR "recognizes positive duties that impose specific demands on both State agents and non-State actors. recognizes positive duties "that impose specific requirements on both State agents and third parties acting with the State. and third parties acting with their tolerance or acquiescence and who are responsible for the detention" (paragraphs 81 and 71, respectively). In this regard, see: I/A Court H.R., Case of Juan Humberto Sánchez v. Colombia. Case of Juan Humberto Sánchez v. Honduras, Judgment of June 7, 2003, Preliminary Objections, Merits, Reparations and Costs; I/A Court H.R., Case of Maritza Urrutia v. Honduras, Judgment of June 7, 2003, Preliminary Objections, Merits, Reparations and Costs. Case of Maritza Urrutia v. Guatemala, Judgment of November 27, 2003, Merits, Reparations and Costs.

June 7, 2003, Preliminary Objections, Merits, Reparations and Costs. Case of Maritza Urrutia v. Guatemala, Judgment of November 27, 2003, Merits, Reparations and Costs.

The IACHR Court acknowledges a criteria of State obligations that are first divided, in general, into obligations to respect and obligations to ensure rights, which then include, in particular, the obligations to protect, to create institutions to carry out research, punish and redress, and to promote human rights (these last three are conceptually included within the obligation to guarantee rights).

As for negative obligations, the clearest example is the obligation to respect, which implies that the authorities should not carry out actions that violate human rights: this dimension acknowledges the classic view of rights as individual spheres that keep the State at a distance and restrict its power vis-à-vis individuals.[9]

Regarding positive obligations, both the obligation to protect (to ensure that people's rights are not violated either by the authorities or by private parties) and the obligation to guarantee (consisting of the adoption of State measures—all those necessary—to create the conditions for the effective enjoyment of the rights), unfold a set of actions that go beyond the classic vision of the State and give it an active role not only as a central actor for the full exercise of rights, but also to ensure that other non-State subjects do not hinder them and instead comply with them.[10]

The framework of positive obligations focuses on the role of private subjects in the respect and guarantee of human rights. The focus of this paper is on the work of companies and platforms that have a real possibility—or power—to alter the flow of information and affect rights such as access to information and freedom of expression.[11] A clear example of this type of obligation can be found in the framework of the *Guiding Principles on Business and Human Rights*, which is probably the most important instrument in this area and sets out in detail the duties of companies and other private parties to respect, protect and remedy human rights violations. This framework not only acknowledges the general obligation to respect rights, but also specific obligations to *act with due diligence* and to *be transparent*, as well as the obligation to *make reparations* for rights violations within the framework of their competencies.[12]

---

9 I/A Court H.R., Case of Velásquez Rodríguez v. Honduras. Case of Velásquez Rodríguez v. Honduras, Merits, Judgment of July 29, 1988, Series C No. 4, para. 165.

10 Salazar Ugarte, Pedro. The Constitutional Reform of Human Rights. A Conceptual Guide, Mexico, Belisario Domínguez Institute, 2014, pp. 112-117. The positive nature of these state obligations requires effective action against private individuals, ranging from taking measures, for example, against a company that pollutes the environment, to those necessary for companies to respect privacy and freedom of expression (issues particularly relevant to the work at hand).

11 The discussion on the "horizontal effectiveness" of rights raises an issue that is usually displaced or minimized in academic and political criticism: the fact that some companies have sufficient capacity and power to be considered as legally bound by the set of human rights to which they are related; that is, that some non-state actors not only can—and indeed do—violate human rights, but are also concretely obliged to carry out certain types of actions to respect and guarantee them. In this regard, see: Ziemele, Ineta. "Human Rights Violations by Private Persons and Entities: The Case-Law of International Human Rights Courts and Monitoring Bodies", European University Institute-Academy of European Law, EUI AEL; 2009/08; Nolan, Aoife (2014), "Holding non-state actors to account for constitutional economic and social rights violations: Experiences and lessons from South Africa and Ireland", I-CON (2014), Vol. 12 No. 1, pp. 61-93; Chorny, Vladimir. "The violation of hu- man rights by non-state subjects: a comprehensive view of rights." *Latin American Journal of Political Philosophy*, March 2018.

12 Human Rights Council. *Guiding Principles on Business and Human Rights*, United Nations, 2011.

It is easy to think of cases where companies have these obligations *vis-à-vis* freedom of expression on the Internet. Content moderation is probably one of the most interesting obligations, and the ISHR is particularly relevant about these relationships because it acknowledges that freedom of expression must be guaranteed to any person without any discrimination, within a complex framework of duties and responsibilities that depend on the specific situation in which the right is exercised and the technical procedure used to express and disseminate that expression.[13]

Freedom of expression thus has a particular flexibility that must be taken into account. A good example of its flexibility can be found in what is known as specially protected speech and speech that does not have the enhanced defense of the right to freedom of expression.

b. Protected speeches and special protected speeches

The specific aspects of the right to freedom of expression protect the different types of expression regardless of their form, content or means of communication, as stated in Article 13 of the ACHR, which provides that:

> "Everyone has the right to freedom of thought and expression. This right includes freedom to seek, receive, and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing, in print, in the form of art, or through any other medium of one's choice."[14]

All expressions (oral, written, artistic, etc.) are protected "from the beginning" (often referred to as *ab initio* coverage), which means that there is a presumption that all expressions are protected even if they may be considered shocking, offensive or disturbing. As a general rule, it is a right subject to a very limited regime of exceptions, expressly and specifically defined in international law by means of concrete and specific prohibitions.[15]

The obligation of States to be neutral with regard to the contents that are conveyed within the framework of this right is the result of the presumption of coverage and is also an effect of the need to ensure that, in principle, there are no individuals, groups, ideas or means of expression that are previously excluded from public debate.[16] It is this statement of freedom of expression that results in the prohibition of prior censorship.

---

13 IACHR, Office of the Special Rapporteur for Freedom of Expression. Inter-American Legal Framework on the Right to Freedom of Expression. OEA/Ser.L/V/II IACHR/RELE/INF.2/09, December 30, 2009, para. 18.

14 American Convention on Human Rights, San José, Costa Rica, November 12-22, 1969, Organization of American States.

15 Center for the Study of Law, Justice and Society, Dejusticia. *The right to freedom of expression*, Colombia, 2017, p. 59. This does not mean that this assumption always applies or that it is an absolute right, but rather that the limitations of the right, also as a general rule, must be applied after the manifestation has been expressed.

16 IACHR, Office of the Special Rapporteur for Freedom of Expression. Inter-American Legal Framework on Freedom of Expression. OEA/Ser.L/V/II. IACHR/RELE/INF.2/09. December 30, 2009, para. 30.

Inter-American doctrine has broadly classified specially protected speech into three types of speech:

i. Political speech and speech on matters of public interest

In a democratic society, the importance of public discussion related to the political sphere and matters of general interest narrows the margin of legitimate restrictions on political criticism and demonstrations related to matters of public interest. Both the IACHR and the Inter-American Court have promoted this doctrine by explaining that the exercise of democracy requires the highest possible level of public discussion on the functioning of society and the State in all its aspects, that is, on matters of public interest. Hence, the proper development of democracy requires the widest possible circulation of reports, opinions and ideas on matters of this nature.[17]

The American Convention on Human Rights acknowledges extended protection for this type of expression, something that has been consistently reiterated by its main interpretative body (the Inter-American Court of Human Rights). Expanded protection implies that there are clear obligations for States to strictly refrain from setting limits on the forms of expression, on the one hand, but also to explain that persons participating in public discussion must have a higher threshold of tolerance for criticism.[18]

ii. Speeches on public officials and candidates for public office and other public figures

When the expressions of persons are directed at public officials, private persons voluntarily involved in public affairs or candidates for public office,[19] the formula that leads to acknowledging a higher threshold of tolerance for criticism is repeated,[20]and this means that the obligations of abstention on the part of the State (in terms of restrictions and limitations to the exercise of the right) are also present.

All these groups of subjects (public officials, candidates and individuals involved in public affairs) participate voluntarily under this regime of strong public scrutiny, in which criticism of their actions by the public functions as an accountability mechanism that is part of the broader notion of democratic control. Democratic controls are justified in order to keep the exercise of public

---

17 I/A Court H.R., Case of Kimel v. Argentina. Judgment of May 2, 2008. Series C No. 177, paras. 57 and 87; I/A Court H.R., Case of Claude Reyes et al. v. Chile. Judgment of September 19, 2006. Series C No. 151, paras. 84, 86 and 87; I/A Court H.R., Case of Palamara Iribarne v. Chile. Judgment of November 22, 2005. Series C No. 135, para. 83; I/A Court H.R., Case of Herrera Ulloa v. Costa Rica. Judgment of July 2, 2004. Series C No. 107, para. 127.

18 I/A Court H.R., Case of Palamara Iribarne v. Chile. Judgment of November 22, 2005. Series C No. 135, para. 83; I/A Court H.R., Case of Herrera Ulloa v. Costa Rica. Judgment of July 2, 2004. Series C No. 107, para. 125; IACHR. Arguments before the Inter-American Court in the Case of Herrera Ulloa v. Costa Rica. Transcribed in: I/A Court H.R., Case of Herrera Ulloa v. Costa Rica. Judgment of July 2, 2004. Series C No. 107, para. 101.2.c.

19 I/A Court H.R., Case of Kimel v. Argentina. Judgment of May 2, 2008. Series C No. 177, para. 86; I/A Court H.R., Case of Palamara Iribarne v. Chile. Judgment of November 22, 2005. Series C No. 135, para. 82.

20 I/A Court H.R., Case of Kimel v. Argentina. Judgment of May 2, 2008. Series C No. 177, paras. 86-88; I/A Court H.R., Case of Palamara Iribarne v. Chile. Judgment of November 22, 2005. Series C No. 135, paras. 83-84; I/A Court H.R., Case of "The Last Temptation of Christ" (Olmedo Bustos et al.) v. Chile. et al.) v. Chile. Judgment of February 5, 2001. Series C No. 73, para. 69; I/A Court H.R., Case of Ivcher Bronstein v. Peru. Judgment of February 6, 2001. Series C No. 74, paras. 152 and 155; I/A Court H.R., Ricardo Canese Case. Judgment of August 31, 2004. Series C No. 111, para. 83; I/A Court H.R., Case of Herrera Ulloa v. Costa Rica. Judgment of July 2, 2004. Series C No. 107, paras. 125-129; I/A Court H.R., Case of Claude Reyes et al. Judgment of September 19, 2006, Series C No. 151, para. 8.

power under review, through the obligation of transparency and maximum publicity that mandates all actions of the State.[21]

If people participate in the public sector, it is not only to be expected that they will be subject to strong public scrutiny, but also that they will be subject to a higher level of criticism because their capacity to respond is also greater, whether due to their public outreach, their social influence or their access to the media (as also acknowledged by the Inter-American Court in the Tristán Donoso v. Panama case, to mention just one example).[22]

iii. Speeches expressing essential elements of personal identity or dignity

Freedom of expression is not only a right in itself but also a right that enables other rights and functions as a tool to strengthen the identity and dignity of individuals. Its role as an enabler and necessary condition for rights such as personal identity means that speech related to this type of expression is also protected in a reinforced manner.

An example that has been used repeatedly within the ISHR is that of the rights of indigenous peoples to express themselves and receive information in the language that shapes their identity, since their own language is one of the central elements to be taken into account in the shaping of the identity of individuals and groups, and they can express themselves through it. In addition, for indigenous peoples, the transmission of their culture and worldview, which differentiates them from the rest of the non-indigenous population, also involves the use of language and the shaping of their cultural identity (based on it).[23]

Expression on issues related to sexual diversity, the acknowledgment of sexual or gender identity and the importance of these rights to avoid any type of censorship of expressions related to these issues. The Court, in its Advisory Opinion 24/17, identified cases of indirect censorship when a legal system did not recognize gender identity and when gender expressions that deviated from cisnormativity or heteronormativity standards were punished or censored (even indirectly), since in such cases the message was conveyed that those who fell outside the "traditional" standards did not receive the same consideration and respect or the same legal protection and acknowledgment of their rights.[24]

---

21 IACHR, Office of the Special Rapporteur for Freedom of Expression. Inter-American Legal Framework on Freedom of Expression. OEA/Ser.L/V/II. IACHR/RELE/INF.2/09. December 30, 2009, para. 40.

22 I/A Court H.R., Case of Tristán Donoso v. Panama. Preliminary Objection, Merits, Reparations and Costs. Judgment of January 27, 2009 Series C No. 193, para. 122.

23 I/A Court H.R., Case of López Álvarez v. Honduras. Judgment of February 1, 2006. Series C No. 141. 141. para. 169.

24 I/A Court H.R., Gender identity, and equality and non-discrimination of same-sex couples. Advisory Opinion OC-24/17 of November 25, 2017, Series A No. 24; Inter-American Commission on Human Rights, Observations presented by the Commission on February 14, 2017, para. 49. See, in the same sense, United Nations, Committee on the Rights of the Child, General Comment No. 20 (2016) on the realization of the rights of the child during adolescence, December 6, 2016, CRC/C/GC/20, para. 34, and Office of the United Nations High Commissioner, Living Free & Equals. What States are doing to tackle violence and discrimination against lesbian, gay, bisexual, transgender and intersex people, New York and Geneva, 2016, HR/PUB/16/3, pp. 86-87.

c. Non-protected speeches

At the other extreme of the exercise of freedom of expression, there is another series of expressions that are openly and categorically prohibited and that do not have the coverage given to protected expressions. In these cases, it is possible to take more restrictive measures and even to censor in advance exceptional situations reserved for very specific cases, such as the sexual abuse of minors, also known as "child pornography".

In accordance with international human rights law and specifically the provisions of Article 13 of the ACHR, there are three types of speech that are not protected by freedom of expression.

i. Incitement to violence

The exercise of freedom of expression can be sanctioned when it is carried out in an abusive manner. Both at the doctrinal and jurisprudential levels, there is broad agreement that when the conduct of inciting violence (inciting to commit crimes, to break public order or to affect national security) is carried out, it is acceptable to establish sanctions even in the sphere of criminal law.

Article 13.5 of the ACHR specifically states that: Any propaganda for war and 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."

However, in these cases, it must be clear that there is evidence of a present, certain, objective and compelling character that the alleged incitement conduct was not the mere manifestation of an opinion (however harsh, unfair and disturbing) and that it had not only a clear intent to commit a crime but also the present, real and effective possibility of achieving its objectives.[25]

ii. Direct and public incitement to genocide

As with incitement to violence, direct and public incitement to genocide is prohibited both in international and inter-American standards on freedom of expression, as well as in other specialized international agreements that regulate actions related to the crime of genocide, such as the Convention on the Prevention and Punishment of the Crime of Genocide.[26]

The text of this agreement considers as genocide the killing of members of a group, serious injury to their physical or mental integrity, the intentional infliction of conditions of life calculated to bring about their physical destruction (in whole or in part), measures intended to prevent births within the group, and the forcible transfer of children from the group, when such conduct is carried out with the intent to destroy, in whole or in part, a national, ethnical, racial or

25 I/A Court H.R. Compulsory Membership in an Association of Journalists (arts. 13 and 29 American Convention on Human Rights). Advisory Opinion OC-5/85 of November 13, 1985. Series A No. 5, para. 77; IACHR, Office of the Special Rapporteur for Freedom of Expression. Inter-American Legal Framework on Freedom of Expression. OEA/Ser.L/V/II. IACHR/RELE/INF.2/09. December 30, 2009, para. 58.

26 IACHR, Office of the Special Rapporteur for Freedom of Expression. Inter-American Legal Framework on Freedom of Expression. OEA/Ser.L/V/II. IACHR/RELE/INF.2/09. December 30, 2009, para. 59.

religious group.[27] The conduct that is criminally punishable, in this sense, is the "direct and public instigation to commit [it]."[28]

iii. Sexual abuse of minors (child pornography)

There is an international consensus to firmly prohibit "child pornography". The best interests of children and adolescents are inevitably harmed by a discourse that is violent towards them and violates their rights,[29] which are acknowledged and protected by the State (as well as by society and the family).[30]

"Child pornography" is internationally condemned because it is a form of sexual exploitation and abuse, and incitement and coercion in any act of a sexual nature as well as in any pornographic show or material is punishable.[31]

The Inter-American Human Rights System is very clear: except for these limitations, which are specific and reserved for extremely serious cases of illegitimate exercise of the right of expression, all expressions must be subject to the regime of subsequent responsibilities that privileges and protects their dissemination and establishes a system of later sanctions that may be established in other cases in which the exercise of the right affects the legitimate interests of other persons, but always outside and beyond the censorship regime.

d. Limitations on freedom of expression

Up to this point it is clear that all expressions have a presumption of protection under freedom of expression, except for the three exceptions discussed above. The enhanced protection of freedom of expression is justified because democratic societies uphold the conviction that it is valuable to have as many elements as possible to think, be informed and express the ideas and feelings of individuals in the public sphere. However, no right is absolute and there may be situations under which freedom of expression may be limited; limitations that must follow a series of strict conditions to be considered in conformity with the law.

This means that for a limitation on freedom of expression to be considered "legitimate" or "valid" (depending on whether we make a judgment on its political or legal dimension), it must be subjected to the conditions established by Inter-American law (of the ISHR). The rule on the limits to freedom of expression applies both to state authorities (of all branches of government) and to non-state subjects (whether they are private individuals performing state functions or with public funding, or whether they do so themselves).[32]

---

27 Convention on the Prevention and Punishment of the Crime of Genocide, UNGA, December 9, 1948, Article II.

28 Convention on the Prevention and Punishment of the Crime of Genocide, UNGA, December 9, 1948, Article III, paragraph c).

29 IACHR, Office of the Special Rapporteur for Freedom of Expression. Inter-American Legal Framework on Freedom of Expression. OEA/Ser.L/V/II. CIDH/RELE/INF.2/09. December 30, 2009, para. 59.

30 ACHR. Article 19.

31 Convention on the Rights of the Child, Article 34.

32 Center for the Study of Law, Justice and Society, Dejusticia. The right to freedom of expression, Colombia, 2017, p. 96.

As noted above, the system of protections and limits of this right is based on the prohibition of prior censorship and, as a general rule, the system of subsequent liability operates for all expressions (with the exceptions noted above). The double facet of this right owes its raison d'être to the fact that, in a democratic society, it is acceptable to consider speeches that are valuable for its sustainability and improvement, and that there are others that are considered disvaluable because they mean just the opposite.[33]

For the rule of prohibition of prior censorship, it is not acceptable to establish any requirement, condition or prior authorization of expressions (except in cases of non-protected speech, given that these are essentially vulnerable interests that justify a qualified safeguard); for the rule of subsequent liability, there is a parameter known as the *tripartite test*, through which we can analyze whether a particular limitation to freedom of expression is *valid* or not.

Based on Article 13.2 of the ACHR, the Inter-American system has developed—with the help of the Inter-American Court—a method that serves to know what steps must be followed if a lawful limitation of the right to freedom of expression is sought (and, in the legal aspect, to determine whether such limitation is valid or not, whether it violates the right or not, etc.).[34]

In the first place (step 1), the limitation must be previously embodied in a law (formally and materially), and must be expressly and strictly defined therein in relation to one of the legitimate purposes acknowledged by the ACHR itself (the so-called "democratic objectives").[35] Democratic objectives are delimited by the ACHR and pertain to the rights or reputation of others, the protection of national security, public order, public health or morals. The interpretation of these objectives must always be "democratic" in the sense that these interests should not be taken as being superior to freedom of expression but rather as being articulated with it in order to maximize it and strengthen the democratic system as a whole.[36]

When the rights of others are involved, the proper interpretation must first start from the fact that it is clear that those rights have been harmed or threatened, and then assess the degrees to which this is so against the privileged weight of freedom of expression.[37]

---

33 Salazar Ugarte, Pedro and Gutiérrez Rivas, Rodrigo. *El derecho a la libertad de expresión frente al derecho a la no discriminación*, Mexico, Instituto de investigaciones Jurídicas-UNAM, 2008, p. 28.

34 IACHR, Office of the Special Rapporteur for Freedom of Expression. Inter-American Legal Framework on Freedom of Expression. OEA/Ser.L/V/II. CIDH/RELE/INF.2/09. December 30, 2009, para. 67.

35 IACHR. Arguments before the Inter-American Court in the case of Ricardo Canese v. Paraguay. Transcribed in: I/A Court H.R., Case of Ricardo Canese v. Paraguay. Judgment of August 31, 2004. Series C No. 111, paras. 72. s) to 72.u).

36 Advisory Opinion OC-5/85 of the IACHR Court, Compulsory Membership in an Association of Journalists, November 13, 1985. In addition to this, the IACHR has developed the criterion that limitations must be compatible with the "democratic principle", which implies that they must: i) incorporate the just demands of a democratic society; ii) be compatible with the preservation and development of democratic societies in accordance with Articles 29 and 32 of the American Convention on Human Rights. and development of democratic societies in accordance with Articles 29 and 32 of the ACHR; and iii) be interpreted with reference to the legitimate needs of democratic societies and institutions. In this regard see: IACHR. Annual Report 2009. Report of the Office of the Special Rapporteur for Freedom of Expression. Chapter III (Inter-American Legal Framework of the Right to Freedom of Expression). OEA/Ser.L/V/II. Doc. 51. December 30, 2009. Para. 67.

37 IACHR, Office of the Special Rapporteur for Freedom of Expression. Inter-American Legal Framework on Freedom of Expression. OEA/Ser.L/V/II. IACHR/RELE/INF.2/09. December 30, 2009, para. 77.

The IACHR Court has been very clear in pointing out that it is contradictory "to invoke a restriction on freedom of expression as a means to guarantee it, because it is to ignore the radical and primary character of that right as inherent to each human being individually considered, although it is also an attribute of society as a whole."[38] It is also unacceptable to demand that the exercise of freedom of expression be tied to a condition of truthfulness, because if this were the case, it would open the door to abuses of information controls that would affect the right of access to information of all individuals.[39]

When it comes to the interest in "public order," the Court has treated this principle as "the conditions that ensure the harmonious and normal functioning of institutions on the basis of a coherent system of values and principles",[40] but that, in turn, this principle requires that "The widest possible circulation of news, ideas and opinions, as well as the broadest access to information by society as a whole, are ensured".[41]

When it comes to "national security" the same logic applies. An all-encompassing interpretation of such a limitation is incompatible with democratic societies. On the contrary, modern democracies require that this interest be interpreted in the light of the primary character of freedom of expression and the need to have the most and best information available on public matters of interest to society.

Regarding the Internet, this point has been particularly emphasized in light of the information related to state surveillance programs (and related reports), where the IACHR Rapporteurship for Freedom of Expression has been clear in stating that it is not legitimate to limit this information under the category of national security when private information of dissidents, journalists or human rights defenders is intercepted, captured or used for political purposes or to prevent or compromise their investigations or complaints.[42]

The limitation established in the law must also be clear and precise. All restrictions that do not comply with these requirements imply a violation of the first element of the tripartite test and are considered contrary to the international framework, particularly because they open a wide margin of discretion for the authorities and because they enable a space for arbitrariness that in some cases may lead to disproportionate responsibilities or even to censure.[43]

---

38 I/A Court H.R., Compulsory Membership in an Association of Journalists (Arts. 13 and 29 American Convention on Human Rights). Advisory Opinion OC-5/85 of November 13, 1985, Series A No. 5, para. 7.

39 I/A Court H.R., Compulsory Membership in an Association of Journalists (Arts. 13 and 29 American Convention on Human Rights). Advisory Opinion OC-5/85 of November 13, 1985, Series A No. 5, para. 77.

40 I/A Court H.R., Compulsory Membership in an Association of Journalists (Arts. 13 and 29 American Convention on Human Rights). Advisory Opinion OC-5/85 of November 13, 1985, Series A No. 5, para. 64.

41 I/A Court H.R., Compulsory Membership in an Association of Journalists (Arts. 13 and 29 American Convention on Human Rights). Advisory Opinion OC-5/85 of November 13, 1985, Series A No. 5, para. 69.

42 IACHR, Office of the Special Rapporteur for Freedom of Expression. Freedom of Expression and the Internet. OEA/Ser.L/V/II. IACHR/RELE/INF.11/13, December 31, 2013, para. 60.

43 IACHR, Office of the Special Rapporteur for Freedom of Expression. Inter-American Legal Framework on Freedom of Expression. OEA/Ser.L/V/II. IACHR/RELE/INF.2/09. December 30, 2009, para. 70.

In the case of expressions on the Internet, limitations that have problems of vagueness and ambiguity can also generate a silencing effect that leads to the infringement of the right (due to the uncertainty of what is valid to do and what is not), and it can "have a special impact on this growing universe of people whose incorporation into the public debate is one of the main advantages offered by the Internet as a space for global communication."[44]

Secondly (step 2), limitations must meet three conditions: they must be suitable, necessary and proportional. To say that a limitation is necessary to safeguard a democratic objective implies analyzing whether or not that measure can be achieved in the least restrictive way (because the measure that limits freedom of expression the least should always be chosen). To say that a limitation is suitable means that it effectively solves the problem in question (not one that maintains or aggravates it, for example). To say that a limitation is proportional is to ensure that freedom of expression is not excessively sacrificed in relation to the good being protected; in other words, that there is a proportional relationship between the cost that the limitation implies for the right to freedom of expression and the cost of the right to freedom of expression.[45]

The necessity of the measure must not be equated with an idea of utility or opportunity,[46] but must be a strong necessity that makes it impossible to protect the objective by a less restrictive means, while at the same time, once it is acknowledged that this is so, it must not be limited beyond what is essential (it must be limited to the maximum extent possible).[47]

Adequacy functions as an instrument to evaluate compliance with the purpose of the measure, which must always be limited within the framework of democratic interpretation and in harmony with freedom of expression.[48]

---

44 IACHR, Office of the Special Rapporteur for Freedom of Expression. Freedom of Expression and the Internet. OEA/Ser.L/V/II. IACHR/RELE/INF.11/13, December 31, 2013, para. 58.

45 I/A Court H.R. Case of Herrera Ulloa v. Costa Rica. Preliminary Objections, Merits, Reparations and Costs. Judgment of July 2, 2004. Series C, No. 107, para. 121; Case of Gomes Lund et al ("Guerrilha do Araguaia") v. Brazil. Preliminary Objections, Merits, Reparations and Costs. Judgment of November 24, 2010; and Case of Claude Reyes et al. v. Chile. Merits, Reparations and Costs. Judgment of September 19, 2006. Series C No. 151.

46 I/A Court H.R., Compulsory Membership in an Association of Journalists (Arts. 13 and 29 American Convention on Human Rights). Advisory Opinion OC-5/85 of November 13, 1985, Series A No. 5, para. 46; I/A Court H.R., Case of Herrera Ulloa v. Costa Rica. Judgment of July 2, 2004, Series C No. 107, para. 122; IACHR. Annual Report 1994. Chapter V: Report on the Compatibility between Contempt Laws and the American Convention on Human Rights. Title IV. OEA/Ser. L/V/II.88. doc. 9 rev. 17 February 1995.

47 I/A Court H.R., Case of Kimel v. Argentina. Judgment of May 2, 2008. Series C No. 177, para. 83; I/A Court H.R., Case of Palamara Iribarne v. Chile. Judgment of November 22, 2005. Series C No. 135, para. 85; I/A Court H.R., Case of Herrera Ulloa v. Costa Rica. Judgment of July 2, 2004. Series C No. 107, paras. 121-122; I/A Court H.R., Compulsory Membership in an Association of Journalists (Arts. 13 and 29 American Convention on Human Rights). Advisory Opinion OC-5/85 of November 13, 1985, Series A No. 5, para. 46.

48 I/A Court H.R., Case of Kimel v. Argentina. Judgment of May 2, 2008. Series C No. 177.

Proportionality requires the person who establishes the measure to intervene as little as possible with the exercise of freedom of expression, while verifying that this affectation effectively benefits the protected interest.[49] To find out if this is the case, the person reviewing the measure must evaluate the degree of impairment of the right (serious, medium, low), the importance of the protected interest (high, medium, low) and the cost-benefit of this balance to see if the restriction is justified (in terms of proportionality).[50]

Thirdly (step 3), the assessment of the damage and the measure must always be contextual, which means that the assessment must not be made in general, but must be based on the specific circumstances of the case in question.[51] When analyzing the context, a "public interest test" must also be carried out on the information related to the expression to be limited (and thus know the degree of protection it has). Information related to the State, government management, transparency, accountability of public officials and other information of a similar nature meets this public interest standard, and therefore obtains the enhanced protection mentioned above.[52]

Finally, and for cases involving the right to freedom of expression on the Internet, there is a fourth step that must be fulfilled that is what the IACHR calls the "digital systems perspective", which means that when assessing the validity of a limitation to freedom of expression on the Internet, the impact that such a measure has on the functioning of the Internet in general must be taken into account, particularly in terms of its key characteristics of being a decentralized, free and open network.[53] Limitations on freedom of expression on the internet affect the entire network, no longer just the exercise of the specific right, and it is therefore important to think about the consequences that enabling a limitation could have on the very design of the internet.[54]

---

49 I/A Court H.R., Case of Eduardo Kimel v. Argentina. Judgment of May 2, 2008. Series C No. 177, para. 83; I/A Court H.R., Case of Palamara Iribarne. Judgment of November 22, 2005. Series C No. 135, para. 85; I/A Court H.R., Case of Herrera Ulloa v. Costa Rica. Judgment of July 2, 2004. Series C No. 107, para. 123; I/A Court H.R., Compulsory Membership in an Association of Journalists (Arts. 13 and 29 American Convention on Human Rights). Advisory Opinion OC-5/85 of November 13, 1985. Series A No. 5, para. 46; IACHR. Arguments before the Inter-American Court in the case of Herrera Ulloa v. Costa Rica. Transcribed in: I/A Court H.R., Case of Herrera Ulloa v. Costa Rica. Judgment of July 2, 2004. Series C No. 107, para. 101.1.

50 I/A Court H.R., Case of Kimel v. Argentina. Judgment of May 2, 2008. Series C No. 177, para. 84.

51 I/A Court H.R., Case of Kimel v. Argentina. Judgment of May 2, 2008. Series C No. 177, para. 51; I/A Court H.R., Case of Tristán Donoso v. Panama. Preliminary Objection, Merits, Reparations and Costs. Judgment of January 27, 2009, Series C No. 193, para. 93.

52 I/A Court H.R., Case of Kimel v. Argentina. Judgment of May 2, 2008. Series C No. 177, paras. 57 and 87; I/A Court H.R., Case of Claude Reyes et al. v. Chile. Judgment of September 19, 2006. Series C No. 151, paras. 84, 86 and 87; I/A Court H.R., Case of Palamara Iribarne v. Chile. Judgment of November 22, 2005. Series C No. 135, para. 83; I/A Court H.R., Case of Herrera Ulloa v. Costa Rica. Judgment of July 2, 2004. Series C No. 107, para. 127; I/A Court H.R., Case of Herrera Ulloa v. Costa Rica. Judgment of July 2, 2004. Series C. No. 107, para. 106; IACHR (2009), Inter-American Juridical Framework on Freedom of Expression, OAS, para. 113.

53 IACHR, Office of the Special Rapporteur for Freedom of Expression. Freedom of Expression and the Internet. OEA/Ser.L/V/II. IACHR/RELE/INF.11/13, December 31, 2013, para. 63.

54 Center for the Study of Law, Justice and Society, Dejusticia. The right to freedom of expression, Colombia, 2017, p. 281.

e. Prohibition of prior censorship and indirect restrictions

The ISHR strictly prohibits prior censorship with the exception of public performances in which children and adolescents are protected in accordance with Article 13.4 of the ACHR.[55] The IACHR Court has been emphatic in stating that when there is a measure of prior censorship, the right to freedom of expression is *radically violated*, and democracy in general is affected.[56]

It is important to understand that the rules prohibiting censorship apply to both direct actions and indirect measures, whether they are directed at the means or inputs that a media outlet requires in order to disseminate information or at the way in which information is disseminated on the Internet (e.g. removal of links or moderation of content).[57-58]

Article 13.3 of the ACHR states that this right should not be restricted by "indirect methods or means, such as the abuse of government or private controls over newsprint, radio broadcasting frequencies, or equipment used in the dissemination of information, or by any other means tending to impede the communication and circulation of ideas and opinions." But these measures are not exclusive to others which, nowadays and in the light of new technologies, could also constitute means of indirect censorship. For this reason, the Inter-American Court has expressly stated that this statement is not exhaustive and that those indirect means or channels that could have these effects must be evaluated.[59]

Just as the right of expression can be violated through various means, it is also important to understand that this can happen with different subjects: not only the state can limit freedom of expression, but also private individuals. The ACHR (via Article 13(3)) requires States to protect individuals from control or interference by private parties that results in limiting the right to freely express.[60]

States parties to the ACHR may therefore also be liable for violating the Convention when they allow, encourage or fail to act against measures by private parties that violate freedom of expression.[61] At this intersection (of public and private), the obligation of non-discrimination that all measures must respect is particularly relevant. It is not valid for any limitation to foster

---

55 I/A Court H.R., Case of Kimel v. Argentina. Judgment of May 2, 2008. Series C No. 177, para. 54; I/A Court H.R., Case of Palamara Iribarne v. Chile. Judgment of November 22, 2005. Series C No. 135, para. 79; I/A Court H.R., Case of Herrera Ulloa v. Costa Rica. Judgment of July 2, 2004. Series C No. 107, para. 120.

56 I/A Court H.R., Case of Palamara Iribarne v. Chile. Judgment of November 22, 2005. Series C No. 135, para. 68; I/A Court H.R., Compulsory Membership in an Association of Journalists (Arts. 13 and 29 American Convention on Human Rights). Advisory Opinion OC-5/85 of November 13, 1985. Series A No. 5, para. 54.

57 I/A Court H.R., Case of "The Last Temptation of Christ" (Olmedo Bustos et al.) v. Chile. Judgment of February 5, 2001. Series C No. 73; IACHR, Office of the Special Rapporteur for Freedom of Expression. Inter-American Legal Framework on Freedom of Expression. OEA/Ser.L/V/II. IACHR/RELE/INF.2/09. December 30, 2009, para. 147.

58 I/A Court H.R., Case of Ivcher Bronstein v. Peru. Judgment of February 6, 2001. Series C No. 74, paras. 158-163.

59 I/A Court H.R., Case of Ríos et al v. Venezuela. Preliminary Objections, Merits, Reparations and Costs. Judgment of January 28, 2009. Series C No. 194, para. 340; I/A Court H.R., Case of Perozo et al. v. Venezuela. Preliminary Objections, Merits, Reparations and Costs. Judgment of January 28, 2009. Series C No. 195, para. 367.

60 I/A Court H.R., Case of Perozo et al. v. Venezuela. Preliminary Objections, Merits, Reparations and Costs. Judgment of January 28, 2009. Series C No. 195, para. 367; I/A Court H.R., Case of Ríos et al. v. Venezuela. Preliminary Objections, Merits, Reparations and Costs. Judgment of January 28, 2009. Series C No. 194, para. 240.

61 I/A Court H.R., Compulsory Membership in an Association of Journalists (Arts. 13 and 29 American Convention on Human Rights). Advisory Opinion OC-5/85 of November 13, 1985. Series A No. 5, para. 48.

or perpetuate prejudice or intolerance towards vulnerable groups, whether such measures are established by private individuals or by the state.[62]

f. Public officials and freedom of expression

The IACHR has pointed out that in addition to administrative and legislative measures that may violate the right to freedom of expression, States may impact on this right through public statements made by their public officials.

If public speech increases or results in the vulnerability of a group or individual, freedom of expression is violated: if a government speaks out in the media in a way that intimidates or restricts the ability to exercise the right, it creates a situation of risk (which it should prevent in the first place) to the right of that group or individual.[63] The Court has made an assessment of the power and inequality in which some groups are situated to indicate that public officials should refrain from speech that increases the "relative vulnerability" of groups at risk.[64] In addition, States have a number of duties that they must take into account when making declarations, such as:[65]

- Duty to make statements in certain cases, in the performance of their legal and constitutional duties, regarding matters of public interest.
- Special duty to reasonably verify the facts on which their statements are based.
- Duty to ensure that their statements do not amount to human rights violations.
- Duty to ensure that their statements do not constitute arbitrary interference—direct or indirect—with the rights of those who contribute to the public discourse through the expression and distribution of their thoughts.
- Duty to ensure that their statements do not interfere with the independence and autonomy of judicial authorities.

62 IACHR. Annual Report 1994. Chapter V: Report on the Compatibility between the Disaccomplishment Laws and the American Convention on Human Rights. Title III. OEA/Ser. L/V/II.88. doc. 9 rev. 17 February 1995.

63 I/A Court H.R., Case of Perozo et al. v. Venezuela. Preliminary Objections, Merits, Reparations and Costs. Judgment of January 28, 2009. Series C No. 195, para. 161; I/A Court H.R., Case of Ríos et al v. Venezuela. Preliminary Objections, Merits, Reparations and Costs. Judgment of January 28, 2009. Series C No. 194, para. 149.

64 I/A Court H.R., Case of Ríos et al. v. Venezuela. Preliminary Objections, Merits, Reparations and Costs. Judgment of January 28, 2009. Series C No. 194, para. 145; I/A Court H.R., Case of Perozo et al. v. Venezuela. Preliminary Objections, Merits, Reparations and Costs. Judgment of January 28, 2009. Series C No. 195, para. 157.

65 Cf. in IACHR, Office of the Special Rapporteur for Freedom of Expression. Inter-American Legal Framework on Freedom of Expression. OEA/Ser.L/V/II. IACHR/RELE/INF.2/09. December 30, 2009, paras. 201-205.

g. Freedom of expression and the Internet

The intrinsic characteristics of the Internet have made it a true facilitating and enabling tool for other rights, which is why it has been recognized in some countries as an autonomously shaped human right.[66]

In 2011, the United Nations (UN) Special Rapporteur on Freedom of Opinion and Expression, the Representative on Freedom of the Media of the Organization for Security and Cooperation in Europe (OSCE), the Organization of American States (OAS) Special Rapporteur on Freedom of Expression and the Special Rapporteur on Freedom of Expression and Access to Information of the African Commission on Human and Peoples' Rights (ACHPR), the Organization of American States (OAS) Special Rapporteur on Freedom of Expression and the Special Rapporteur on Freedom of Expression and Access to Information of the African Commission on Human and Peoples' Rights (ACHPR), drafted a joint statement on Freedom of Expression and the Internet in which they emphasized in item 1 c) that:

> "Approaches to regulation developed for other means of communication—such as telephony or broadcasting—cannot simply be transferred to the Internet but, rather, need to be specifically designed for it."[67]

The logic behind this rule is that it is not possible to handle the internet in the same way as other media because it is a form of communication with particularities that require specific treatment to keep it a free and open space. The architecture of the internet has, for example, the element of net neutrality, which has been broadly defined as facilitating "access to and dissemination of content, applications and services freely and without distinction. At the same time, the absence of disproportionate entry barriers to offer new services and applications on the Internet is a clear incentive for creativity, innovation and competition.[68]

In the same line, the IACHR in its report on Freedom of Expression and the Internet reiterated that:

> "The Internet has been developed using design principles which have fostered and allowed an online environment that is decentralized, open and neutral. It is important for all regulation [...] to maintain the basic characteristics of the original environment, strengthening the Internet's democratizing capacity and fostering universal and non-discriminatory access".[69]

---

66 For example, Mexico, with the right of access to information technologies in Article 6 of the Mexican Constitution.

67 UN. OSCE. OAS. ACHPR. Joint Declaration on Freedom of Expression and the Internet. 2011. Available at: https://www.oas.org/en/iachr/expression/showarticle.asp?artID=849&lID=1

68 IACHR, Office of the Special Rapporteur for Freedom of Expression. Freedom of Expression and the Internet. OEA/Ser.L/V/II. IACHR/RELE/INF.11/13, December 31, 2013, para. 27; Council of Europe. Committee of Ministers. Declaration of the Committee of Ministers on network neutrality. 29 September 2010. Point 3; Belli, Luca. Council of Europe Multi-Stakeholder Dialogue on Network Neutrality and Human Rights. Outcome Paper. CDMSI (2013) Misc 18. 3-6 December 2013. Para. 16-17.

69 IACHR, Office of the Special Rapporteur for Freedom of Expression. Freedom of Expression and the Internet. OEA/Ser.L/V/II. IACHR/RELE/INF.11/13, December 31, 2013, para. 11.

In addition:

> "Any measures which could, in one way or another, affect the access to and use of the Internet must be interpreted according to the primacy of the right to freedom of expression, at all times, especially in regard to speech that is protected pursuant to the terms of Article 13 of the American Convention."[70]

Thus, the protection of the principle of net neutrality is essential to ensure the plurality and diversity of information flowing through the internet. It is also clear that the architecture of this "network of networks" is essential to balance and enhance a democratic public debate, which is inclusive and pluralistic in nature (with "information pluralism" at its core)."[71]

In this sense, all legislation related to freedom of expression and the internet must therefore take into account its characteristics and consider that the controls and measures that constitute limits to freedom of expression and free access to the internet must comply with the standards established by the ISHR that were previously developed, in addition to those resulting from the structural specificities of the internet.

70 Ibid. para. 14.

71 I/A Court H.R., Case of Kimel v. Argentina. Case of Kimel v. Argentina. Merits, Reparations and Costs. Judgment of May 2, 2008. Series C No. 177. para. 57; I/A Court H.R., Case of Fontevecchia and D`Amico v. Argentina. Case of Fontevecchia and D`Amico v. Argentina. Merits, Reparations and Costs. Judgment of November 29, 2011. Series C No. 238. para. 45.

II. The liability of Internet intermediaries for expressions of third parties

This section will address the role played by (primarily private) intermediaries on the Internet and the way in which they handle third-party expressions that are published, hosted, streamed or linked through such services.

The discussions on the liability of intermediaries have increased in Latin America in recent years, based on various legislative proposals in the region, the introduction of free trade treaties and European regulatory frameworks on the subject. This discussion is essential because of its centrality for the exercise of rights such as freedom of expression, but also for a broader set of rights that are exercised on the internet (strongly related to the development of different online services).[72]

a. The role of intermediaries on the Internet

Intermediaries are private actors that provide a range of services such as access and interconnection; transmission, processing and routing of traffic; hosting and accessing material published by third parties; referencing content or searching for material on the network; carrying out financial transactions; and connecting users through social networking platforms (among others).[73] The IACHR points out that although there are different ways of classifying them, the most relevant ones are:[74]

- Internet Service Providers (ISPs)
- Web hosting providers
- Social networking platforms
- Search engines

To a large extent, intermediaries are responsible for driving the social impact of freedom of expression, which is why their actions often want to be controlled through the imposition of accountability on them and the position they occupy and the role they play. These intermediaries have emerged as the points through which it is (technically) possible to exercise control over content on the internet.[75]

---

72 Del Campo, Agustina; Schatzky, Morena; Hernández, Laura; Lara, Juan Carlos. Looking South. Towards new regional consensus on Internet intermediary liability, Al Sur, April 2021, pp. 4-8.

73 United Nations. General Assembly. Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue. A/HRC/17/27. 16 May 2011. Para. 38. Available for reference at: http://ap.ohchr.org/documents/dpa-ge_s.aspx?m=8

74 OEA/Ser.L/V/II. IACHR/RELE/INF.11/13, December 31, 2013, para. 91.

75 United Nations. General Assembly. Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue. A/HRC/17/27. 16 May 2011. Para. 74. Available for consultation at: http://ap.ohchr.org/documents/dpa-ge_s.aspx?m=8

b. The principle of non-liability of intermediaries

The importance of preventing violations of the right of expression in its individual and—particularly—social dimension has been recognized through the Joint Declaration on Freedom of Expression and the Internet, established by the United Nations (UN) Special Rapporteur on Freedom of Opinion and Expression, the Organization for Security and Cooperation in Europe (OSCE) 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 (ACHPR) Special Rapporteur on Freedom of Expression and Access to Information:[76]

> "2. Intermediary Liability
>
> a. No one who simply provides technical Internet services such as providing access, or searching for, or transmission or caching of information, should be liable for content generated by others, which is disseminated using those services, as long as they do not specifically intervene in that content or refuse to obey a court order to remove that content, where they have the capacity to do so ('mere conduit principle').
>
> b. Consideration should be given to insulating fully other intermediaries, including those mentioned in the preamble, from liability for content generated by others under the same conditions as in paragraph 2(a). At a minimum, intermediaries should not be required to monitor user-generated content and should not be subject to extrajudicial content takedown rules which fail to provide sufficient protection for freedom of expression (which is the case with many of the 'notice and takedown' rules currently being applied)."

The Joint Declaration's emphasis on the role of non-liability of intermediaries is not a coincidence but reflects the fact that their privileged place to exercise control over content circulating on the Internet makes them a target often sought by governments to interfere with the flow of information. The pressure results from the fact that it is easier to identify and control these actors than those directly responsible for the expression they seek to inhibit or control.[77]

However, the declaration reflects the international consensus that exists on the rejection of strict liability models for intermediaries (which implies holding intermediaries liable for illegitimate or illegal content generated by third parties).[78] One of the main reasons for this consensus is the difficulties of reviewing all content circulating on, for example, an intermediary's platform, and the burden of presuming that avoiding potential harm to a third party is an action that is within the limited control that intermediaries actually possess. Consensus supports that in-

76 UN. OSCE. OAS. ACHPR. Joint Declaration on Freedom of Expression and the Internet. 2011. Available at: https://www.oas.org/en/iachr/expression/showarticle.asp?artID=849&lID=1

77 IACHR, Office of the Special Rapporteur for Freedom of Expression. Freedom of Expression and the Internet. OEA/Ser.L/V/II. IACHR/RELE/INF.11/13, December 31, 2013, para. 93.

78 IACHR, Office of the Special Rapporteur for Freedom of Expression. Freedom of expression and the Internet. OEA/Ser.L/V/II. IACHR/RELE/INF.11/13, December 31, 2013, para. 95.

termediaries should not be legally bound by obligations to monitor user-generated content in order to stop and filter unlawful speech.[79]

The IACHR provides a useful analogy to explain the anti-democratic impact of holding intermediaries objectively responsible for the circulation of information generated by third parties: holding an intermediary responsible in this sense, in the context of an open, pluralistic, universally accessible and expansive network, would be like holding telephone companies responsible for the telephone threats that one person makes to another, causing uncertainty or other types of damage.[80]

The UN Special Rapporteur on freedom of expression argued, similarly, that holding intermediaries responsible for content disseminated or created by their users seriously undermines the right to freedom of expression because it results in a form of private censorship,[81] that arises as a self-protective response by intermediaries (to avoid being sanctioned) that is excessively broad, lacking in transparency and due process.[82]

For these reasons, the IACHR holds that liability of intermediaries for expressions of a third party that are unlawful should only proceed when ordered by a judicial authority that operates with sufficient guarantees of independence, autonomy and impartiality, and that is capable of assessing the rights at stake in order to provide the necessary guarantees to the user (whereby resolutions or recommendations of mechanisms or bodies of an administrative nature would be excluded in principle).[83]

---

79 IACHR, Office of the Special Rapporteur for Freedom of Expression. Freedom of Expression and the Internet. OEA/Ser.L/V/II. CIDH/RELE/INF.11 /13, December 31, 2013, para. 96.; United Nations (UN) Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Representative on Freedom of the Media of the Organization for Security and Cooperation in Europe (OSCE), Organization of American States (OAS) Special Rapporteur on Freedom of Expression, and Special Rapporteur on Freedom of Expression and Access to Information of the African Commission on Human and Peoples' Rights (ACHPR). June 1, 2011. Joint Declaration on Freedom of Expression and the Internet. Point 2 (b); Court of Justice of the European Union. Scarlet Extended SA v. Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM). C-70/10. Judgment of November 24, 2011. Paras. 49-53; Court of Justice of the European Union. Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v. Netlog NV. C-360/10. Judgment of February 16, 2012. Paras. 47- 51.

80 IACHR, Office of the Special Rapporteur for Freedom of Expression. Freedom of Expression and the Internet. OEA/Ser.L/V/II. IACHR/RELE/INF.11/13, December 31, 2013, para. 97.

81 There is a certain degree of agreement that initiatives that have been presented in recent years in different countries (particularly in Europe) are highly problematic for the concerns outlined here and thinking within the inter-American human rights framework. Thus, a recent analysis points out that: "Regulatory initiatives in recent years have been criticized mainly for their adverse effects on human rights, with particular attention to the right to freedom of expression. to freedom of expression. Threats of liability, with significant monetary fines—or worse, imprisonment, as in the case of the Australian law—coupled with the obligation to resolve in extremely short periods of time, create an incentive for excessive removal of content known as 'private censorship'. of content known as "private censorship". In the face of these pressures, the fear is that the platforms will remove content that is allegedly or manifestly illegal and, in many cases, completely legal, violating the protection of the right to freedom of expression recognized in international instruments. international instruments. Del Campo, Agustina; Schatzky, Morena; Hernández, Laura; Lara, Juan Carlos. Mirando al Sur. Towards new regional consensus on Internet intermediary liability and content moderation on the Internet, Al Sur, April 2021, p. 30.

82 UN. UN Special Rapporteur on Freedom of Expression, A/HRC/17/27, para. 40.

83 IACHR, Office of the Special Rapporteur for Freedom of Expression. Freedom of Expression and the Internet. OEA/Ser.L/V/II. IACHR/RELE/INF.11/13, 31 December 2013, para. 106.

c. Section 230 of the United States Communications Decency Act of the United States of America

The principle of non-liability of intermediaries has its origins in the legislation of the United States of America (USA). As a result of the influence that US regulation has had on the development of the Internet, one of the most relevant legal instruments concerning intermediary liability is Section 230 of the Communications Decency Act, which was added to the US Telecommunications Act.[84]

This section, on the one hand, recognizes the principle of non-liability of intermediaries for the expressions of their users. However, it also addresses the absence of liability for content moderation actions voluntarily taken by intermediaries ("good Samaritan" rule, as explained below):

1. "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.

2. No provider or user of an interactive computer service shall be held liable on account of:

   a. any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected; or

   b. any action taken to enable or make available to information content providers or others the technical means to restrict access to material described in paragraph (1).[85]

i. The Origin of Section 230

The First Amendment to the US Constitution has extensive jurisprudence concerning the difference between the party distributing speech (such as a television station, a printing press or a radio program) and the third party broadcasting such speech.[86] According to case law, a distributor is not legally liable for what a third party has said as long as he did not know or should not have known about the content giving rise to such liability.[87]

---

84 This provision articulates, of course, with the First Amendment of the US Constitution, which has a central value in the US constitutional scheme and has sometimes even come to be regarded as "absolute". The scope of the First Amendment is in principle for the State, but it certainly also includes companies and, in the case at hand, Internet intermediaries. Section 230 establishes the moderation regime of content on the Internet in a manner articulated with the First Amendment since the early 1990s. See: Del Campo, Agustina; Schatzky, Morena; Hernández, Laura; Lara, Juan Car- los. Mirando al Sur. *Towards new regional consensus on Internet intermediary liabilit*, Al Sur, April 2021, p. 17.

85 Available at: https://www.law.cornell.edu/uscode/text/47/230#f_3.

86 The enhanced protection of section 230 has led to the academic community to consider that the standard established therein exceeds the scope of the First Amendment, becoming a law or statute that enhances freedom of expression (speech-enhancing statute), because it reaches "not only defamatory content but any complaint based on the content of third parties". Del Campo, Agustina; Schatzky, Morena; Hernández, Laura; Lara, Juan Carlos. Mirando al Sur. Towards new regional consensus on Internet intermediary liability, Al Sur, April 2021, p. 18, citing: Goldman, Eric, "Why section 230 is Better Than The First Amendment," Notre Dame Law Review Reflection, 2019.

87 Cfr. Koseff, Jeff. "The Twenty-Six Words That Created The Internet". Cornell University Press. New York. 2019. p. 11-35.

This doctrine emerged and was, of course, conceived at a time when the internet did not exist and discussions about the right to freedom of expression focused on the role of the mass media. As such, it encountered difficulties when users began to use the internet. The creation of Section 230 of the CDA is largely explained by two Supreme Court cases: *Cubby Inc v CompuServe and Stratton Oakmont Inc v Prodigy Services Co.*

The *Cubby* case concerns a situation in which an alleged slander was made towards the Cubby company, by means of a newsletter called "Rumorville", on an internet forum called Compuserve. *Compuserve* had no "editorial control" over the content before it was published. The Court decided that since *Compuserve* did not actively review its site, the nature of the platform was that of a distributor and not a publisher, and therefore it should not be subject to such liability.[88]

However, in the Stratton case, where an individual posted "defamatory" comments against a well-known brokerage firm on a forum of the company *Prodigy*, the Court controversially decided that the company was legally responsible for acting as an editor of such forums. The Court argued that the company's work should be considered as editorial work since the company itself acknowledged that it moderated its own content and actively deleted some posts on its forum.

The latter decision was highly controversial in the media and was criticized by a number of experts who were concerned about the incident, as in their view the decision opened the door to arbitrariness in determining the legal liability of platforms. Many people considered that Prodigy's moderation did not involve editorial work and that the main problem with the court decision was that, given the ambiguity it created, it constituted a precedent that would encourage intermediaries to moderate and reduce malicious content, so as not to run the risk of being charged with legal liability.[89]

The concerns that were raised by the Supreme Court decisions translated into a legislative discussion in the US Congress. The congressmen who pushed for this regulation were looking for a way to generate the appropriate incentives to moderate content and achieve industry development. In analyzing the legal dilemma, they concluded that over-regulation of intermediaries severely obstructed the creation and development of new online services.[90]

After discussion, a paragraph was established which sets out, in section 230 c, two main points:

1. The legal liability for third-party content published on interactive platforms lies with the provider of this information and not with the platform.

2. Unilateral moderation measures carried out by platforms in "good faith" are permissible and do not give rise to legal liability.

88 Cubby, Inc. v. CompuServe, Inc. 776 F. Supp. 135 (1991). Available at: https://law.justia.com/cases/federal/district-courts/FSupp/776/135/2340509/.

89 Kosseff Jeff. "The Twenty-Six Words...", op. cit., p. 53.

90 Id. p.60

What is the purpose of both principles? To create positive incentives for intermediaries. What Congress aimed to do in adding these paragraphs was not to discourage the expansion of internet intermediaries, in particular by providing incentives for those who can block violent or offensive content to do so under the "good Samaritan" principle outlined above.[91] Subsequently, in court, the Supreme Court recognized that the two most important points of this provision were: i) the protection of bona fide measures removing "indecent" content and ii) the protection of freedom of expression.[92]

ii. The principle of non-liability of intermediaries in section 230

As mentioned above, the distinction between the subject publisher and the subject author or sender of a message had been one of the most important issues for jurisprudence related to the First Amendment of the US Constitution, which is daily considered as the cornerstone of freedom of expression in this country. The Zeran case maintained this approach by stating that platforms would not be legally liable even if they had knowledge of a third party's production of objectionable knowledge.[93]

The logic behind the ruling was that only in this way would intermediaries not have an incentive to remove any type of message that was notified to them or that was found circumstantially. Otherwise, as would be the case with a strict liability system, platforms would be under pressure to set up surveillance systems that would result in content censorship, even if this was "collateral".

Collateral censorship occurs when a private entity can control the speech of its users through moderation systems.[94] If regulation allows intermediaries to be held legally responsible for third-party content, these companies take an active role in censoring any expression that might carry the slightest risk of a lawsuit.[95] In other words, faced with the risk of liability and doubt about certain expressions, corporations only gain incentives to monitor and control, and not to protect and guarantee freedom of expression.

Precisely in this type of situation, section 230 protects users' freedom of expression by preventing such perverse incentives. This view was shared by the Supreme Court in the case of Reno v. ACLU, which argued the constitutionality of several provisions of the Communications Decency Act regarding the protection of minors from "pornographic or indecent" content. The Supreme Court established that the internet is a radically different medium from traditional mass media, such as radio or television, and that the difference lies in the fact that while the latter limits entry to certain content creators, the internet allows any user to publish and share content online.

---

91 Klonick, Kate. "The new governors: The people, rules, and processes governing online speech". *Harvard Law Review*, 131, 2018, p. 1605.

92 Id. p. 1608, in *Zeran v. America*.

93 Kosseff Jeff. "The Twenty-Six Words..." op. cit., p. 86-102.

94 See Michael Meyerson. Meyerson, Michael I. "Authors, Editors, and Uncommon Carriers: Identifying the 'Speaker' Within the New Media" (1995). Notre Dame Law Review, Vol. 71, No. 1, p. 79, 1995, Available at SSRN: https://ssrn.com/abstract=1327090.

95 Idem.

Given that the Internet has a free and open architecture, which is fundamentally different from that of traditional media (restricted and limited), the equalizing and freedom of expression maximizing potential lies in that architecture staying that way. In this case, Justice Stevens recognized that for the internet to maintain its growth (in tune with its free speech-enhancing effect), minimal intervention in speech in general was necessary, so he held that the clauses and penalties relating to the concept of "indecency" were unconstitutional, and then left the rest of section 230 intact.

Section 230 in its current state is strongly focused on facilitating the economic and technological rise of US companies. In doing so, it gives a wide margin that is in the interest of the right to freedom of expression because it seeks to avoid censorship of lawful expression, which could be carried out by intermediaries seeking to avoid unnecessary risk (as opposed to a different system of strict liability).[96] For these reasons, the first part of section 230 embodies one of the main foundations of the modern Internet, which is essential for the exercise of freedom of expression online, regardless of all the related debates regarding the obligations that some, mainly private, intermediaries have towards freedom of expression and other rights that are exercised in the digital environment.

iii. The non-liability for unilateral measures with respect to content moderation

Aiming to favor the development of the Internet, as well as to encourage private platforms to take action to remove abusive expressions online, in addition to the rule of non-liability of intermediaries for third-party expressions, Section 230 incorporates a provision related to unilateral content moderation measures. In this case, Section 230(2) excludes from legal liability unilateral moderation measures that are in good faith and on content that may be considered objectionable or undesirable (as mentioned above, the "good Samaritan" rule).

The freedom to manage third-party-generated content granted by the "good Samaritan" rule allows intermediaries the flexibility to ensure that their service or platform is suitable and attractive to the majority of people, that it is economically viable, especially when the intermediary has a business model based on advertising, and in general, to remove content deemed harmful without fear of legal repercussions.

In this way, the "good Samaritan" rule incorporated in Section 230 presumes to generate benefits for the public interest and for intermediaries by seeking to strike a balance between the lack of incentives for censorship guaranteed by the non-liability of intermediaries for content generated by third parties and the presence of incentives to act against "abusive" content published, hosted or linked on the platforms or services offered by such intermediaries.

In this way, Section 230 ensures that intermediaries cannot be legally considered as "publishers", and be subject to liability as such, for the simple fact of moderating content (as long as such moderation is carried out in good faith).[97]

---

96 Keller, Daphne. "El "derecho al olvido" de Europa en América Latina", in Del Campo, Agustina (coord.) Hacia una Internet libre de censura II: perspectivas en América Latina. University of Palermo, School of Law, Center for Studies on Freedom of Expression and Access to Information, Buenos Aires, 2017, p. 180. Available at:: https://www.palermo.edu/cele/pdf/investigaciones/Hacia_una_internet_libre_de_censura_II.pdf

97 Gillespie, Tarleton. "Custodians of the Internet". Yale University Press. United States. 2018. p.30-31

d. The principle of non-responsibility of intermediaries in commercial treaties

The principle of no intermediary liability has started to be included in trade agreements such as the Mexico-United States-Canada treaty, known as T-MEC, which replaces the North American Free Trade Agreement (NAFTA), and which, unlike NAFTA, includes a series of provisions and conditions to be met by the countries party to the agreement in its chapter 19 on "Digital Commerce".[98]

Paragraph 17 of that chapter acknowledges, in similar terms to Section 230 described above, the principle of non-liability of intermediaries for content generated by third parties that these platforms host or process, in the following terms:

> No Party shall adopt or maintain measures that treat a provider or user of an interactive computer service as an information content provider to determine liability for damages related to information stored, processed, transmitted, distributed or made available by the service, except to the extent that the provider or user, in whole or in part, created or developed the information.

Free trade agreements have added interests and provisions that are in tension not only with local regulations in different countries, but also with Inter-American standards. In part because of this, both civil society and academia specializing in human rights in the digital environment have conducted campaigns and used judicial means to challenge the implementation of some of the provisions of trade agreements that jeopardize freedom of expression on the internet and that relate to the principle of non-accountability of intermediaries.[99]

98 The lack of regulation in different Latin American countries means that one of the regulatory frameworks for intermediaries is brought in "from outside" with the provisions on intellectual property and digital commerce in free trade agreements. In the case of treaties with the United States of America, the most common is to find regulations that refer to the Digital Millennium Copyright Act (DMCA) of this country, which includes the figure of notice and takedown. The truth is that although there are differences between the way in which the different countries in Latin America include these provisions and there is no uniformity, many of these rules coincide. For a complete analysis of this point in most of the region, see: Del Campo, Agustina, et. al. Mirando al. To see how the mechanism established in the DMCA functions as an exception to section 230 in terms of copyright law (which has been heavily criticized even with its limited scope), see specifically pp. 19-20.

99 Del Campo, Agustina, et. al. "Mirando al Sur...". op. cit., pp. 10-11.

III. Content moderation

Content moderation, as previously discussed, defined as the "organized practice of reviewing user-generated content posted on websites, social networks or other platforms",[100]has been considered essential to the functionality of Internet platforms. For experts such as Tarleton Gillespie, not only would a platform be dysfunctional without moderation, but moderation is an indispensable element of a platform's existence.[101]

However, as will be explained in this section, content moderation in practice poses complex challenges that have raised questions about the assumptions made by provisions such as Section 230 explained above.

An example of this is the goal of maintaining the flow of information on the Internet. How should the need to exclude harmful content and ensure the fulfilment of this goal be understood?[102] In many cases, the tension between the two is inevitable and leads to conflicts or dilemmas in moderation, which can only be avoided by balancing the interests at stake.

When content moderation is examined this way, it is easier to understand the two most extreme (and opposite) positions in the debate about the "what and how" of moderation: those who believe that platforms have been too permissive of harmful content (such as child pornography, online harassment or hate speech that incites violence) and those who say that platforms have overstepped their powers and intervened too much in the public discussion.[103]

To understand them better, it is necessary to understand the incentives that platforms have when moderating, the moderation rules, the moderation procedures that are in place and the consequences of content moderation on freedom of expression.

a. Objectives and justifications for content moderation

All platforms moderate. Although many of them avoid moderation being very noticeable, it is inevitable. For others, due to their characteristics or the service they provide, the moderation system is one of the main features of their business.

Some of the main reasons for moderation are:

- Corporate image:

  While it is true that many platforms create moderation rules and systems as a matter of social responsibility, it is undeniable that many others do so in order to fit their corporate identity.[104]

---

100 Roberts, Sarah. "Behind The Screen: Content Moderation in the Shadows of Social Media". Yale University Press. United States. 2019. P. 33.

101 Id. p. 21.

102 Gillespie, Tarleton. "Custodians of the...". op. cit. p. 10.

103 Idem, p. 11.

104 Klonick, Kate. "The New Governors..." op. cit., p. 1625.

Not every platform is an open forum for decentralized discussion on any topic. For example, while Facebook's mission is to make the world more connected and open, to create a space where friends and family can live together, find communities and grow businesses,[105] Reddit aims to host small communities for any topic that its users are passionate about.

- Economic reasons:

  Economic rewards are perhaps the most important reason for the way many platforms operate. Their main objective serves a business model that seeks to maximize the profits possible from making the user stay on their platform and thus increase their advertising revenues.[106]

  One case that illustrates this logic is Twitter's policy against hate speech that incites violence (which is also seen as being in line with a social interest in strengthening public debate), which was generated by the negative reaction of users to the threats, targeted harassment and violence against feminists and journalists in the controversial Gamergate case.[107]

b. Moderation rules

In order to carry out effective moderation, platforms must set clear rules and guiding principles that enable their users to understand the community standards by which the platform will generate its communication environment, as well as to organize the moderation systems and the people employed to moderate content. These are the rules of the game by which a platform relates to the public that participates in it and should therefore be public and clear to all participants.

Publicity and clarity of moderation rules are essential for intermediaries to operate within a transparent framework in which they can be held accountable. They are also essential to limit the arbitrariness with which such platforms can (and often do) act.

There are two key texts for platforms that involve the publication of third-party content and that enable public scrutiny of their actions: community standards and terms and conditions.

The terms and conditions are a contract stating the obligations of the user and the platform; they outline methods of dispute resolution, what content is appropriate and also refer to legal responsibilities, intellectual property or any reference that may avoid litigation.[108]

Community norms are usually set out as user-driven documents that are written in clear language so that they can be understood by all users. They specify expected and unacceptable behaviors within the social network;[109] their "values", their vision and the kind of interactions they encourage in their spaces with their users are more specifically detailed.

---

105 Meta. "Company Info". Accessed December 6, 2021. Available at: https://about.fb.com/company-info/.

106 Klonick, Kate. "The New Governors..." op. cit., p. 1627.

107 107 Idem, p. 1629.

108 Gillespie, Tarleton. "Custodians of the...", op. cit. p. 46.

109 Idem.

Although each platform is different, normally minimal prohibitions are set for content such as spam, explicit pornography, hate speech that incites violence, harassment or illegal content in accordance with the regulations of the access point. The risks and consequences of poorly defining these bans will be addressed below.

Also some platforms, for example Reddit, require a minimum quality of content to fit the purpose of each section (subreddit), their internal rules and the specific functionality of that site.[110] Other platforms such as Wikipedia require that posts comply with requirements such as neutrality, quality of sources, avoidance of plagiarism and encyclopedic relevance.[111]

Content moderation rules play a very important role, because they act as a source for assessing what the platforms are committed to do (the rules they are required to follow), on the one hand, and also because they make it possible to assess whether or not they meet certain minimum standards, such as transparency and the right to freedom of expression.

c. Moderation Procedures

Content moderation takes place at several stages:

1. ***Ex Ante* content moderation:** human moderators or automated systems review the content before it is published.[112]

This method is also known for the automatic upload filters that some platforms such as Twitter or Facebook have in place to check mainly that what is going to be published is not child pornography or involves the "improper" use of copyrighted material.

2. ***Ex Post* content moderation:** the review of content is done after it has been disseminated.

After the content has been posted, content can be reviewed by moderators in case another user is *flagging*,[113] or a human moderator, automated system or third party makes a report of such content.[114]Some platforms like Facebook have a system designed to filter out reports made by users.

Post-moderation can be carried out in different ways:

- **Reactive moderation:** the moderator reviews content passively or moderates only until requested to do so by a third party.

110 Ibid, p. 64.

111 Wikipedia. Policies and Conventions. Accessed October 15, 2021, Available at.: https://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines

112 Klonick, Kate. "The New Governors…", op. cit., p. 91.

113 Mechanism by which platform users themselves can report content they consider inappropriate for review by a moderator.

114 Roberts, Sarah. "Behind The Screen.". op.cit., p. 33.

Reactive moderation is the most common method used by many platforms for content moderation.[115]Some platforms use flagging systems so that it is the users themselves who, as a community, report a post that they consider inappropriate.

Most complaints are filtered at an early stage to avoid unnecessary workload for the various levels of moderation. For example, on Twitter or Facebook the user is asked to categorize their complaint to see if it would be appropriate in the first place or if it was just the user's personal dislike.

- **Proactive moderation:** moderators proactively search for content that does not comply with the terms and conditions.[116]

Active moderation is the method normally used to deal with certain speech that is not protected by the right to freedom of expression and can be either automated or manual. An example of such moderation is the moderation of speech from terrorist or extremist groups.

Both reactive and active moderation can be carried out by two types of moderators:

- **Humans:** moderators who are trained to review different content depending on the platform they are on.

- **"Algorithms":** automated systems using machine learning that are trained to search for certain content and remove it if it breaches terms and conditions.[117]

It is important to note that automated moderation has the dilemma that speech is generally delivered in a particular context that contains issues that need to be assessed, such as intentionality and the historical, cultural and social circumstances of the particular case, so that content can often be suppressed or moderated incorrectly by systems because of their inability to take that context into account.[118]

The contextual dilemma of expressions is what explains the need for mechanisms to appeal and review the moderation carried out by automated systems, and to question whether content moderation is carried out by them alone.

115 Klonick, Kate. "The New Governors..." op. cit., p. 1638.

116 Id. Errors arising from the decontextualization and interpretation of the different variables to be taken into account (and also from the biases inherent in the systems) can undoubtedly enhance discrimination against vulnerable groups. In this regard, see: Del Campo, Agustina; Schatzky, Morena; Hernández, Laura; Lara, Juan Carlos. Mirando al Sur. Towards new regional consensus on the responsibility of Internet intermediaries, Al Sur, Abril 2021, p. 31.

117 Idem.

118 Roberts, Sarah. "Behind The Screen", op. cit., p. 34.

d. Effects of moderation on freedom of expression and other rights

The main dilemma in the content moderation debate lies in the difficulties of finding an ideal balance between moderation that is almost non-existent or moderation that is disproportionate and ends up damaging fundamental interests such as freedom of expression.

Next, we will discuss the two extremes of the pendulum and their consequences, mainly those that affect freedom of expression but also rights such as non-discrimination, the right to physical and mental integrity, as well as the rights of children and adolescents.

i. Practical considerations and limits of different moderation methods: What happens if there is no moderation?

It is impossible to claim that there is a platform or website that does not engage in some form of content moderation. Particularly because there are specific requirements in law to remove content that is not protected by freedom of expression, as noted above.

Moderation exists and it should not be debated about as if it were a simple problem. The main consequences of not moderating can be summarized as follows:

- **Spam**
  Unsolicited automated or coordinated messages that are sent repeatedly in order to grab a user's attention by flooding the information channels.[119]

  The spam content can vary from a strategy to sell products to a criminal scheme to gain access to users' personal data or accounts.

- **Online harassment, threats and violence**
  Online violence is a recurrent issue on online platforms, especially against vulnerable groups. One fact that reflects this situation is that young women between 18 and 30 years of age are the most affected by these aggressions, as well as the fact that 40% of this violence is committed by people known to the survivors and 30% by strangers.[120]

  Group dynamics increase this type of behavior,[121] and can lead to coordinated attacks on their victims, which inevitably hinder or violate the right to freedom of expression of the individuals or groups subject to harassment.[122]

119 Internet Society. What Is Spam. Accessed 31 December 2018. Available at: https://www.internetsociety.org/resources/doc/2014/what-is-spam/#_ftn2.

120 Luchadoras, Article 19, Cimac, et. al. " Online violence against women in Mexico", 2017, p. 16. Available at: https://r3d.mx/wp-content/uploads/180125-informe_violencia_en_linea_mx-v_lanzam.pdf.

121 Sunstein, Cass. *Republic.com* 2.0, Princeton University Press, 2009. p. 60.

122 Keats Citron, Danielle. Hate Crimes in Cyberspace, Harvard University Press, United States of America, 2014. pp. 193-197.

- **Child pornography**
  Child pornography, as mentioned in the section on limits to freedom of expression, is not protected by the right to freedom of expression but is expressly prohibited speech.

  There are numerous cases that show the real problem of child pornography and the measures that both authorities and intermediaries have taken to fight it. In the United States, for example, there has been an increase in cases of child exploitation and child pornography since 2012.[123] Both the FBI and the Department of Justice, as well as different platforms, have collaborated to develop strategies to counteract these crimes.

  It is worth stressing that US law makes an exception to the principle of intermediary immunity in matters relating to child pornography. Therefore, automated systems have been implemented for the detection of child pornography images.

  An example of such a system is Photo DNA developed by Microsoft.[124] Photo DNA uses a hashing of images that can be matched to a database of photographs collected from child pornography databases. This facilitates the detection of pages or publications containing such images and assists in the prosecution of this crime.

- **Sexually explicit content**
  Discussions around the publication of sexually explicit content are central to content moderation. Sites such as YouTube or Instagram have decided that their platform does not host pornographic content. For example, YouTube bans any sexually explicit content that generates sexual gratification but allows content with nudity when it is for educational, artistic or health purposes (similar to Instagram), and may restrict content that is not sexually explicit but has "sexual innuendo" for certain age groups.[125]

  There are gray areas, as YouTube points out when trying to define sexual content, as well as certain issues that may over-censor content protected by freedom of expression (discussed in the next section).

  The reasoning behind such measures is that many platforms want access to specific groups of people, and that the inclusion of sexually explicit content could drive many users away from their platform.

  The dissemination of sexual content without consent is also an issue that platforms need to address. Some platforms such as Twitter or Facebook offer a means of appeal when such content is uploaded, however, such mechanisms need to be expedited, transparent and effective.[126]

---

123 Dube Ryan. "Unfortunate Truths about Child Pornography and the Internet", Make Use Of. Accessed 7 December 2012. Available at: https://www.makeuseof.com/tag/unfortunate-truths-about-child-pornography-and-the-internet-feature/.

124 International Centre for Missing and Exploited Children. "Giving law enforcement the tools it needs to fight child sexual exploitation". Available at: https://www.icmec.org/train/law-enforcement/technology-tools/.

125 YouTube. "Nudity and Sexual Content Policies". Available at: https://support.google.com/youtube/answer/2802002?hl=en&ref_topic=9282679#zippy=%2Cother-types-of-content-that-violate-this-policy%2Cage-restricted-content.

126 For more information see: "Keats Citron Danielle "Hate crimes in cyberspace", op. cit.; Goldberg, Carry. "Nobody's Victim", op. cit.; Luchadoras MX, Article 19, APC, et. al. "Online violence against women in Mexico. Available at: https://r3d.mx/wp-content/uploads/180125-informe_violencia_en_linea_mx-v_lanzam.pdf.

- **Graphic or explicit violence**
  There is a fairly widespread consent on social media platforms on the prohibition of explicit violent content. While the extent varies between platforms, the overall aim is to prevent such content from being used to promote violence.[127]

  For example, terrorist or organized crime organizations have used social media outreach as an advertising tool to recruit members, as well as to advertise their menace, strength and to remind the public of their power.[128] Organized crime has a significant presence in Mexico, particularly in relation to drug trafficking cartels. For example, the communication strategy that various organized crime groups maintain on the *TikTok* platform was recently documented.[129]

  However, this view does not exist without nuances either. In the following section we will analyze the difficulty of reviewing publications under such broad criteria. In particular, when they pursue a legitimate objective by serving the public interest or protecting other human rights.

ii. The effects of vagueness or ambiguity in moderation criteria

There is no moderation that does not deal with nuances, and we have previously pointed out how some criteria such as "sexually explicit content" or "violent content" are concepts that can be controversial in their interpretation and that it is generally up to the platform to resolve such controversy

For this reason, it is particularly important to discuss the nuances and, when they exist, the cases in which moderation should take into account particular elements that avoid reaching a moderation that ends up affecting the freedom of expression of platform users. Here are some of these situations.

- **Graphic or violent content of public interest**
  In this case, the exception to the rule banning the dissemination of violent content is evident when a video is of public interest. Although it is shocking content or content that may be unpleasant for many people, its relevance lies, for example, in reporting crimes against humanity (or other types of crimes) that are silenced by governments or other subjects.

  An example of this was the first video of the Syrian war, uploaded to YouTube in 2011, showing a video of the body of teenager Hamza al-Khatib being beaten and burned. Hamza had been arrested while attending the protest against the government of Bashar al-Assad, so this video sparked public outrage and the teenager became the symbol of the Syrian Revolution. However, YouTube decided to remove it from its platform for being against its graphic content policy.

  This decision was controversial because of the high public interest the images represented for people in Syria, questioning what kind of power these platforms had to decide when

127 Gillespie, Tarleton. "Custodians of the…", op. cit., pp. 54-55.

128 Fernandez M., Alberto. "Here to stay and growing: Combating ISIS Propaganda Networks", Broo- kings Institution. Available at: https://www.brookings.edu/wp-content/uploads/2016/07/IS-Propaganda_Web_English_v2-1.pdf.

129 Lopez, Oscar. Mexican cartels invade TikTok". The New York Times. Accessed November 28, 2020. Available at: https://www.nytimes.com/es/2020/11/28/espanol/america-latina/cartel-tiktok.html.

something was relevant to society and when it was not. Following multiple protests against the platform's decision, YouTube decided to keep the video on its platform with an age-restriction filter.[130]

Another clear example is the case of Mexico, where similar controversies have also been reported, such as the video posted on Facebook showing the execution of a teacher by a drug cartel.[131]

This video was shared by different users to condemn the violence.[132] Public interest in them would be to reject the denialist stance of the Mexican government towards drug violence. The other videos of reports were uploaded by citizens or journalists who used social networks to denounce the violence that continued to take place in the north of the country. Such violence was dismissed by the Presidency of the Republic alleging that citizens were suffering from "collective hysteria". Therefore, the only way of reporting the situation of violence was through social networks.

However, the video of the teacher's execution was criticized in other countries because it was unnecessarily graphic for a platform like Facebook where highly impressionable people could watch it. Although Facebook had initially decided to keep it on the platform, it later decided to remove it.[133]

- **Sexual content or nudity**
  Defining what is sexually explicit content is complicated, to say the least. Even though a platform may not seek to host pornographic content to reach a specific audience, it is relevant to ask, what are the standards to consider explicit sexual content or not? What happens, for example, with sex workers who offer services on some platforms because it is a safer way than doing it on the street?

  Regarding the first question about what sexually explicit content is or not, some platforms make a catalog with body parts that are exhibited, others point out that it is content that seeks to generate sexual satisfaction in people, and some make catalogs of body parts that they consider "inappropriate" to be exhibited on their platform.

  From the above, it is necessary to mention the high level of subjectivity that exists at the time of moderating this type of content. A person with more conservative criteria or perceptions may censor expressions protected by freedom of expression, the free expression of sexuality or even the right to health.

130 Kaye, David. "Speech Police. The Global Struggle to Govern the Internet," Columbia Global Reports, 2019, pp. 22-23.

131 Grant, Will. "Facebook beheading video: Who was Mexico's Jane Doe?," BBC News, Accessed November 4, 2013. Available at: https://www.bbc.com/news/magazine-24772724.

132 Kelion, Leon. "Facebook lets beheading clips return to the social network," BBC News, Accessed October 23, 2013. Available at: https://www.bbc.com/news/technology-24608499.

133 Memott, Mark. "Facebook removes beheading video, says it will tighten rules". NPR. Accessed October 23, 2013. Available at: https://www.npr.org/sections/the-two-way/2013/10/23/240190936/facebook-removes-beheading-video-says-it-will-tighten-rules.

One example of the problems with this level of subjectivity is that of activist groups promoting the normalization of breastfeeding. In 2008, they uploaded photos breastfeeding their babies on their Facebook profiles as part of a virtual protest against the policy banning the publication of breastfeeding images that had visible nipple.[134]

In 2015,[135] Facebook removed once again the image of a mother breastfeeding her child, a situation that again led to complaints from users and ended with the platform modifying its community standards to make it explicit that breastfeeding images are allowed on its platforms,[136] this time regardless of which part of the breast is displayed in the publications.

- **Hate speech, polarizing or shocking speeches**

  Stigmatized speeches towards vulnerable groups are reprehensible and may end up affecting the rights of people who are part of such groups. However, these speeches cannot be limited to *banned words*, since language evolves continuously and is also used in different ways depending on the context; that is, words lack meaning when they are taken out of a specific context.

  For example, there are derogatory racial slurs with a homonym to another every-day term. There are also insulting words toward vulnerable groups that have been appropriated by them, such as the insults toward the LGBTIQA+ community that are now used among people who are members of these groups.

  Another example is that sometimes controversial words can be used in artistic spaces, for journalistic purposes or even on occasions where the insult is already detached from its original concept, so they do not have a stigmatizing effect on vulnerable groups.

iii. The link between concentration and the impact on human rights

The impact of the decisions to moderate online content regarding the right to disseminate, receive or seek information depends, to a large extent, on alternatives for users to exercise their right to freedom of expression on a different platform or service.

That is, it could hardly be argued that removing a publication from an intermediary significantly affects their right to freedom of expression or the free flow of information if a user has publication alternatives with the same or greater possibility of reach.

On the contrary, when a dominant Internet platform makes a moderation decision on its platform, the inability or difficulty of disseminating, receiving or searching for certain information with the same reach through another platform has a decisive impact on the right to freedom of expression.

Therefore, when analyzing the complex reality of content moderation on the Internet, it is essential to distinguish between content moderation carried out by platforms that, due to their

134 Sweeney, Mark. "Mums furious as Facebook removes breastfeeding photos", The Guardian. Accessed December 30, 2008. Available at: https://www.theguardian.com/media/2008/dec/30/facebook-breastfeeding-ban.

135 Idem.

136 Facebook. "Does Facebook allow posting of breastfeeding mothers?". Accessed February 20, 2021. Available at: https://www.facebook.com/help/340974655932193.

scale or other factors, can significantly limit the scope of an expression, and content moderation carried out by intermediaries without such power.

e. The difficulty of moderation at scale

So far, we have described the general problems resulting from the moderation of content on platforms, however, not all platforms have the same type of moderation and more importantly: not all platforms are large ones.

When the legislature suggests fighting "malicious" content found on platforms, they often refer only to market-dominant platforms such as Facebook or Google. Unfortunately, the misguided view of reducing the Internet to a few platforms can provoke or aggravate barriers to competition for emerging platforms, in favor of dominant companies that in their early days benefited from the lack of this type of strict regulation, which allowed them to grow and obtain the dominant position they have today.

To avoid the obstacles that limit competitiveness and affect users of digital platforms, it is necessary to understand the main problems of moderation at scale and to understand that moderation is a zero-sum game: there will always be someone who ends up dissatisfied with the moderator's final decision.[137]

Masnick's Impossibility Theorem[138] states that large-scale content moderation is impossible to do perfectly (100%). Masnick argues that in moderation there will always be someone who wins and someone who loses, so there will never be a scenario in which everyone is satisfied with the outcome. The dilemma becomes more complicated as the number of users and posts to moderate increases:

> Getting 99.9% of content moderation decisions at an "acceptable" level probably works fine for situations when you're dealing with 1,000 moderation decisions per day, but large platforms are dealing with way more than that. If you assume that there are 1 million decisions made every day, even with 99.9% "accuracy", you're still going to "miss" 1,000.[139]

On the other hand, the technical and human resources needed to carry out content moderation are costly. 350 million photos are uploaded daily on Facebook, not counting other types of posts.[140] It takes an army of moderators and automated systems to handle these massive amounts of information every minute for the platform to operate.

---

137 Goldman Eric, Miers Jess. "Why Internet Companies Can't Stop Awful Content," Social Science Research Network. Rochester, NY. January 1, 2020. p. 3. Available at: https://doi.org/10.2139/ssrn.3518970.

138 Masnick Mike. "Masnick's Impossibility Theorem: Content Moderation at Scale Is Impossible To Do Well", Techdirt. Accessed October 5, 2021. Available at: https://www.techdirt.com/articles/20191111/23032743367/masnicks-impossibility-theorem-content-moderation-scale-is-impossible-to-do-well.shtml.

139 Idem.

140 Cooper Smith. "Facebook Users Are Uploading 350 Million New Photos Each Day", Business Insider. Accessed October 12, 2021. Available at: https://www.businessinsider.com/facebook-350-million-photos-each-day-2013-9.

For example, Facebook has 30,000 people working on platform security, of which 15,000 are moderators on full-time contracts, not counting the people hired as service providers so that it can "go global". The average pay of an employee at Facebook is $240,000 per year, while that of a service provider is only $28,800 per year. Facebook is a company that in 2019 reported earning $6.9 billion a year in revenue.[141] Even Zuckerberg announced in 2019 that he would invest more than $3.7 billion for security issues on the platform, and even mentioned that it was much more than Twitter's total annual earnings.[142]

The fact that Facebook can boast about its security budget against the total earnings of another platform shows the disparities that exist even among the largest platforms. Therefore, moderation criteria should not be set that are only feasible for companies as large as Facebook or Google. The large platforms may pay the costs to comply with their new legal obligations because they were built with those platforms in mind in the first place, but these rules would pose a barrier to market entry for future emerging platforms or alternative social space projects that focus on user-generated content.[143]

Another problem with regulatory initiatives that consider platforms as homogeneous is that they overlook the fact that moderation is subjective and that there are different types of moderation. It is a common mistake to think that all moderation processes are done automatically or by the army of moderators who are reviewing every single post every second. There are platforms such as Reddit, where they have general guidelines for their entire platform and have teams to review the misconduct of a subgroup,[144] but they also have volunteer moderators in each subreddit who are dedicated to ensuring that the people affiliated with that subgroup comply with the agreed rules of that specific space.[145]

The main reasons for considering differentiated regulation based on size may be to (1) hold accountable the large firms that have done the most damage, (2) reduce the barrier to entry for competitors, and (3) aim for fairness in a market of competitors with large differences in profit and size.[146]

The current *Digital Services Act* (DSA) proposal being discussed in Europe provides an exemption for micro and small intermediary companies. The regulation aims to avoid disproportionate burdens on emerging companies unless these companies have a similar reach or impact to a large platform.

---

141 Newton, Casey. "The Secret Lives of Facebook Moderators in America," The Verge. Accessed February 25, 2019. Available at.: https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

142 Lauren Feiner Rodriguez, Salvador. "Mark Zuckerberg: Facebook Spends More on Safety than Twitter's Whole Revenue for the Year," CNBC. Accessed May 23, 2019. Available at: https://www.cnbc.com/2019/05/23/facebook-fake-account-takedowns-doubled-q4-2018-vs-q1-2019.html.

143 Eric Goldman, Jess Miers. "Why Internet Companies..." op. cit., p. 4.

144 Reddit. "Content Policy - Reddit". Accessed October 5, 2021. Available at: https://www.redditinc.com/policies/content-policy.

145 These volunteer moderators also have guidelines to follow that are imposed by the platform. In this regard, see: https://www.redditinc.com/policies/moderator-guidelines.

146 Goldman, Eric; Miers, Jess. "Regulating Internet Services by Size," SSRN Scholarly Paper, Social Science Research Network. Rochester, NY. Accessed May 1, 2021, p. 2. Available at: https://papers.ssrn.com/abstract=3863015.

The DSA sets greater responsibilities on very large platforms, and defines a "very large platform"[147] as those that have a monthly number of active users in the European Union equal to or greater than 45 million. This is under the principle of proportionality, so that although the requirements are greater and stricter, it is also true that large companies have the budget and infrastructure to comply with these obligations.[148]

There are different types of metrics for the size of a platform:[149]

- By age of the company
- By number of employees
- By market capitalization
- By revenue
- By user consumption: which in turn can be divided into user consumption per month, registered users or page views.

Goldman and Miers consider that there is no categorical answer to which specific metric should be used, as each may have its disadvantages if applied categorically. However, the authors propose that the following factors should be taken into account:[150]

1. Metrics must be published and constantly audited.
2. The metrics must have a clear definition of the organization, material boundaries and economic boundaries that constitute each platform. For example, know on which corporate structure is being measured (example: it is not the same to measure only Google's service as Google, than to measure it together with its other services such as Gmail, Alphabet, etc.). Just as it is not possible to measure only by the content generated by users without considering the market capitalization of the platform. The clear example is Wikimedia, which can be considered a huge platform due to the large amount of content used by its editors and other users, but does not receive as much revenue and has a very small team.
3. The period of the measure. Regulators should specify the time frame over which time frame over which the metric is to be measured.
4. Use different metrics to avoid market volatility and false positives.

---

147 European Commission, "Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and Amending Directive 2000/31/EC" (2020), article 25, para. 1, https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-parliament-and-council-single-market-digital-services-digital-services.

148 European Commission, "Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and Amending Directive 2000/31/EC" (2020), 7–11, https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-parliament-and-council-single-market-digital-services-digital-services.

149 Goldman, Eric; Miers, Jess. "Regulating Internet Services by Size." op.cit. p. 2-3.

150 Id., p. 4-5.

f. Jurisdictional Aspects of Content Management

In this paper we have referred to a general framework of standards that have a diverse application in local jurisdictions and that may change according to the rigidity of each legal system. In addition, some governments seek to influence or regulate platforms to request the removal of content not only from their jurisdiction but from all other countries, supported by their legislation, judicial processes or by influencing platforms with their public power. For example, in November 2006, the Thai government announced that it would block YouTube for everyone using a Thai IP unless Google removed 20 videos that went against a law prohibiting insulting the king (punishable by 15 years' imprisonment).[151]

For Nicole Wong, a Google employee, it was a cultural shock because while some images were clearly against community norms, there were other cartoons that were just images manipulated in Photoshop. However, in the Thai cultural context there is widespread fondness for the king, so they decided to remove the videos within Thailand's geographic boundaries.[152]

Another similar incident occurred in Turkey: a parody show insinuated that Mustafa Kemal, the founder of modern Turkey, was a homosexual. As a result, a judge ordered that all Turkish users be blocked from accessing YouTube. While the video was later removed voluntarily, the government demanded that Google remove several more offensive videos from the platform.[153]

Google agreed to remove the videos that the company believed did indeed violate Turkish law, but only from Turkish jurisdiction. A year later, the Turkish government demanded that it ban access to these videos worldwide, Google refused, and the Turkish government blocked access to the YouTube platform throughout the country.[154]

Clearly there are logistical problems for platforms in trying to implement different moderation models in different countries. Many of the dominant platforms that developed in the U.S. are based on First Amendment principles, but when these companies reached global levels, they discovered that jurisdiction-specific content filters complicate moderation, and so they decided to promote the same set of standards and expand them as they occurred according to the legal pressures of the government in question.[155]

When a platform decides to strictly impose a series of moderation principles globally, without taking into account the human rights and cultural context of each country, it causes a series of violations of the rights to freedom of expression and access to information.

The Supreme Court of Canada case of Google Inc. v. Equustek solutions Inc. illustrates how a country's decision affects the right to freedom of expression. In this case, the dispute arose because the company Equustek sued the company Datalink for claiming that one of its products violated the intellectual property of the first company. Equustek required Google to de-index Datalink's pages that were used to do business online. After a court ordered Datalink to cease operating and doing business online, Google removed the links from its Canadian domains but

151 Klonick, Kate. "The New Governors...," op. cit., p. 1623.

152 Idem.

153 Ibid, p. 124.

154 Idem, citing Jeffrey Rose, "The Delete Squad," New Republic, April 29, 2013.

155 Keller, Daphne. "Who do you Sue...". op. cit., p. 8.

refused to remove the global domain results. Equustek applied through a court for an interlocutory injunction to de-index globally; Google appealed to the Supreme Court of Canada, which ruled in the company's favor.[156] The Canadian court's reasoning was as follows:

1. Google was required to comply with the order to stop facilitating Datalink's damage to Equustek. The Canadian court argued that a court can order an injunction that is binding on the infringer's conduct anywhere in the world because "the internet has no borders, its natural habitat is global.[157] Therefore, the Canadian court decided that the injunction must have a global impact to ensure its effectiveness.

2. While Google argued that the decision to remove the content internationally could result in international liability on the part of the Canadian state by violating the jurisdiction of other states and affecting freedom of expression, the Court dismissed this argument as a theoretical assumption because "most countries would recognize the violation of property rights and would see the legal liability in selling pirated products.[158]

3. On the issue of infringement of freedom of expression, the Court decided that in the event that Google had evidence that such an injunction violated the laws of another jurisdiction, including the right to freedom of expression, it could consult with the British Columbia courts to modify the injunction to the particular case.[159]

The decision was widely criticized for resulting in a clear infringement of the jurisdiction of other countries and in the infringement of human rights such as freedom of expression and access to information.[160] In short, the Canadian court created a precedent that favors commercial interests over freedom of expression in different jurisdictions and can be used to justify restrictions on human rights on a global scale.

For these reasons, the decision was challenged in the District Court of Northern California, where it was ruled that the injunction ordered by the Canadian court could not be enforced in the US, because of the immunity that Section 230 of the CDA grants Google in US territory.

The platforms also carry out global takedowns that affect freedom of expression. An example in the US is Yahoo! Inc. v La Ligue Contre Le Racisme et L'Antisémitisme, or Sikhs for justice v Facebook. The Yahoo Inc. v La Ligue Contre Le Racisme case is among the first to articulate this misunderstanding as "protecting the rights of users by preventing platforms from removing U.S. speech based on foreign law". The case addresses whether a U.S. court can enforce an order from France to stop Yahoo's search engine from displaying Nazi-related items, not whether Yahoo can voluntarily comply.

---

156 "Google Inc v. Equustek Solutions Inc. (Equustek I)," Global Freedom of Expression. Accessed October 11, 2021. Available at: https://globalfreedomofexpression.columbia.edu/cases/equustek-solutions-inc-v-jack-2/.

157 Google Inc. v. Equustek Solutions Inc., 2017 SCC 34, para. 41 (Supreme Court of Canada, decided June 28, 2017).

158 Ibid, para. 44.

159 Ibid, para. 45-48.

160 Aaron Mackey Ranieri Corynne McSherry, and Vera. "Top Canadian Court Permits Worldwide Internet Censorship," Electronic Frontier Foundation. Accessed June 28, 2017. Available at: https://www.eff.org/deeplinks/2017/06/top-canadian-court-permits-worldwide-internet-censorship; "Global Internet Takedown Orders Come to Canada: Supreme Court Upholds International Removal of Google Search Results - Michael Geist". Accessed 11 October 2021. Available at: https://www.michaelgeist.ca/2017/06/global-internet-takedown-orders-come-canada-supreme-court-upholds-international-removal-google-search-results.

Keller points out that, in fact, Yahoo voluntarily decided to comply with the French government's order while it was litigating the case in U.S. courts. Thus, "a company that fears that its foreign assets will be lost, or its employees arrested, or that does not want to lose access to a lucrative foreign market, may find good reason to follow foreign court orders and do so globally if asked to do so by the court."[161]

The case of Sikhs for justice v. Facebook is another example of a platform censoring valid speech to avoid confrontation with a country's jurisdiction. Sikhs for Justice (SFJ) is a human rights organization engaged in advocacy for the independence of Punjab in India. The organization had a Facebook page that it used for activism, organizing advocacy campaigns, and promoting the right of self-determination for Sikh people in Punjab.[162] In May 2015, Facebook blocked the page in India at the request of the Indian government. Sikhs for Justice asked the platform to return their account and provide an explanation for the block but the platform refused. The organization sued Facebook for damages arguing that the platform was responsible for racial discrimination.[163]

However, the District Court judge dismissed the organization's claims, noting that Section 230 protects moderation decisions, including the decision not to publish SFJ's content, so it could not be considered "discriminatory," but a decision that the platform was entitled to make.

The same happened in Zhang v. Baidu, where a group of pro-democracy activists in China sued the search company Baidu for blocking a variety of pro-democracy political speech in China in the United States at the request of the Chinese government. A federal judge controversially ruled that the platforms' decision about what content remained on and was removed from its pages was protected by the First Amendment, even though it would be used to censor speech in other jurisdictions.[164]

The Court of Justice of the European Union (CJEU) has also decided relevant cases regarding global removals and de-indexing of content. For example, the case of Glawischnig-Piszczek v. Facebook Ireland was decided by the Third Chamber of the Court, which deals with the de-indexing of unlawful content and the territorial scope of this decision.[165]

The facts of the case show that in 2016, a Facebook user shared on his personal page an article from a digital magazine that discussed Austrian politician Glawischnig-Piszczek. The user then posted, in connection with the article, a comment that the plaintiff considered to be damaging to her reputation and defamatory.[166]

---

161 Keller, Daphne. "Who do you Sue...". op. cit., p. 8-9.

162 Sikhs For Justice "SFJ," INC. v Facebook, INC. Case No. 15-CV-02442-LHK (Northern District of California District Court, November 13, 2015).

163 Idem.

164 "Zhang v. Baidu.Com, Inc.," Global Freedom of Expression. Accessed October 5, 2021. Available at: https://globalfreedomofexpression.columbia.edu/cases/zhang-v-baidu-com-inc/.

165 "Glawischnig-Piesczek v. Facebook Ireland Limited," Global Freedom of Expression. Accessed October 10, 2021. Available at: https://globalfreedomofexpression.columbia.edu/cases/glawischnig-piesczek-v-facebook-ireland-limited/.

166 Ibid, para. 12.

The Austrian politician sued Facebook before the Commercial Court of Vienna for failing to remove the comment. The Commercial Court ordered Facebook not to allow the publication or distribution of photographs of the plaintiff if they were accompanied with the exact text or words of equivalent meaning to that of the original comment. The appellate court upheld this decision but limited its scope: only content identical to the comment could be removed. The case reached the Austrian Supreme Court, which decided to refer the case to the CJEU for an interpretation of the Digital Commerce Directive legislation relevant to the case.[167]

The Third Chamber of the CJEU ruled that the Digital Commerce Directive does not prevent a member state from requesting a service provider to remove or block content that has been declared unlawful or content that is equal or equivalent to such unlawful information. Regarding the geographical applicability of that decision, the court indicated that the directive does not rule on any territorial limitation, so that each member state could determine the geographical scope of the restriction, as long as it was within the framework of the relevant international law.[168]

The CJEU also ruled on the scope of content removal in CNIL v. Google. The French personal data protection authority (CNIL) fined Google for failing to globally de-index information about an individual.[169]

The Grand Chamber of the International Court of Justice of the European Union ruled that European legislation was silent on the geographic scope of application for de-indexing orders. The Court found that the "Right to de-index" is not recognized globally. In this regard, the European Court emphasized that this right is not absolute and must be weighed against other fundamental rights in accordance with the principle of proportionality.

The Court of Justice established that in principle de-indexing should be possible in the jurisdiction of all member states, but given that privacy protections are not uniform in the European Union, it was up to the national courts to decide the scope of de-indexing. Finally, the Grand Chamber did not rule on whether Google could ever be obliged to carry out a global de-indexing, leaving it to each national court to decide whether this is appropriate.

---

167 Ibid, para. 14-20.

168 Ibid, para. 27-53.

169 "Google LLC v. National Commission on Informatics and Liberty (CNIL)," Global Freedom of Ex- pression. Accessed October 10, 2021. Available at: https://globalfreedomofexpression.columbia.edu/cases/google-llc-v-national-commission-on-informatics-and-liberty-cnil/.

IV. Transparency and accountability

Access to information has been recognized as a necessary right to fight corruption, to know about human rights violations—by State authorities and individuals—and to guarantee transparency as a fundamental tool for all democracies.

Transparency is essential especially in matters of public interest—including those involving restrictions or violations of human rights—because these enjoy a reinforced protection of the right of access to information. This guarantee is usually associated with a responsibility of the States; however, it is becoming increasingly important that companies—especially those with a leading position in their sector—and whose actions have an impact on the enjoyment of human rights, make regular transparency reports public for all people.

Along with the evolution of the right to information and transparency, a series of mechanisms have also been designed to ensure that States comply with their obligations to respect, protect and fulfill human rights and to recognize the role of companies as specialized agencies of society that perform specialized functions and must comply with all applicable laws and respect human rights.[170]

Under a balanced understanding of the duty of States to protect against human rights violations committed by third parties—including companies—in view of the right of access to information, there is a proactive responsibility that falls on companies to disclose those acts or omissions that have an impact on human rights.

On the obligations of access to information for intermediaries, various international organizations, such as the United Nations[171] and the OAS[172] rapporteurships for freedom of expression, have made pronouncements, which have specifically acknowledged that:

> "Private actors should ensure that their terms of service and community guidelines are sufficiently clear, accessible and in line with international human rights standards and principles, including the conditions under which they may interfere with the right to freedom of expression or privacy of users. In this context, companies should seek to ensure that any restrictions arising from the application of terms of service do not unlawfully or disproportionately restrict the right to freedom of expression."[173]

---

170 United Nations. Guiding Principles on Business and Human Rights. 2011. P. 1.

171 United Nations. General Assembly. Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue. A/HRC/17/27. 16 May 2011. Para. 48. Available for consultation at: http://ap.ohchr.org/documents/dpa-ge_s.aspx?m=8

172 United Nations Special Rapporteur on Freedom of Opinion and Expression, Representative of the Organization for Security and Cooperation in Europe on Freedom of the Media and OAS Special Rapporteur on Freedom of Expression. December 21, 2005. Joint Declaration on the Internet and on Anti-Terrorism Measures; United Nations Special Rapporteur on the Protection and Promotion of the Right to Freedom of Opinion and Expression and Special Rapporteur for Freedom of Expression of the Inter-American Commission on Human Rights. December 21, 2010. 2010. Joint Declaration on Wikileaks. Item 5; United Nations (UN) Special Rapporteur on the Protection and Promotion of the Right to Freedom of Opinion and Expression and Special Rapporteur for Freedom of Expression of the Inter-American Commission on Human Rights of the OAS. June 21, 2013. Joint Declaration on surveillance programs and their impact on freedom of expression. Item 11.

173 IACHR, Office of the Special Rapporteur for Freedom of Expression. Freedom of Expression and the Internet. OEA/Ser.L/V/II. IACHR/RELE/INF.11/13, December 31, 2013, para. 112.

In addition, the duties of intermediaries in terms of transparency are a fundamental part to know about potential acts of corruption and human rights violations by State authorities so, according to the IACHR, intermediaries should:

> "The law provides that companies should be sufficiently protected to make public requests made by state agencies or other legally authorized actors that interfere with the right to freedom of expression or privacy of users. It is a good practice, in this sense, that companies regularly publish transparency reports in which they disclose at least the number and type of requests that may entail restrictions to the right to freedom of expression and privacy of users."[174]

a. The Santa Clara principles

The second iteration of the Santa Clara principles (CSP) propose a set of principles aimed at promoting meaningful transparency and accountability with respect to content moderation carried out by Internet intermediaries.[175]

The CSPs set out a series of foundational and operating principles, as well as principles addressed to governments and other state actors. The foundational principles are general, cross-cutting principles that all companies, regardless of business model, age and size, should consider when conducting content moderation, including:

1. **Human rights and Due Process:** Companies should ensure that human rights and due process considerations are integrated at all stages of the content moderation process and should publish information outlining how this integration is made.

2. **Understandable Rules and Policies:** Companies should publish clear and precise rules and policies relating to when action will be taken with respect to users' content or accounts

3. **Cultural Competence:** Companies should ensure that their rules and policies, and their enforcement, take into consideration the diversity of cultures and contexts in which their platforms and services are available and used,

4. **State Involvement in Content Moderation:** Companies should recognize the particular risks to users' rights that result from state involvement in the development and enforcement of companies' content moderation rules and policies.

---

174 IACHR, Office of the Special Rapporteur for Freedom of Expression. Freedom of Expression and the Internet. OEA/Ser.L/V/II. IACHR/RELE/INF.11/13, December 31, 2013, para. 113. United Nations. General Assembly. Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue. A/HRC/17/27, May 16, 2011, para. 46. Available at: http://ap.ohchr.org/documents/dpage_s.aspx?m=85; Global Network Initiative. "A Call for Transparency from Governments and Telecommunications Companies"; Global Information Society Watch. "Don't censor censorship: Why transparency is essential to democratic discourse". As an example, see also: Google. Transparency Report; Twitter. Transparency Report. Communicate fearlessly to build trust, Microsoft. "Law Enforcement Requests Report.

175 Santa Clara Principles. The Santa Clara Principles On Transparency and Accountability in Content Moderation. Available at: https://santaclaraprinciples.org/.

5. **Integrity and Explainability:** Companies should ensure that their content moderation systems, including both automated and non-automated components, work reliably and effectively. This includes pursuing accuracy and nondiscrimination in detection methods, submitting to regular assessments, and equitably providing notice and appeal mechanisms, to only use the ones that assure high confidence levels and to offer transparency and independent supervision.

The operational principles, on the other hand, establish more detailed expectations for larger or more mature companies with respect to specific stages and aspects of the content moderation process.

In contrast to the minimum standards established in the first iteration (Numbers, Notification and Appeal), this second wave of principles provides greater specificity, with precision as to what information is required to ensure meaningful transparency and accountability.

This second iteration of the Santa Clara Principles expands the scope of where transparency is required with respect to what is considered "content" and "action" taken by a company. The term "content" refers to all user-generated content, paid or unpaid, on a service, including advertising. The terms "action" and "actioned" refer to any form of enforcement action taken by a company with respect to a user's content or account due to non-compliance with their rules and policies, including (but not limited to) the removal of content, algorithmic downranking of content, and the suspension (whether temporary or permanent) of accounts.

The SCPs propose 3 operational principles:

1. **Numbers** (Transparency). Companies should report the number of deleted posts and accounts suspended permanently or temporarily due to violations of their content policies.

2. **Notification**. Companies must notify each user whose content is removed or whose account is suspended of the reason for the removal or suspension.

3. **Appeal**. Companies should provide a meaningful opportunity to timely appeal any content removal or account suspension.

i. Transparency

On the principle of transparency, the SCPs establish a series of minimum requirements that companies must disclose to ensure that society respects the right of access to information and sufficient guarantees to supervise that their moderation does not affect the right to freedom of expression in an arbitrary or disproportionate manner.

Specifically, the SCPs provide that intermediaries must make public multiple categories of statistical information, including:

- Total number of flagged posts and accounts.

- Total number of deleted posts and suspended accounts.

- Number of flagged posts and accounts, and number of deleted posts and suspended accounts, by category of the rule that was violated.

- Number of flagged posts and accounts, and number of deleted posts and suspended accounts, by content format in question (e.g., text, audio, image, video, live streaming).

- Number of flagged posts and accounts, and number of removed posts and suspended accounts, depending on the source; i.e., governments, trusted reviewers, individual users, different types of automated detection.

- Number of flagged posts and accounts, and number of deleted posts and suspended accounts, depending on the source, i.e. governments, trusted reviewers, individual users, different types of automated detection.

- Number of posts and accounts flagged, and number of posts deleted and accounts suspended, by location of affected trusted reviewers and users (where relevant).

The SCPs recommend that the suggested data should be provided in a regular report, ideally quarterly, in a database-readable and openly licensed format.

ii. Notification

The importance of companies notifying users when their expression should be limited is based on detailed guidance from the company to the community so that users are aware of what content is prohibited. Examples of permissible and impermissible content and the guidelines used by reviewers or moderators should be included. Companies should also provide an explanation of how automated detection is used for each category of content.

When providing a user with notice about why their post has been actioned, companies should ensure that notice includes:

- URL, excerpt of content and other information sufficient to allow identification of the removed content.

- The specific clause of the guidelines that the content was found to violate.

- How the content was detected and removed (flagged by other users, trusted flaggers, automated detection, or external legal or other complaints).

- Specific information about the involvement of a state actor in flagging or ordering action.

Notices should be available in a durable form that is accessible even if a user's account is suspended or terminated. Notification of the results of the review, and a statement of the reasoning sufficient to allow the user to understand the decision.

iii. Appeal

The SCPs anticipate that a minimum—desirable—in appeals consists of:

- A process that is clear and easily accessible to users, with details of the timeline provided to those using them, and the ability to track their progress.
- Human review by a person or panel of persons who were not involved in the initial decision.
- The person or panel of persons participating in the review being familiar with the language and cultural context of content relevant to the appeal.
- An opportunity for users to present additional information in support of their appeal 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.

In the long term, independent review processes may also be an important component for users to be able to seek redress.

V. Recommendations on the regulation of content moderation by dominant internet intermediaries

Given the inter-American standards on freedom of expression and in view of the complexities and practical obstacles to the moderation of online content that have been developed, we believe it is essential that the following recommendations be taken into account when considering the creation of self-regulation, co-regulation or state regulation schemes for content moderation:

1. Non-liability of intermediaries for expressions of third parties

Any regulatory scheme must start from the general premise that intermediaries should not be held responsible for the expressions of third parties in circumstances in which they have not been involved in the modification of such content, since otherwise there are strong incentives for a content moderation prone to censorship of legitimate expressions.

This also implies that there should be no obligation to proactively monitor or filter content.

2. Differentiated approach and clear delimitation of the intermediaries to which the regulation would be applicable

The intermediaries to which any regulation would be applicable must be precisely delimited, ensuring that the parameters used to define regulated entities are adequately strict so as to be applicable only to those intermediaries that, due to their size, number of users, revenue level, market share, or any other relevant criteria, have the real capacity to significantly impact the flow of information.

A differentiated approach and a clear delimitation of the regulated entities is essential to prevent the regulation from having anti-competitive effects, i.e., favoring dominant stakeholders with greater economic, technical and administrative capacity to comply with the regulation, excluding or generating disproportionate burdens on smaller or less experienced intermediaries, causing greater concentration and affecting plurality and diversity in the supply of Internet services.

3. Policies in line with Human Rights

All self-regulation, co-regulation or state regulation schemes should encourage the adoption of content policies aimed at users that are in line with inter-American human rights standards.

State regulation should refrain from imposing content removal obligations, except with regard to the categories of non-protected speech strictly defined by the Inter-American system. That is, content that involves child sexual abuse, public and direct incitement to genocide and propaganda in favor of war, and advocacy of national, racial or religious hatred that constitutes incitement to violence.

A particularly damaging factor is the imposition of content moderation obligations based on vague and imprecise categories, which can have an inhibiting effect on users or even grant broad discretion to the State and private stakeholders to unduly restrict speech.

Policies established by intermediaries on content should be clear, precise and accessible to users. This way, users should be able to understand which types of content are banned and will be removed or will suffer other consequences, such as down-ranking, suspension or termination of the account.

Intermediaries should apply their content policies in a consistent and non-discriminatory manner. They should ensure that they evaluate the automated and non-automated methods used in content moderation in order to detect bias, errors or poor quality of decision making.

Companies should only use automated processes to identify and remove content or suspend accounts when they are used with human review mechanisms or there is sufficiently high confidence in the quality and accuracy of those processes.

The design, implementation and evaluation of content policies must include special consideration of the differentiated impact that the policies may have on specific groups, especially in terms of gender, race, language, disability, age, among others, as well as in terms of the context, such as elections, social protests or violence by the State or organized crime groups.

4. Limiting restrictions on content required by the law of other countries that are not compatible with inter-American human rights norms

States should avoid demanding that content removal decisions are applied globally, in particular those that are not compatible with inter-American human rights standards.

Intermediaries who are required to carry out legal obligations to remove content that are not compatible with Inter-American human rights standards should endeavor to judicially challenge such requirements and/or geographically limit the effects of such removals so that they do not apply to users located in countries that are part of the Inter-American system.

5. Transparency

Transparency about content moderation decisions made by intermediaries should be ensured as much as possible. Intermediaries should regularly publish sufficient statistical information to enable users, researchers and society to evaluate the effects of content moderation decisions.

The published statistical information should be disseminated at multiple levels and be structured in open formats. The Santa Clara Principles 2.0 provide detailed guidance on the statistical information that should be published on content moderation based on unilateral decisions by the intermediary, as well as those in response to a request from an authority.

6. Notifications

Self-regulatory, co-regulatory or state regulatory schemes should ensure that intermediaries notify users directly affected by a content moderation decision of the reasons for that decision.

Notifications should contain enough information to enable the infringing party to assess the relevance or lawfulness of the decision, including information that clearly identifies the allegedly infringing content, the regulatory basis on which the decision is based, the method of detection of the content (whether it is automated, reported by other users or requested by an authority) and in the case of removals motivated by authority requests, the legal basis and the identification of the requesting authority.

Notifications should be opportune, accessible and clearly state the appeal mechanisms available to the alleged infringer.

7. Appeal

Self-regulatory, co-regulatory or state regulatory schemes should provide for intermediaries to establish internal appeal mechanisms for content moderation decisions.

The appeal mechanisms should reverse a decision to remove content, suspend an account, or take any other action arising from the implementation of the intermediary's content policies. Where appropriate, they should incorporate reparations in accordance with the Inter-American human rights framework for cases that affect the human rights of users on their platforms.

Intermediaries must provide accessible, timely and sufficient information so that users affected by a content moderation decision or with a legitimate interest in such decision may have access to appeal mechanisms. This implies knowing details about the process, the communication channels for following it and the approximate time for its resolution, which should be as short as possible, especially in contexts such as elections or protests, where delay may render the decision ineffective in the appeals process.

8. Disaggregation of content moderation decision making

It is recommended that intermediaries consider establishing independent mechanisms to review content moderation decisions and policies. For example, mechanisms like Facebook's Oversight Board or the Social Network Councils proposed by the organization ARTICLE 19 can help avoid conflicts of interest and provide greater legitimacy to the content moderation decisions of dominant intermediaries.

The implementation of such mechanisms should ensure diversity and equitable participation of diverse groups in society, including geographic, gender, racial or ethnic diversity, among other categories.

Intermediaries must guarantee the financial sustainability and independence of disaggregated decision-making mechanisms with respect to content moderation, for which these mechanisms must act considering the transparency parameters previously developed.

9. Other measures to reduce concentration and promote plurality and diversity on the Internet

The power concentration of some intermediaries increases the importance of their content moderation decisions for the flow of information, which is why it is essential that States adopt measures to promote competition, plurality and diversity on the Internet.

Regulation and the work of the competition authorities must lead to the adoption of the necessary measures to prevent or sanction anti-competitive behavior by dominant intermediaries on the Internet, including measures such as interoperability and the divestment of assets.

Likewise, it is essential to guarantee the principle of net neutrality and the prohibition of zero rating offers based on commercial criteria that lead to anti-competitive effects and concentration in some intermediaries.

Adopting measures to reduce concentration and promote plurality and diversity on the Internet may require legal or regulatory reforms or the effective implementation of existing regulations by various regulatory authorities.

10. Multisectoral participation in the definition of policies and evaluation of practices.

The definition of self-regulation, co-regulation and state regulation schemes related to content moderation should ensure the open participation of all interested parties, including civil society, industry and international organizations, so that deficiencies or deficiencies in such schemes can be promptly noticed.

At the same time, multisectoral mechanisms for monitoring and evaluating content moderation implemented by intermediaries should be contemplated.

AlSur