Online Content Moderation Governance

Perceptions on the role of actors and regimes

INSTITUTE
FOR RESEARCH
ON INTERNET
AND SOCIETY

Online Content Moderation Governance

Perceptions on the role of actors and regimes

AUTHORSHIP

Fernanda dos Santos Rodrigues Silva
Júlia Maria Caldeira Gertrudes

INTERNAL REVIEW

Gustavo Ramos Rodrigues

EXTERNAL REVIEW

Pedro de Perdigão Lana

TRANSLATION

Paulo René da Silva Santarém

GRAPHIC DESIGN, COVER, LAYOUT AND FINISHING

Felipe Duarte

EDITORIAL PRODUCTION

IRIS - Institute for Research on Internet and Society

HOW CITE IN ABNT

SILVA, Fernanda dos Santos Rodrigues; GERTRUDES, Júlia Maria Caldeira. **Online Content Moderation Governance:** perceptions on the role of actors and regimes. Belo Horizonte: Institute for Research on Internet and Society 2023. Available at: <https://bit.ly/47yTvuL>. Accessed: dd mmm yyyy.

iris

INSTITUTE
FOR RESEARCH
ON INTERNET
AND SOCIETY

DIRECTORS

Ana Bárbara Gomes

Paloma Rocillo

MEMBERS

Bruni Emanuele | Administrative Analyst

Felipe Duarte | Head of Communication

Fernanda Rodrigues | Head of Research

Glenda Dantas | Researcher

Júlia Caldeira | Researcher

Luiza Dutra | Pesquisadora

Paulo Rená da Silva Santarém | Researcher

Rafaela Ferreira | Researcher

Thais Moreira | Communication Analyst

Victor Barbieri Rodrigues Vieira | Researcher

Wilson Guilherme | Researcher

irisbh.com.br

SUMMARY

1. Introduction

When it comes to content moderation, the platform's practice is usually the first thought that comes to mind. In the role of intermediaries, they are the ones who establish and apply the first norms, defined in terms of use and community guidelines, in order to control their online environment and guarantee users' freedom of expression, while seeking to protect them from exposure to harmful content. In this sense, due to the power they have, it is possible to state that these technology companies assemble a dynamic of centralized governance that is usually divided into three powers in a rule of law: legislative, executive and judiciary.

While setting the rules for users to communicate, interact and stay in that community, the platforms also apply these rules and make decisions based on them. In a context of self-regulation, it is intermediaries who become the focus of online content governance. However, looking at the internet governance ecosystem itself, it is possible to observe that other equally relevant actors are capable of influencing the power conferred on platforms.

State norms, for example, add up not only from a specific regulation for the sector, but also by bringing definitions of illegal activities that must be pursued and held accountable in any context. Civil society and the technical-scientific community present critical contributions and help the subject public debate evolution, by participating in the construction of international recommendations on moderation, such as the Manilla and Santa Clara Principles. Even users play an important role in bringing feedback from the usage of a given platform.

More recently, actors such as private and independent instances of moderation decisions review – for example, Meta's Oversight Board[1] – have also joined this scenario by sharing responsibility for making this activity more transparent and in line with international human rights parameters. In fact, the decentralization of decision-making by platforms can help ensure they are not the only ones to be activated when an error or misunderstanding is found in the process of intervening in third-party content.

[1] The Oversight Board is an "independent oversight committee, created especially to receive appeals from users dissatisfied with the company's decisions. (...) Under its bylaws, the Oversight Board can only review cases involving content that has been removed for violating Facebook policies". In: ARCHEGAS, João Victor; GODOY, Miguel Gualano de. The Jurisdiction Limits of Facebook Oversight Board: From Marbury *vs.* Madison to Facebook *vs.* Trump ("*Os limites da jurisdição do Facebook Oversight Board: de Marbury v Madison para Facebook v Trump*"). **Jota Info,** 02 Feb. 2021. Available at: https://www.jota.info/opiniao-e-analise/artigos/os-limites-da-jurisdicao-do-facebook-oversight-board-02022021. Accessed: 20 Dec. 2022. [Translation Note: in order to ease the English reading as much as possible, we opted to translate even the surnames of laws, names of organizations, as well as titles of articles, books and news, etc., while always displaying the Portuguese original text in italics and between parentheses or brackets, where parentheses already exist]

In Brazil, the discussion around moderation has gained even more prominence since the 2018 elections, in which the presidential race was influenced by the massive dissemination of fake news,[2] with social networks and instant messenger apps as main propagation vehicles. This trend, which had already been observed in the US 2016 electoral context, with Donal Trump election,[3] underscored the need to look at platforms more carefully and seek to understand how to face the proliferation of disinformation content.

In view of this, it is necessary to better understand aspects of the interaction between the different actors involved in online content regulation, and this understanding is essential to think about the regulatory models possibilities. From creating State norms to elaborating legal effectless recommendations, there are more mechanisms, other than those derived from the platforms, which can impact moderation procedures and, therefore, demand attention and a deeper study.

In this context, this research sought to identify the beliefs, perceptions and arguments surrounding different sectors' role in content moderation regulation (such as governments, companies and civil society participation in elaborating rules and standards) and the obstacles and advantages of each regulatory model type in adapting moderation to human rights. For this, interviews were carried out with specialists or professionals engaged with the online content regulation subject, coming from different sectors.

Thus, in order to present the results found, this work contains four more chapters, in addition to this introduction. The first is intended to present the theoretical framework aimed at situating the existing debate around content moderation regimes today, especially in the Brazilian scenario. The second is devoted to descripting the methodology used in this research, as well as its possible limitations. The third chapter is dedicated to interviews content analysis. And the last chapter is purposed to present the final considerations of the authors.

[2] GAMA, Sophia. Disinformation War: fake news in 2018 elections ("*Guerra de desinformação: as fake news nas eleições de 2018*"). **Curitiba City Hall ("*Câmara Municipal de Curitiba*"),** 15 Jul. 2022. Available at: https://www.curitiba.pr.leg.br/informacao/noticias/guerra-de-desinformacao-as-fake-news-nas-eleicoes-de-2018. Accessed: 19 Dec. 2022.

[3] MARS, Amanda. How did misinformation influence the presidential elections? ("*Como a desinformação influenciou nas eleições presidenciais?*") **El País,** 25 Feb. 2018. Available at: https://brasil.elpais.com/brasil/2018/02/24/internacional/1519484655_450950.html. Accessed: 19 Dec. 2022.

2. Content moderation management: between actors and regimes

As already mentioned, content moderation management is not restricted to the role played by online platforms. The accountability for controlling what is posted online is also shared with other internet governance actors, such as the government sector, civil society and the technical-scientific community, which can also influence the regulatory moderation process.

Thus, the regulation ecosystem can derive from different sources: the creation of State or international norms issued by public authorities, whether directly on platforms activities or on how they should proceed in regulating content; standards and guidelines with a recommendation character, issued by groups of non-governmental agents; besides the social platforms' own community policies and terms of use.

The greater or lesser degree of control exercised both by the different agents and by the different normative sources will, in turn, impact on identifying the current moderation regime model in a given context, which may be self-regulatory, co-regulatory or hetero-regulatory. With that in mind, this section will briefly address the role of the different actors involved in managing content moderation, as well as the rules that make up this normative environment.

2.1. Actors involved in content moderation management and the different regimes

Different authors sought to conceptualize the platform's governance and the agents involved in this ecosystem. For Gorwa, it engages not only platform users and companies, but also, fundamentally, political actors from different government branches, other stakeholders and advocacy groups, such as non-governmental organizations for privacy and digital rights, the scientific community and investigative journalists.[4]

However, the role played by the private sector is notorious, especially given the power intermediaries have over their own platforms organization. In this context, Gorwa points out the tendency for self-regulation initiatives with a single actor, which may or may not derive from government pressure.[5] These initiatives, in turn, are usually

[4] GORWA, Robert. What is platform governance? **Information, Communication & Society,** v. 22, n. 6, pp. 854-871, 2019, p. 857. Available at: https://www.tandfonline.com/doi/full/10.1080/1369118X.2019.1573914. Accessed: 24 Aug. 2022.

[5] GORWA, Robert. The platform governance triangle: conceptualising the informal regulation of online content. **Internet Policy Review,** v. 8, n. 2, jun./2019. Available at: https://policyreview.info/articles/analysis/platform-governance-triangle-conceptualising-informal-regulation-online-content. Accessed: 30 Aug. 2022.

translated into platform community policies or terms of use, containing rules about what content is allowed or prohibited in that space and guiding measures such as range restriction or removal of user posts.

For Laura DeNardis, this power over online discourse available to intermediaries can be considered as a "privatization of freedom of expression".[6] Whether due to government requests for censorship, such as for law enforcement or political gain; whether due to concerns about institutional reputation or contractual values and norms with the end user, in different contexts it is private regulations that replace (or add to) laws, norms and governments in conditioning freedom of expression in the public sphere.[7]

This concentration of responsibilities in the digital environment – to determine, manage and execute rules – highlights the power these private actors have and directly impacts people's lives, even being called "the new governors of online speech"[8] by some authors. According to Kadri and Klonick, in addition to the three powers common to the State, the platforms would also have a fourth function, as they also act as press or editors, that is, "controlling access to discourse on behalf of speakers and listeners".[9]

Proposals like Meta's Oversight Board appear as alternatives to balance this scale, seeking inclusion of other actors besides the companies themselves to influence the content moderation process. Meta's Oversight Board is an external supervisory body derived from years of pressure for instruments permitting Facebook's moderation decisions review.[10] According to Gorwa, however, it may be necessary to view the creation of this council as yet another private governance arrangement, also noting that the company's discourse has moved away from "The Supreme Court of Facebook" initial idea.[11]

[6] DENARDIS, Laura. **The Global War for Internet Governance.** Yale University Press, 2014, p. 157. https://www.jstor.org/stable/j.ctt5vkz4n.9

[7] *Ibidem*, p. 158.

[8] KLONICK, Kate. The new governors: the people, rules, and processes governing online speech. **Harvard Law Review,** v. 131, n. 6. pp. 1598-1669, abr. 2018. P. 1603. Available at: https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech/ Accessed: 30 ago. 2022.

[9] "Platforms are both the governors, setting speech policies and adjudicating speech disputes, and the publishers, controlling access to speech on behalf of speakers and listeners". *In:* KADRI, Thomas E.; KLONICK, Kate. Facebook V. sullivan: Public figures and newsworthiness in online speech. **Southern California Law Review**, v. 93, n. 37, pp. 37-99, 2019, p. 94. Available at: https://scholarship.law.stjohns.edu/cgi/viewcontent.cgi?article=1292&context=faculty_publications. Accessed: 30 Aug. 2022.

[10] KADRI, Thomas E.; KLONICK, Kate. Facebook V. sullivan: Public figures and newsworthiness in online speech. **Southern California Law Review**, v. 93, n. 37, p. 37-99, 2019. Available at: https://scholarship.law.stjohns.edu/cgi/viewcontent.cgi?article=1292&context=faculty_publications. Accessed: 30 Aug. 2022.

[11] GORWA, Robert. The platform governance triangle: conceptualising the informal regulation of online content. **Internet Policy Review,** v. 8, n. 2, jun./2019, p. 9. Available at: https://policyreview.info/articles/analysis/platform-governance-triangle-conceptualising-informal-regulation-online-content. Acesso em 30 Aug. 2022.

Oversight Board page describes its purpose as

> (...) to promote free expression by making principled, independent decisions regarding content on Facebook and Instagram and by issuing recommendations on the relevant Facebook company content policy.
>
> (...) The board is not designed to be a simple extension of Facebook's existing content review process. Rather, it will review a select number of highly emblematic cases and determine if decisions were made in accordance with Facebook's stated values and policies.[12]

It would be possible to verify, in theory, a kind of power substitution or division in relation to Meta's platforms' content moderation, especially regarding the paradigmatic cases review, which will serve as basis for general recommendations to the company. In a similar project, Twitter also have had a Trust and Safety Council, composed of different organizations and experts, operating through thematic advisory groups in areas such as "Online Safety and Harassment, Human and Digital Rights, Suicide Prevention and Mental Health, Child Sexual Exploitation, and Dehumanization", in addition to content governance itself.[13]

Although it cannot be denied self-regulation exercise has specific advantages, such as the possibility of carrying out interventions in certain situations more quickly than via legislation,[14] this process is also private companies' goodwill dependent and many times it does not have procedures that are transparent or explicit enough to allow effective oversight by civil society.[15] That is why other actors' actions can serve to help fill these gaps.

Although they do not have coercive power, entities from the third sector and the academia contribute not only through scientific content elaboration and production on the matter, but also through supervision of platform activities and participation in discussion spaces that often result in recommendation documents. The Santa

[12] **Oversight Board.** 2022. Available at: https://oversightboard.com/. Accessed: 30 Aug. 2022.

[13] TWITTER. **Trust and Safety Council.** 2022. Available at: https://about.twitter.com/pt/our-priorities/healthy-conversations/trust-and-safety-council. Accessed: 30 Aug. 2022.

[14] GORWA, Robert. What is platform governance? **Information, Communication & Society,** v. 22, n. 6, pp 854-871, 2019, p. 863. Available at: https://www.tandfonline.com/doi/full/10.1080/1369118X.2019.1573914. Accessed: 24 Aug. 2022.

[15] GORWA, Robert. What is platform governance? **Information, Communication & Society,** v. 22, n. 6, pp 854-871, 2019, p. 863. Available at: https://www.tandfonline.com/doi/full/10.1080/1369118X.2019.1573914. Accessed: 24 Aug. 2022.

Clara Principles,[16] as well as the Manila Principles,[17] are just some of the best-known examples of initiatives that started from civil society to collaborate with the online content moderation scenario.

Despite the absence of a binding effect, in fact these recommendations help to promote a healthier virtual space and respect for human rights, whether by influencing the legislative scenario, by exerting pressure on the private sector or by creating good practice parameters. Regarding the impacts of Santa Clara Principles first version, for example, the Director of International Freedom of Expression at the Electronic Frontier Foundation, Jillian York, highlighted having realized technology companies would have improved, mainly, in relation to offering ways of contestation of moderation decisions for users, which would demonstrate a breakthrough.[18]

Regarding the government sector, this assistance is given mainly through regulation, but it can also occur via supervision exercise, both by the Judiciary and the Executive Power. In the absence of more specific regulation, currently the content moderation scenario is heavily influenced by the Brazilian Civil Rights Framework for the Internet (*Marco Civil da Internet*) – MCI, which, despite not expressly mentioning the term "moderation", provides intermediaries liability for third-parties content exclusively in the case of non-compliance with a specific court order, with the legal exception of some content, such as the non-consensual dissemination of intimate images. Thus, it turns out this intermediary liability regime indirectly regulates the matter by influencing the structure of incentives for content removal or non-removal.

The State's participation in this sphere translates into a regime of moderation guided by co-regulation, in which there is active encouragement, support and eventual public authorities monitoring of self-regulation.[19] In this sense,

> mechanisms can have different levels of formality. Co-regulatory initiatives, in a strict sense, often include a formal regulatory element — such as a law or administrative decision — that acts as a framework and governs the activities of the actors involved, including rules and consequences of different kinds.[20]

[16] EFF – Electronic Frontier Foundation et al. **Santa Clara Principles on transparency and accountability in content moderation**. Available at: https://santaclaraprinciples.org/pt/cfp/. Accessed: 20 Dec. 2022.

[17] EFF – Electronic Frontier Foundation et al. **Manila Principles on Intermediary Liability**. Available at: https://manilaprinciples.org/principles.html. Accessed: 20 Dec. 2022.

[18] WHITELAW, Ben. Jillian C. York on the newly revised Santa Clara Principles. **Everything in Moderation,** 15 Dec. 2021. Available at: https://www.everythinginmoderation.co/jillian-c-york-santa-clara-principles/. Accessed: 02 Jan. 2023.

[19] ACCESS NOW. **26 recommendations of content governance**: a guide for lawmakers, regulators, and company policy makers. Available at: https://www.accessnow.org/cms/assets/uploads/2020/03/Recommendations-On-Content-Governance-digital.pdf. Accessed: 19 Dec. 2022.

[20] ACCESS NOW. **26 recommendations of content governance**: a guide for lawmakers, regulators, and company policy makers. p. 11. Available at: https://www.accessnow.org/cms/assets/

The creation of a State norm more directed to the moderation process, however, is not an easy task, as it presents a series of challenges. The German NetzDG act, for example, faced this problem and demonstrated a group of difficulties that can be generated by government regulation, such as incentives for digital platforms to carry out excessive moderation and the delegation to the private sector of the definition of what is legal and illegal.[21]

Even so, the role of the government sector remains relevant in setting limits for this activity and ensuring minimum guarantees to users. In a context in which most social platforms are from global North countries, the role of national regulators can help these companies provide a more personalized service that is also attentive to the particularities of the countries they are intended for, such as those located in the Global South. However, choosing the most appropriate moderation regime for the reality of each nation-state is one of the major discussions in the current scenario.

Finally, within the scope of moderation regime models, there would still be the possibility of adopting a hetero-regulatory regime , in which the norms would be drawn up by the public sector without the participation of the platforms themselves. This modality, however, would face difficulties such as, for example, the government's lack of expertise on each platform dynamics, which could jeopardize creating norms achievable or coherent to its operation. Nonetheless, it should be noted, in this case, the possibility of the government itself establishing commissions of experts or calls for public consultations to assist in the preparation of regulations, as carried out, for example, in the process of PL 21/2020,[22] regarding artificial intelligence regulation in Brazil.[23]

uploads/2020/03/Recommendations-On-Content-Governance-digital.pdf. Accessed: 19 Dec. 2022.

[21] ESTARQUE, Marina; ARCHEGAS, João Victor. **Social networks and content moderation: creating rules for public debate from the private sphere** (“*Redes sociais e moderação de conteúdo: criando regras para o debate público a partir da esfera privada*”). Institute of Technology and Society (“Instituto de Tecnologia e Sociedade”) – ITS: Rio de Janeiro, 2021, p. 20. Available at: https://itsrio.org/pt/publicacoes/redes-sociais-e-moderacao-de-conteudo. Accessed: 19 Dec. 2022.

[22] AGÊNCIA SENADO. **Suggestions for artificial intelligence law can be sent until May 13** (“*Sugestões para lei sobre inteligência artificial podem ser enviadas até 13 de maio*”). 05 Apr. 2022. Available at: https://www12.senado.leg.br/noticias/audios/2022/04/sugestoes-para-lei-sobre-inteligencia-artificial-podem-ser-enviadas-ate-13-de-maio. Accessed: 02 Jan. 2023.

[23] AGÊNCIA SENADO. **Artificial Intelligence: jurists commission delivers report on Tuesday** (“*Inteligência Artificial: comissão de juristas entrega relatório nesta terça*”). 05 Dec. de 2022. Available at: https://www12.senado.leg.br/noticias/materias/2022/12/05/inteligencia-artificial-comissao-de-juristas-entrega-relatorio-nesta-terca. Accessed: 02 Jan. 2023.

2.2. Regulatory model for content moderation in Brazil

As already mentioned, Brazil currently lacks legislation that is more specific on content moderation. Marco Civil da Internet, approved almost a decade ago, did not need to delve into the topic at the time, considering most urgent problems back then were aimed at establishing rights and minimum guarantees for Internet users.

Article 19 is the only one with which the legislator sought to exempt intermediaries from liability for third-parties content posted on their platforms, under the condition that there is no specific court order determining any action. A study carried out by Hartmann and Monteiro pointed out this was the point with the majority of proposals for alteration through bills since Marco Civil approval.

According to the authors,

> About **33% of the projects (16 PLs) propose, to some extent, to change platforms liability regime** – either completely eliminating the dynamics of judicial notification established by art. 19, or inserting new exceptions to this regime, in addition to cases of copyright and revenge pornography, already indicated in art. 19, § 2°, and in art. 21.
>
> (...) In general, the main concerns of parliamentarians conveyed in the bills are the **spread of false news, terrorist attacks and offensive content, which instigate bodily injury or suicide** .[24] (emphasis added)

Thus, it appears there is already a great concern to update what was provided by MCI in accordance with new demands that emerged over the last few years, especially after episodes such as 2018 and 2022 presidential elections. Various researches are trying not only to understand the nature and characteristics of these phenomena,[25] but also to assess possible limitations to the dissemination of this content type during election campaigns.[26]

[24] HARTMANN, Ivar; IUNES, Julia. Fake news in the pandemic and social emergency context: the social network platforms' duties and responsibilities in moderating online content between theory and legislative propositions ("*Fake news no contexto de pandemia e emergência social: os deveres e responsabilidades das plataformas de redes sociais na moderação de conteúdo online entre a teoria e as proposições legislativas*"). **RDP**, Brasília, v. 17, n. 94, p. 388-414, Jul./Aug. 2020. Available at: https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/download/4607/Hartmann%3B%20Iunes%2C%202020. Access:19 Dec. 2022.

[25] DOURADO, Tatiana Maria Silva Galvão. Fake news in Brazil's 2018 presidential election ("*Fake news na eleição presidencial de 2018 no Brasil*"). 308 f. Thesis (PhD) - Graduate Program in Communication and Contemporary Cultures ("*Tese (Doutorado) – Programa de Pós-Graduação em Comunicação e Culturas Contemporâneas*"), Bahia Federal University ("*Universidade Federal da Bahia*"), Salvador, 2020. Available at: https://repositorio.ufba.br/handle/ri/31967. Accessed: 22 Jan. 2023.

[26] SARLET, Ingo Wolfgang; SIQUEIRA, Andressa de Bittencourt. Freedom of expression and its limits in a democracy: the so-called "fake news" case in social networks during the election period in Brazil ("*Liberdade de expressão e seus limites numa democracia: o caso das assim chamadas 'fake news' nas redes*

This is the context in which it was proposed the PL 2.630/2020, also known as "*PL das Fake News*" (Fake News Bill), with the initial objective of combating desinformation, but based on the creation of provisions encouraging users mass surveillance, identifying accounts and collecting message forwarding records.[27] Despite the legitimacy of the goal of confronting disinformation, studies have drawn attention to the risks rules in this sense could bring to users' privacy and personality rights.[28]

Throughout the legislative process, however, the proposal underwent a series of changes, which ended up making it a kind of general regulation aimed at digital platforms. Regarding the content moderation process, the Bill has, for example, rules aimed at building a due process for the moderation procedure and the requirement to publish transparency reports, in addition to the creation of the Internet Transparency and Responsibility Council , a body dedicated to monitoring platforms compliance with the law.

Despite not being the focus, other legislation, such as the General Data Protection Law and the construction of a Legal Framework for Artificial Intelligence Regulation, in particular, – through PL 21/2020, which is under discussion in the Federal Senate – may also have an impact on the online content regulation sphere. Regarding the latter, the rules aimed at explainability, transparency and accountability of AI systems are more prominent. In a context marked by the widespread use of artificial intelligence systems for moderation rules enforcement, the aforementioned aspects of a possible AI regulation should significantly impact the content moderation issue.

These norms have the power to directly affect the use of automated technologies for content moderation, since such systems may present problems involving algorithmic bias, discrimination and technical failures in their operation. Still, considering that these tools work to limit the discourse published online, also affecting the freedom of expression of individuals, standards that facilitate the understanding and auditing of these systems emerge as essential.

sociais em período eleitoral no Brasil"). **Institutional Studies Magazine ("*Revista Estudos Institucionais*"),** v. 6, n. 2, pp. 534-578, may/aug. 2020. Available at: https://www.estudosinstitucionais.com/REI/article/view/522/511. Accessed: 19 Dec. 2022.

[27] COALIZÃO DIREITOS NA REDE. **Technical Note on the June 26, 2020 report to Bill 2,630/2020** ("*Nota técnica sobre o relatório de 26 de junho de 2020 ao Projeto Lei nº 2.630/2020*"). 28 Jul. 2020. Available at: https://direitosnarede.org.br/2020/06/28/nota-tecnica-sobre-o-relatorio-de-26-de-junho-de-2020-ao-projeto-lei-no-2-630-2020/. Accessed: 19 Dec. 2022.

[28] KURTZ, Lahis Pasquali; DO CARMO, Paloma Rocillo Rolim; VIEIRA, Victor Barbieri Rodrigues. **Transparency in content moderation: national regulatory trends** ("*Transparência na moderação de conteúdo: tendências regulatórias nacionais*"). Belo Horizonte: Instituto de Referência em Internet e Sociedade, 2021. Available at: https://bit.ly/3xjAUka. Accessed: 19 Dec. 2022.

3. Methodology

This section proposes to address the methodology used for the research development. Thus, the steps were divided into the following topics: i) questions script elaboration; ii) sample composition of interviewed people; iii) procedures for interviews carrying out; and iv) categories and codes for interviews content analysis.

Considering the interviewees secrecy guarantee, encryption mechanisms were used to safeguard the information confidentiality during all stages of the research. To this end, the researchers used GpgEx and Kleopatra softwares – to create keys and encrypt files – as well as a private messaging application, with end-to-end encryption – for transferring and sharing encrypted files.

3.1. Interview type definition and questions script elaboration

First, it should be noted that the authors chose to carry out a semi-structured interview, in which the elaboration of the script of questions served less as a rigid model to be followed and more as a reference guide for debating with the interviewees. Thus, participants were not always obliged to answer all the questions in the script. These variations occurred due to the perception certain questions had already been answered in a spaced manner throughout the interview, or due to the trajectory of the interviewee, for whom asking a certain question might not be relevant to their field of activity. Furthermore, freedom was given for the interviewee to choose whether to answer the question or move on to the next question.

That said, the main preliminary steps to preparing the script, which is attached in Appendix I, were meetings with the project consultant and other IRIS researchers, who also had carried out previous interviews research (namely, for the production of the *paper* "Cryptography and Criminal Investigations: what do the professionals who move the debate in Brazil think").[29] These moments, in turn, were important for defining some procedures used in the following phases, such as choosing the platform for interview conduction and creating the Free and Informed Consent Form (TCLE).

As for the script elaboration, 18 questions were formulated, based on previous research carried out from a systematic review of foreign and national literature about the different regimes of content moderation. With an approximate duration of 50 minutes, the interview objective was to understand the interviewees' conceptions about the issued themes. Thus, these were divided into 3 blocks, namely:

[29] The aforementioned work authors, Ana Bárbara Gomes and Gustavo Rodrigues, collaborated by sharing suggestions and contributions to the interview script, as well as with regard to encryption and encoding strategies for the interviews, which is why we express here our thanks. We also thank the methodological consultant for the project, Lucas Caetano Pereira, for his material contributions to this work, from the script preparation to the final writing phase, after the interviews.

> I. *Presentations and general themes*: 2 questions, aiming to be a moment of the interviewee' presentation and trajectory better understanding in relation to content moderation ;
>
> II. *Content moderation and interaction between the different actors involved*: 5 questions, exploring issues on the involved actors role in the addressed practices;
>
> III. *Regulatory models*: 11 questions, focusing on regulatory models and on perceptions about PL nº 2.630/2020.

3.2. Sample composition of interviewed people

In order to define who would be selected to compose the sample of participants, the search began with researching people involved with discussions on the topic of content moderation and its subtopics. The search included professional and academic spaces, highlighting activities such as authorship of scientific papers, participation in theme related events, legal contributions (such as participation in public hearings regarding theme related bills) and other various engagements.

People selection aimed to contemplate the participation of the four main sectors of internet governance (namely: government sector, private sector, technical-scientific community and third sector). At this point, two observations are important regarding the final sample: a) first, within the government sector, people active not only in the Executive Branch were considered, but also in other spheres of the public sector in Brazil in general; and b) besides, although people were identified as belonging to a specific sector, in some cases in interviewee could be linked to more than one sector, due to their performance. Despite this, priority was given throughout the text to linking the participant to the sector through which the interview was conducted, which was discretionarily determined by the researchers, according to people's observed performance.

At this point, it is important to highlight that the difficulty -- in establishing an objective set of scientifically valid rules to determine, in a concrete case, the belonging of person X or Y to sector A or B -- is not unknown. There are even researches on the multi-stakeholder nature of internet governance that discuss the fictionality of this division between sectors. Gustavo Rodrigues addresses it, based on the cyborg concept by the philosopher Donna Haraway, as "a material and semiotic fiction that is simultaneously descriptive and normative, real and imagined".[30] For Hofmann,

[30] RODRIGUES, Gustavo Ramos. "The function of a data protection law is to protect everyone, including the one who collects data": the debate on the General Data Protection Law and the ideology of harmony at the Internet Forum in Brazil (*""A função de uma lei de proteção de dados é proteger a todos, inclusive aquele que coleta dados": o debate sobre a Lei Geral de Proteção de Dados e a ideologia da harmonia no Fórum da Internet no Brasil*"). In: **VII ENCONTRO NACIONAL DE ANTROPOLOGIA DO DIREITO**, 2022,

> The difficult match between the stakeholder taxonomy and the political spectrum in Internet governance is at odds with the basic idea of multi-stakeholderism, which assumes that political positions can be aggregated along the lines of formal affiliations.[31]

Nevertheless, the choice to determine the interviewee's sector according to people's observed performance and affiliation, even in the face of eventual participation in multiple sectors, served as a methodological instrument for the analysis of a social field largely structured in the emic plane by the idea of multi-stakeholderism. In this sense, it is important to highlight the heuristic character of this sectoral division, without seeking to reify it as something static, fixed and simply transferable to other contexts in which other classification schemes can operate.

The invitations were then sent by email, divided among the researchers. The Free, Prior and Informed Consent Form was sent via Docusign platform, after acceptance, and the script could be sent at guest's request. The FPIC, in turn, contained information regarding research's purpose, justification, procedures, secrecy, privacy measures, etc. 18 interviews were concluded between the months of May and July, and, due to the departure of one of the researchers, another interview was carried out in December, in order to complement the number of interviewees. From the final sample, 4 people were representatives from the government sector, 3 from the private sector, 5 from the technical-scientific community and 6 from the third sector (according to the abovementioned definition). To maintain the confidentiality of the participants, for each one pseudonyms were generated randomly and through an online platform.

3.3. Procedures for interviews carrying out

The interviews were mainly carried out using the Zoom platform. Out of the 2 exceptions, both were exceptionally submitted in writing, as respondents preferred this participation method due to scheduling conflicts. Also, one of the written interviews was prepared by the guest interviewee's work team.

As for those that were carried out via video call, consent for recording was collected through FPIC, with no denial in this regard. Given the semi-structured nature of the interview, the script was used as a reference, with no rigidity on the questions order, and with freedom regarding theme deepening questions, asked in order to better understand the interviewees answers.

São Paulo. Anais [...]. São Paulo: NADIR/USP, 2021. v. 1. p. 12. Available at: bit.ly/3GBlZYx. Accessed: 11 Jan. 2023.

[31] HOFMANN, Jeanette. Multi-stakeholderism in Internet governance: putting a fiction into practice. **Journal of Cyber Policy,** v.1, n. 1, pp. 29-49. Available at: https://www.tandfonline.com/doi/pdf/10.1080/23738871.2016.1158303. Accessed: 11 Jan. 2023.

The interviews were transcribed by both researchers and the consultant. To aid these activities, the InqScribe program was used, which allows the execution of the audio/video together with the transcription, but doesn't collect transcribed content data.

3.4. Categories and codes for interviews content analysis

For the content analysis phase of the interviews, encodings were performed using Atlas.ti 8, a program that "contributes to the organization of data, optimization of the analytical process, in addition to being malleable for numerous types of qualitative research with different objectives".[32] Thus, the first step was creating the codes that would be used, a phase that researchers carried out jointly, in order to synthesize the most relevant categories and standardize the information that would be collected from the script. This systematic qualitative analysis was inductive, for it didn't start from a predefined set of criteria for creating codes and categories used.

In this way, 83 codes were created (coding scheme in Appendix II), divided among the researchers, so each one coded from that set. Later, spreadsheets with the codifications result were generated by the program, in MS-Excel format. These were condensed into a single spreadsheet for analysis, which was shared in encrypted form among the researchers. From this material, the contents arising from each question in the script were analyzed and the topics that form the next chapter of this work were written, through an interpretative approach and the narrative recreation of the found arguments.

3.5. Research limitations

Before starting results analysis, it should be noted this research has some limitations. First, the present work does not present representative results of a certain sector of internet governance or any other population segment. This was due to the unsystematic character of the sample selection method, which did not produce an empirical universe representative of the populations to be analyzed.

Second, it is also worth highlighting that it was difficult to carry out an equal number of interviews between the different sectors, either through contact or availability of agendas. Therefore, the technical-scientific sector and the third sector are overrepresented in comparison to the others. This may represent a bias regarding the final conclusions.

[32] SILVA JUNIOR, LA; LEÃO, M. B. C. The Atlas.ti software as a resource for content analysis: analyzing science teaching robotics in Brazilian theses ("*O software Atlas.ti como recurso para a análise de conteúdo: analisando a robótica no Ensino de Ciências em teses brasileiras*"). **Ciência & Educação,** Bauru, v. 24, n. 3, p. 715-728, Sep. 2018. Available at: https://www.scielo.br/j/ciedu/a/yBwC9L74v4vD3s4PwVXggsk/?lang=pt. Accessed: 21 Dec. 2022.

And third, the objective of this report was to draw an overview around the main opinions surrounding the content moderation theme among specialized actors involved, with the purpose of contributing to the matter' public debate.

4. Results: interviews analysis

In this chapter, the results found in the research are presented, based on the analysis and discussion of the arguments given by the interviewees within the different themes.

The arguments were analyzed through narrative reconstruction, seeking to present the logical relationships between them and highlighting the most frequent statements. Together, comments are made, accompanied by citations to theoretical frameworks, in order to discuss the statements.

Also, at the end of some topics, mental maps are presented in order to illustrate the main arguments raised by the interviewees.

4.1. On content moderation and its techniques

The following arguments were found based on the interviewees' answers to questions 3 and 13 of the script, which addressed the main functions and activities of content moderation, as well as the evaluation around the use of automated mechanisms (artificial intelligence) for this activity and the need for human review.

4.1.1. On content moderation main functions and actions

To begin, the first questions in the script – with a content-oriented nature – proposed to address what would be, for the interviewees, the main functions of content moderation and what actions were included in its scope. Given the great correspondence, answers could be divided into large groups.

For the listed functions, the division was between: **i. Virtual space sanitization (argument mentioned 14x)** and **ii. Platforms product (8x)**. As for actions understood as content moderation, the division was: **i. Content curation (10x)** and **ii. Intervention (13x)**. Thus, given the common points, but also the distinct and particular issues under the interviewees' conceptions, it becomes valid to discuss each topic individually.

Function: Virtual space sanitization

Digital platforms can serve different purposes. Although currently major social networks have successfully served the most varied interests on a single platform, the web is made up of a myriad of portals and communities. Consequently, behavior standards definition is needed to ensure users safe and harmonious navigation, in addition to restricting or even preventing acts considered inappropriate. In this sense, despite the large platforms having members of different nationalities, ethnicities and creeds, some rules of coexistence are universal so the user experience is possible and, moreover, satisfactory.

Based on this assumption, one point presented was that the promotion of a "healthy" character for the virtual space is up to content moderation practices. From this perspective, the act of "sanitizing" the network could be defined as the obstruction of unpleasant, harmful or even illegal content, with the main examples being hate speech, child sexual abuse and exploration material, terrorism, racism, misinformation, spam, online violence and acts such as cyberbullying and doxing, In this vibe, the guarantee of this basic filter was referenced through expressions such as: determining a civilizing minimum, maintenance of a users receptive space, and materializing a minimally habitable and conversation conducive or interaction subject environment.

Along with the obstruction of such content, some interviewees also linked the act of virtual space "sanitizing" to **security guarantee (3x), civil participation (1x) and users' free speech (2x)**. In this context, Maria Teixeira, a technical-scientific community member, states:

> I think the primary function of content moderation would be **to avoid or constrain criminal activities in general**. So crimes in every sense: racism crimes, in short, crimes of enticing minors, this type of thing that is more supervised, so to speak. But the parallel functions of this content moderation are precisely **identifying how these contents are not only criminal in this sense, they are more contents that seem harmless, but that behind them there are many other things**. They are lying contents, they are contents that can destroy the public image of certain individuals. So I think the function of content moderation, one of the functions, a parallel function besides constraining crimes, would be **constraining the free speech extrapolation**. So I think that's what content moderation is about. First, **avoid criminal activities and, second, avoid these harmful speeches in general.** Which do not always lead to crimes, but which are an extrapolation of freedom of expression (emphasis added).

Thus, through the interviewee's speech, the relationship is understood between sanitization of the network and the establishment of parameters for users' freedom of expression, while the act of avoiding harmful speeches is configured as a way of guaranteeing a respectful space.

Function: Product of platforms

Content moderation is a practice that, in addition to serving users, serves platforms, making it possible to execute its object and, therefore, guaranteeing its usefulness. In this context, the product function can be defined as a platform mechanism that ensures compliance with internal rules in favor of the current business model (when there is a commercial nature) or simply in agreement with the fulfillment of its initial proposal. Given this, regardless of the platform type, moderation is understood here as a way to ensure its proper functioning. When related to commercial interests, in turn, marketing aspects are considered, related to user targeting advertisements, as mentioned by Fernanda Rezende, an interviewed third sector member. However, it is important to point out that economic interests are not necessary for platforms to feel the need to preserve their rules through content moderation. Caio Fernandes, a government member, illustrates this in his speech when he says that

> (...) the platform has every right to have its rules and has every right to suspend the user who does not comply with the rules. If you have a ballroom dance club, you can't go in havaianas flip-flops because you can't do ballroom dancing in havaianas flip-flops. On the other hand, if it's a nudist club, you can't enter dressed, these are club rules. That doesn't mean there's a crime involved. So, because these things sometimes get mixed up.

In this context, Caio's speech also exemplifies how the product function, in turn, differs from those already exposed, while it aims to ensure not a basic filter, as in the case of sanitization or compliance with external rules and standards, but instead the adequacy to platform's internal rules – which may serve commercial interests or not. In this context, Theo Freitas, a third sector member, exemplifies this conception as follows:

> (...) is a thematic function. So, content moderation, thinking of it as an activity that is **the activity of elaborating and implementing rules, standards, community norms in Internet sectors open to user activity**, open to the generation of content by third-parties, thinking about it, I think the first function is a thematic function (emphasis added).

In a similar sense, Samuel Cardoso, a technical-scientific community member, states that the main function will be the platform proper functioning maintenance, which will vary according to the internal policy, objective and notion of what is "appropriate" for the platform.

Besides, 3 respondents associate the Product function with the application of legal standards and guidelines present in the legal system in which the platform is inserted. This complement, in turn, approaches the Sanitization function; however, it does not fit into this category, since the standards mentioned by the interviewees comply with more general rules present in the respective legal systems. Quoting Samuel again, interviewee who comments on the regulatory feature, in dialogue with the product function, by complementing his previous statement, as follows:

> (...) in this broader idea of adequacy, I am considering that it also encompasses **compliance with legislation, compliance with the promotion of human rights, compliance with ethical standards,** I understand that moderation has this function, to keep it appropriate to the function, to the interests of the platform. And then it will vary, it is, going to the limit, like, a pornography publishing platform, for example, could remove a Dr. Strange movie that was published on the platform and that does not have adult content, right? So, I would also call this moderation, not only for copyright infringement, to curb it, but also to adapt the platform's content to its purpose (emphasis added).

When talking about content moderation, most people immediately think of removal, as if these terms were synonymous. Although removing content from the web is one of the most controversial acts and, perhaps because of that, one of the most remembered, moderating includes a series of activities. Among them, some mentioned by the interviewees were: reduction of visibility (or downranking), labeling, shadowbanning,[33] suspension, ranking, invisibilization, banning and verification.

In this context, it should be noted that the terms Intervention and Curatorship did not reach an exact distinction in their understandings. Given this, the category "Moderation" is perceived sometimes as a removal synonym, sometimes in a more expansive sense.

Action: Intervention

As presented, intervention is here understood as acts of content moderation with automatic action and effect. In this scenario, respondents often associated such activities with the execution of a platform rule that prohibits certain types of content

[33] Also known as "phantom ban" or "shadow ban". It means a significant reduction of user published content reach, as if a ban had occurred, but without any platform notification. For Radsch, "'Shadowbanning' refers to a relatively common moderation practice of lowering a user's visibility, content or ability to interact without them knowing it so that they can continue to use the platform normally, but their content is not visible to anyone else". *In:* RADSCH, Courtney. Shadowban/Shadow Banning. In: BELLI, Luca; ZINGALES, Nicolo; CURZI, Yasmin (eds.). **Glossary of platform:** law and policy terms. Rio de Janeiro: FGV Direito Rio, 2021. https://hdl.handle.net/10438/31365. Although the objective is to remove potentially harmful content from the platforms, this practice affects the users' right to freedom of expression without giving them the possibility of contestation, as well as full defense and contradictory , since they are not notified about the imposed restriction.

or actions in their environment. Thus, Theo Freitas, a third sector member, mentions acts of intervention related to enforcing the platform internal rules and, therefore, to the Product function:

> It's... for example, **having rules saying 'this topic cannot be here'**, then, for example: a platform for movie reviews. And there's something there that isn't a movie review, so that... decides, let's say, to create a rule, to implement a rule for that. Yeah... it guarantees a multiplicity of different spaces and communities, which will be able to discuss different themes and will eventually be able to **exclude matters that do not have a pertinence with that** (emphasis added).

In face of intervention acts contemplated in content moderation, some interviewees mentioned the controversial issues that may come to light with its execution, such as wrongly removing legal content that is in line with the platform's rules. In this context, a recurring example relates to nudity cases, in which platforms automatically remove images they consider inappropriate, not taking into account cultural, artistic, generational aspects, among others, that attribute different meanings to their manifestation. As an example, some interviewees cited the case of the photograph entitled Napalm Girl, which can be understood as nudity content (and therefore automatically removed by platforms); however, it is a historical image, as it records a Vietnam War bombing. As the acts of intervention carried out by the platforms occur predominantly through artificial intelligence, appropriate content is often misinterpreted, given the limitations of AI (a matter further explored in the next topic).

Action: Content curation

Content curation stands out for contemplating activities that not only have instantaneous effects, but that often seek to model user behavior and/or to rank content, thus aiming for more complex and long-term effects. Sometimes it was mentioned by the interviewees as a type of action opposed to intervention, and in other cases, a gender and sub-gender relationship was established.

That said, content curation can serve both sanitization function and platform product function, although it is more related to the second, since it can be linked to different interests, whether they are meeting the internal rules of a platform or the current business model.

In this sense, Samuel Cardoso, a technical-scientific community member, discusses some Curatorship operation schemes:

> (...) in my opinion, content moderation encompasses **from content valuing and ranking by means of propagation limiting or propagation incentive limitations until access raising or restricting to certain content**. So, I don't know, when changing the profile picture or some congratulations reach more **diffusion amplitude** than a sad message, I understand this comes under content moderation, because the platform understands that this is more **suitable**, in this case, to keep people **accessing and using the platform services** (emphasis added).

In this sense, the content adequacy to the platform was also pointed out as being carried out by the curatorship when it comes to the regulatory feature. On the matter, Igor Peixoto, a business sector member, states:

> (...) but of course also from content moderation, when the platforms are **making these rules** and analyzing contents and removing certain contents, **decreasing visibility**, for there are several forms of content moderation to affect the content, they also end up, indirectly, in a way, **reducing the illicit materials incidence**, which is something that can generate **liability** – and then when I say illicit, it then **depends on the State, because they operate on a global scale, so in each State there is its own legal system**, there are different types of content that are considered illegal, so it's kind of like they already are **filtering out what might be considered illegal**. And even a common practice of the platforms, when they receive communication from the State, from either from the Judiciary, or, anyway, from the Executive, request for content removal or content analysis, they first analyze whether it violates any of the policies of the platform, and if it does violate, they remove it and say that, well, "oh, it violated the platform terms" and not necessarily the State legal rules and so on, and then this ends up impairing the analysis. (...) so, in a way it is also a means of **increasing the level of adherence to local legislation** (emphasis added).

The mention Igor makes to the reduction of illicit content through curatorial actions is related to the statement of Bárbara Silveira, a business sector member, when she says that

> The restriction **of a content reach** allows a more subtle approach to the types of **speech that may be considered unwanted**, coming to a better balance between freedom of expression and freedom of information. Ultimately, how attention is directed poses a key issue (emphasis added).

With regard to behaviors modulation by means of content curation, this topic is very much related to emblematic issues debated in the field, such as networks polarization and bubble filters disseminating disinformation and fake news. Thus, when platforms segment audiences and affect content reach, users are impacted in different ways, influencing social and political spheres. On the other hand, platforms benefit from achieving greater engagement in advertising content and/or a longer time of use by individuals, which, in turn, intensifies profit making.

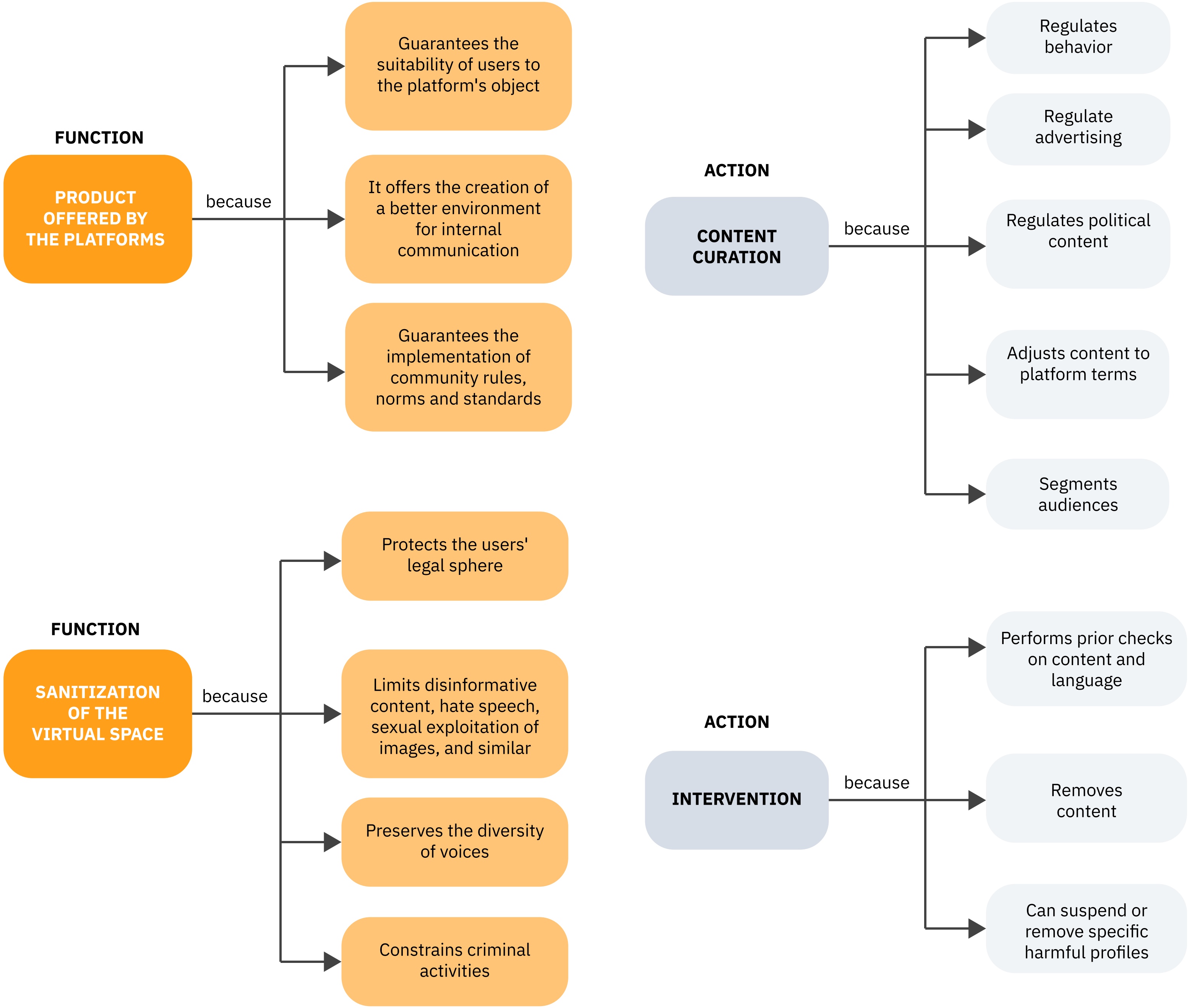
In this sense, Lucas Oliveira, a third sector member, discusses how Curatorship actions "moderate" behaviors in the following excerpt:

> Look, I guess content moderation goes **from removing a message, an image from within a digital platform, to eventually limiting the distribution of that message or choosing to segment which audiences will receive that particular message,** and nowadays you can expand even a little more and talk not only about content moderation, but also about **behavior moderation**, to the way people relate to these new technologies. So you **put a label** in a message to try to prevent people from believing a distorted message, instead of removing that message from the platform, sometimes it has much more to do with a **behavior moderation rather than actually with moderating that content** (emphasis added).

Finally, the more complex feature of actions considered as curation is very much linked to the fact that they are carried out only through Artificial Intelligence and there is not always transparency regarding the guidelines and methods used by the platforms. Some interviewees cited this aspect as a problematic issue, emphasizing the need for content moderation guidelines to be pointed and exemplified in platforms transparency reports.

Check out the mental map below, summarizing the main arguments presented in regard to content moderation actions and functions:

Figure 1 - Main arguments on content moderation actions and functions.



FUNCTION

PRODUCT OFFERED BY THE PLATFORMS

because

Guarantees the suitability of users to the platform's object

It offers the creation of a better environment for internal communication

Guarantees the implementation of community rules, norms and standards

FUNCTION

SANITIZATION OF THE VIRTUAL SPACE

because

Protects the users' legal sphere

Limits disinformative content, hate speech, sexual exploitation of images, and similar

Preserves the diversity of voices

Constrains criminal activities

ACTION

CONTENT CURATION

because

Regulates behavior

Regulate advertising

Regulates political content

Adjusts content to platform terms

Segments audiences

ACTION

INTERVENTION

because

Performs prior checks on content and language

Removes content

Can suspend or remove specific harmful profiles

4.1.2. On the use of artificial intelligence for content moderation

The use of artificial intelligence for content moderation purposes is already a reality for most digital platforms. One of the main arguments in favor of its use was precisely the fact that due to the size of most platforms, which can have up to millions of users and, consequently, a large number of posts per hour, **the scale of the content produced (9x)** would not allow moderation to be done only by humans.

In topics such as child sexual abuse and exploration material, terrorism, nudity and copyright, for example, it was recognized that **automated moderation would have a high degree of efficiency (3x)**, having as one of the reasons the sharing of databases between companies. However, one of the most pointed problems was **the failure of these technologies to analyze context and/or language (7x)**. The position of Samuel Cardoso, a technical-scientific sector member, illustrates well this concern and possible consequences:

> I think that, as we are talking about very small or very large platforms, the issue is contextual and the **automated mechanisms, they fail precisely in detecting finely tuned contexts**. So, I am very concerned about the low percentage being legitimized as a degree of efficiency of these mechanisms, because when we talk about millions of communications, **this low percentage can mean hundreds of thousands of errors, whether false negatives, but mostly false positives.** So I think automated removal or any kind of automated moderation **should be seen more as a precarious procedure, in the sense of something that can be circumvented more quickly from a request**, eventually based on the very complaint or notice by the users themselves. (emphasis added)

In fact, the **possibility to challenge or appeal automated decisions (2x)**, the diversification of contestation channels, as well as the **need to ensure a human review (4x)** appeared as alternatives to mitigate the risks of these tools. For Emanuella Nogueira, third sector member, it is necessary to worry about the correct use of automated moderation more than whether we should use it or not:

> (...) it seems to me that we need to advance what is the correct, adequate use of these mechanisms more than, well, if they should be used or not, because **they have a series of problems and the way to deal with it is to think about these processes and what should be done to improve their relationship also with humans**, so to say. (emphasis added)

In this sense, the interviewee points out the need to carry out an impact assessment on human rights, and concerns with diversity and with the process of these systems development and monitoring. Furthermore, **also mentions the need for greater transparency, adding to other 4 interviewees**, among which it was also defended the provision of precise information on the criteria and procedures of automated decisions, their degree of precision and greater explainability to users, even to gain their trust.

Such a solution meets one of the negative arguments presented around the use of AI, which claims there is little information about the reasons for any moderation and about the review process. Considering the **possibility of biases in these systems was also ventilated by the participants (2x)**, such as, for example, due to the use of already biased databases or by those who designed the algorithm, it is possible to conclude that greater transparency could also help in these cases.

Other initiatives were also mentioned, such as auditable automated content moderation, with explicit anti-discrimination criteria, **moderation teams better educated about contexts and regional specifics (2x),** carrying out risk assessment and the use of artificial intelligence for the purpose of automating the reporting processes and evaluation of content reporting. For Igor Peixoto, a business sector member, it is interesting to prioritize the use of AI in situations where there is a high degree of certainty that a specific content violates community policies.

According to the interviewee, one of the problems with automated moderation is the belief that technology would be the solution to all problems:

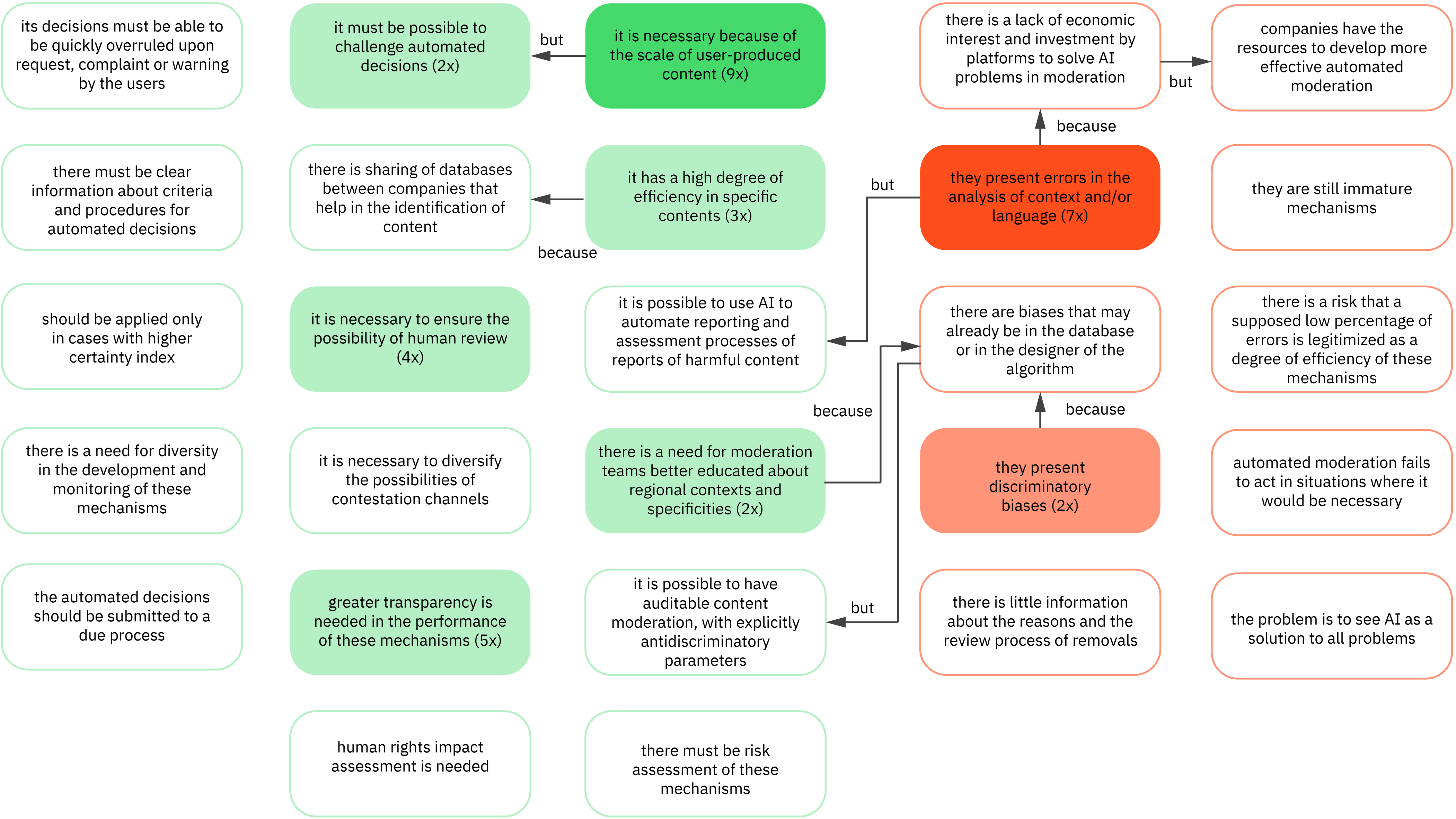
> Also what I see, that I already mentioned, is this **perception that artificial intelligence solves all problems**. “Oh, they can develop a technology that does that”. It’s not that simple and we know that technologies, as I mentioned, they make a lot of mistakes, so we can not be completely dependent on them. (emphasis added)

Caio Fernandes, government sector member, reinforces this perception by stating he considers the automatic mechanisms for moderation are still immature. For Luiz Porto, in some cases it is not even necessary to talk about decisions contrary to the exercise of fundamental rights and freedoms, but about an automated moderation that stopped acting in situations where it would be necessary.

The interviewee states it is also necessary to ensure that automated decisions “follow fair and due process and that through the use of artificial intelligence, artificial intelligence itself is subject to the scrutiny of due process of law as they are established”. However, it is necessary to consider the point made by another participant, in the sense that, currently, there would be little investment and economic interest on the part of platforms in solving problems related to content moderation by AI and making it more effective, even though they have resources for that.

Check out the mental map below, summarizing the main arguments presented in relation to automated mechanisms evaluation:

Figure 2 - Main positive and negative arguments regarding the evaluation of the use of automated mechanisms for content moderation.



its decisions must be able to be quickly overruled upon request, complaint or warning by the users

there must be clear information about criteria and procedures for automated decisions

should be applied only in cases with higher certainty index

there is a need for diversity in the development and monitoring of these mechanisms

the automated decisions should be submitted to a due process

it must be possible to challenge automated decisions (2x)

there is sharing of databases between companies that help in the identification of content

it is necessary to ensure the possibility of human review (4x)

it is necessary to diversify the possibilities of contestation channels

greater transparency is needed in the performance of these mechanisms (5x)

human rights impact assessment is needed

it is necessary because of the scale of user-produced content (9x)

it has a high degree of efficiency in specific contents (3x)

it is possible to use AI to automate reporting and assessment processes of reports of harmful content

there is a need for moderation teams better educated about regional contexts and specificities (2x)

it is possible to have auditable content moderation, with explicitly antidiscriminatory parameters

there must be risk assessment of these mechanisms

there is a lack of economic interest and investment by platforms to solve AI problems in moderation

they present errors in the analysis of context and/or language (7x)

there are biases that may already be in the database or in the designer of the algorithm

they present discriminatory biases (2x)

there is little information about the reasons and the review process of removals

companies have the resources to develop more effective automated moderation

they are still immature mechanisms

there is a risk that a supposed low percentage of errors is legitimized as a degree of efficiency of these mechanisms

automated moderation fails to act in situations where it would be necessary

the problem is to see AI as a solution to all problems

4.1.3. On the need for human review of automated decisions

This topic corresponds to one of the questions that derived from the questioning around the use of automated content moderation mechanisms. People were encouraged to express their views on how they evaluated human review in relation to the use of these tools.

In this sense, also considering those who had already expressed themselves positively in the topic above, there was a wide manifestation **defending the possibility of human review of automated decisions (12x).** In addition to being considered **necessary (6x),** it was pointed out as important for cases in which there are decisions capable of affecting the user's sphere of rights.

In this way, automated moderation could act as a content prioritization mechanism, **performing a first analysis and identifying potentially harmful content (3x),** which would then be forwarded for human review. Such an argument goes along with the respondents who advocated **AI and human moderation activities to be used in a complementary way (2x).**

In this context, it was also argued that human review should be a guarantee for users who **disagree with an automated moderation decision and resort to the platform for review (3x)**. Moreover, it could act as a second step to bar what had eventually escaped the filter of AI mechanisms. At this point, two interviewees argued that there was a loss in not ensuring the mandatory human review through the Brazilian Data Protection General Law (*Lei Geral de Proteção de Dados*) – LGPD.

Another challenge, pointed out by Guilherme Araújo, from the private sector, was that there would be content categories very dependent on context, so there would not yet exist a good and efficient enough technology to dispense with the need for any human review. On the other hand, Lucas Oliveira, from the third sector, pointed out that it is difficult to ensure human review in a scalable way, considering the large volume of information available on the internet all the time. Nevertheless, he pointed out that it is up to not only civil society, but also to the platforms themselves, which have the resources to do so, to conduct research and develop products in this regard, seeking to

> (...) how to insert a human review procedure in decisions that were taken solely and exclusively by algorithms in an automated way, so that we have this control, this second guarantee that in fact that decision it's correct, and unfortunately today we still don't have that guarantee and it ends up leading to distortions (...)

However, among the negative aspects mentioned around the human review, there was concern about the **psychological well-being of workers involved in content moderation (2x)** and the violation of their human rights, due to prolonged exposure to harmful content. It was argued this to be a sub-job, often delegated to developing countries, due to the possibility of companies paying less to these employees and not having to deal with the consequences of this activity. Furthermore, there would be little information on the functioning of moderation teams and their mechanisms, including those for remedying possible damage from human review work.

For Lara Souza, from the third sector, it is also necessary to think about how to insert this debate into the regulation of content moderation:

> How can we – within this discussion, if we are debating the regulation of content moderation, what are the rules and so on –, **how can we also somehow foresee certain guarantees of the rights of those people who will be making this regulation, since we are already seeing – in such a short time that this activity exists – effects within the lives of these people, that I don't think they're frivolous?** So, they are not irrelevant, what I mean. And so, if we already see such a big effect of this circulation of harmful content within the platforms, interfering in our democracy, interfering in our mental health, interfering in the polarization, in the political radicalization – if we already see this within the users who are under moderation, that is, users who are not, well, they are not seeing this type of content because it was moderated, but for it to have been moderated, maybe someone else has seen this content, right? And so, the effect it will have, psychologically, on this group of people, I think this does have to be debated. (emphasis added)

Finally, Theo Freitas, from the third sector, pointed out that only guaranteeing moderation by humans would not solve all the problems of moderation itself, alleging that the claim for this agenda would only pass through a kind of legitimization of the decision made by automated systems.

4.2. On the role and performance of the different internet governance actors involved in content moderation

In this topic, there are the interviewees' answers about their perception regarding which actor would be the most responsible for the management of content moderation, this activity oversight mechanisms available to civil society, evaluation of private instances of moderation decisions review, and the performance of the technical-scientific sector and the third sector. Questions 4 to 7 of the script are contemplated here.

4.2.1. On responsibility for the management of content moderation

When asked about the agent[34] of Internet Governance who holds greater responsibility for the management of content moderation, the vast majority of respondents stated that it was only the private sector, represented by platforms and large technology companies. Thus, the responses were divided between **platforms only (15x), platforms and government sector (1x), government sector only (1x) and joint responsibility across all sectors (1x)**.

The two main arguments used to discuss the power of platforms in terms of content moderation were: availability of **resources (7x) and competence (12x)** to manage such activity. With regard to the first argument, it is understood that such resources cover a series of elements leading such companies to an advantageous position. Among them: i) economic power, which enables the acquisition of qualified technology and, therefores, the ability to deal with the virtual environment large flow of information; ii) position of expertise gained in conjunction with technological development; and iii) regulatory margin, which allows freedom in creating its own rules and guidelines – given the absence of specific guidelines coming from the government sector.

Given this, the relationship between the resources belonging to the platforms and their elements is illustrated by Igor Peixoto, a business sector member, as follows:

> There is the question of **scale. I think it would be very difficult to transfer some of these responsibilities to other entities**, including public entities, because there is an immense amount of content being posted and these platforms, like, mainly the large platforms – because we can also differentiate between the large, medium, small, large platforms – they have a lot of resources also

[34] Among the private sector, government, third sector and technical-scientific community.

> to be able to deal with this issue of scale. And I would also say a lot of expertise, because when dealing with this volume of content over the years, **they learned to deal with greater efficiencies to make the system as a whole work better**. (emphasis added)

In complement, the interviewee also pondered how the platforms, based on their activities and time of operation, hold an amount of information that would be unattainable by other agents. Therefore, there is an informational asymmetry and an impossibility of transferring *know-how* to other sectors.

As far as the second argument presented by the interviewees – the competence of the platforms –, the point was exemplified from the notion that the content moderation management would be a natural attribution of such companies.

Thus, given the position of hegemony and regulatory flexibility occupied by such agents, the implementation of moderation guidelines would be an inherent responsibility of their activities. In this context, Antonio Cavalcanti, government sector member, states that "content moderation is an activity **proper of the network application layer**, that is, of the companies that make content available directly or through third-parties" (emphasis added).

Still living up to this thought, Theo Freitas, third sector member, exemplifies how he understands content moderation management as a responsibility inherent to the activity of platforms as follows:

> Ultimately, I see this activity as an activity of the responsibility of the platforms. Of course, maybe the government has a role, the third sector has a role, the academy has a role... But, in this case, – **just as, for example, the net neutrality of Marco Civil has directly to do with the activity of the telecommunications companies, because they are the controllers of the structure – this is an activity that has totally to do with the performance of the platforms.** (emphasis added)

In addition, it is possible to state that the arguments regarding the availability of resources presented by the platforms and their competence in managing content moderation are directly related. In this sense, Lucas Oliveira, third sector member, mixes both elements when discussing the reasons that, in his opinion, make the private sector the main responsible for content moderation management, when he says:

> (...) First because **they created their own platforms and they know the platform**, including from a **technical point of view**, then they are in the best possible position to moderate those

> services. So when, eventually, a new problem arises, as was the case during covid 19 pandemic, then you have a new phenomenon and we now have to face misinformation about the vaccine, misinformation about the lockdown and about the virus itself as well. **Platforms need to update their moderation protocols, their own standards and rules for using that particular digital space and they have to do this, because they are in the best possible position to do this, and end up also receiving most of the responsibility for moderation precisely because they are there at the forefront of the battle against harmful content.** (emphasis added)

Regarding the issue of legal provision and its relation with the competence of platforms, the relationship developed was in relation to the absence of express guidelines as to the role of the private sector in content moderation and, therefore, the emergence of a large margin of action. However, Fernanda Rezende, third sector member, defended that, despite the absence of a specific law for the definition and application of rules in the field of content moderation, there are legal frameworks in the Brazilian legal system that attribute to platforms the responsibility for the management of such activity, even if indirectly. It would then be the case of the Consumer Defense Code, which characterizes companies as service providers and, therefore, submits them to its liability regime. Fernanda also commented on the regime ensured by Marco Civil da Internet and *Lei Geral de Proteção de Dados*, which would be intrinsically linked to content moderation in view of data collection and processing actions.

Furthermore, the platforms' hegemonic position and their power in managing content moderation before the other agents that make up Internet Governance were also commented in a negative way by some interviewees. In this sense, Caio Fernandes, a government sector member, considers that

> sometimes the big platforms are a **distortion to the original internet, because they concentrated power, concentrated users.** And – because we are lazy to look for things in our... self-indulgent, right? –, so, we end up accepting this kind of thing, but that is not good. It is not good due to the enormous power that they... amass in this process. (emphasis added)

On the other hand, the issue of the platforms' protagonism was also commented on from a different point of view. In this sense, Guilherme Araújo, from the private sector, pointed out that

> There is a very modern debate about whether this responsibility should be the sole responsibility of the platforms or not, but the fact is that – the way it works today – **the responsibility lies solely and exclusively with the platforms. With all that is good about it and all that is bad about it**. (emphasis added)

In view of the high concentration of platforms power, some interviewees positioned themselves as to the **need for a more imposing action on the part of the government sector (7x).** In this sense, Sofia Pires, third sector member, highlights the need for incidence and regulation by the government in favor of mitigating the negligent and abusive character of the private sector actions. In addition, Augusto Cruz, government sector member, argues that platforms themselves should act in moderation; however, it is necessary that National States regulate it in some way, drawing maximum and minimum parameters.

As for the position that platforms share responsibility for managing content moderation with the government sector, Luiz Porto, a technical-scientific community member, presented reflections on the different powers and actions that private and government sectors have as decision makers. In this sense, he points out the State stands out for having an unique decision-making power, attributed by the jurisdiction. From this, it enjoys legal mechanisms that platforms, as representatives of the private sector, lack. In contrast, the large conglomerates have a corporate power of a transnational nature, which guarantees the power to elaborate and apply internal policies; although it is subject to government power.

With regard to the position that the government sector would be the main agent in managing content moderation, Maria Teixeira argues:

> Look, I think **the Powers of the Republic, in our case, share this greater responsibility**. Also because I do not believe that the Self-Regulation Council of the platforms would work (...) I think that the Powers of the Republic need to share this responsibility (...) There needs to be a **development of bills for us to guarantee first the transparency of this content moderation, second the standardization of this content moderation, and to always have the guarantee that it will happen according to the legislation.** (...) I think that the three actors of the Republic are the three most important powers, because they have the power of the **legal constraint**, so to speak, over the platforms. (emphasis added)

Finally, there is the position presented by Bárbara Silveira, representative of the business sector, who argues in favor of a joint responsibility of all sectors regarding the management of content moderation. Bárbara states that, with regard to the role of platforms, it is up to them to look for innovative ways to deal with issues that can overcome the binary approach of removing or maintaining content, and they must also operate directly in changing behaviors or even in prevention through incentives or disincentives.

With regard to action by the government sector, she says that “governments can and should prioritize policies, partnerships, domestic and foreign investments that support digital literacy and defend the Open Internet”. Finally, she argues that society also has a fundamental role in terms of media education for people, “since disinformation only works if there is an audience that engages with it”. In addition, society would also be responsible for overseeing the progress of efforts by governments and companies, besides offering contributions to improve such efforts and demanding that they are in line with international human rights standards.

4.2.2. On possible means of monitoring the content moderation process

This topic aims to analyze the interviewees' speeches when asked about the existence of mechanisms to guarantee the monitoring, by society, in the face of content moderation practices. That is: Are there ways and resources that allow society to monitor how content moderation practices are being performed?

The vast majority of respondents were skeptical of the efficiency or even the existence of such mechanisms. Thus, the observed classifications were: **non-existent (7x)**; **existing, but insufficient (6x)**; **existing, but with reservations on their efficiency (3x);** and **sufficient (1x)**. Finally, one respondent only stated that they are necessary. Among the examples of means used to monitor content moderation practices, **transparency reports were the most cited (6x).**

Thus, the main arguments presented to defend the non-existence of such mechanisms were: the absence of regulatory guidelines that oblige the platforms to develop them and the discretion and freedom of these agents when creating them, since they have an opaque and selective character - and therefore cannot be considered as a means of surveillance by society.

In this sense, Ana Carvalho, member of the third sector, states:

> So, I don't know if it exists, I think **there is no mechanism that guarantees with absolute certainty that civil society will be able to monitor what happens there**. Because, at the same time, they will only divulge what matters, what they want us to know. So, after all, the power over this content moderation activity is with them, with the companies, and very little with us in this monitoring capacity. **As much as we do this immersion in transparency reports, terms of service, privacy policies... I don't think there is a mechanism, even because moderation activities are very unexplained, or not explainable, at least what we see in practice**. There's not much of a "you violated this term of service rule and as a result your account is being suspended for a week" kind of thing. This doesn't exist, it doesn't happen. So we are very much at the mercy of companies. (emphasis added)

In addition, Theo Freitas, also a third sector member, discusses the absence of express guidelines regarding the creation of such inspection mechanisms as follows:

> It would be better if we had a clearer rule saying what can and cannot be done in this activity, and even better would be if we

> had an **independent body or activity that could not get into the discussion of content** – because I think this is very dangerous –, but that could get into this more systemic analysis, let's say, of this activity, along the lines of the DSA, possibly, you know, in European legislation, it's possible. (emphasis added)

In this sense, Samuel Cardoso, a member of the technical-scientific sector, comments that not only in Brazil, but all over the world, there is a tenuous, light and permissive legal regime in relation to the requirement of transparency and accountability. Therefore, there would be no way for society to know whether the appropriate practices regarding content moderation are actually being carried out.

Still regarding the regulatory issue, Stephany das Neves, a technical-scientific community member, points out that "we are still at an earlier stage than that" with respect to discussion and analysis on the subject. In this context, she points out:

> (...) I think that when Marco Civil brought the rule of liability exemption from the **court decision, but without having brought the good Samaritan rule, we ended up not bringing this culture that platforms, on the one hand, need to moderate and can moderate as if it were a faculty of duty**. More than a faculty of duty, it is something almost inherent in the digital service itself. Moderation is part of this service. I think this is not clear in our legal culture, I think **the fact that Marco Civil was silent on this didn't help either**, but I think it's not just Marco Civil, maybe also we as an academic, scientific, technical community, all of us. This topic has just arrived, but we have almost ten years of Marco Civil, 2014, we will do it in a year and a half, in 2024. So I think that **we didn't discuss this topic enough and with that I think in fact we didn't even... it's an activity that is accepted today as something completely normal**. (emphasis added)

With regard to the interviewees who understand the monitoring mechanisms as existing, but insufficient, the main argument to defend such a position was regarding the opaque nature of the manifestations, policies and transparency reports disclosed by the platforms, so that there is no real transparency. Thus, several criticisms were directed at the transparency reports, specifically, such as the comments made by Lucas Oliveira, a third sector member:

> Look, there are some mechanisms, okay? But we still depend a lot on the goodwill of the platforms themselves and that is a problem. What do I mean by that: **these are the platforms that will eventually publish a transparency report, where you will have information about content moderation practices from a systemic point of view and not on a case-by-case basis.**

> (...) And a lot of what we have are the platforms themselves that are making this available on a voluntary basis, so we still need to make a lot of progress in terms of transparency in content moderation. (emphasis added)

Another point presented by Lucas Oliveira, in turn, was regarding the role of the State in demanding "complete and accurate" information from the platforms, since not all of them are of interest to the public. From this, it would be possible for users to achieve a more proactive and demanding attitude towards the private sector.

Antonio Cavalcanti, a government sector member, also opines in favor of the insufficiency of the mechanisms offered by the platforms when he says that:

> (...) although some companies have their internal verification systems, including accepting complaints and appeals against their decisions, these procedures **are neither uniform nor transparent for most users**. Most of the time, the response to these appeals is inaccurate, vague, unsatisfactory or even non-existent. (emphasis added)

As for the interviewees who argued in favor of the existence of monitoring mechanisms, but with reservations regarding their sufficiency, Guilherme Araújo, a private sector member, defends that such means are still in the process of being developed and, although we have not reached the state of the art (*sic*), the current context would be significantly better than that of 4 or 5 years ago. On the other hand, Vicente Moura, from the government sector, says that the sufficiency of monitoring mechanisms depends on a "combined action". In this sense, he says:

> I think that a first step are transparency reports about the number of content withdrawals, the moderations broken down by type, but maybe also the proceduralization of these processes within the platform itself, so eventually **give the user the chance to justify why that post should or should not be moderated, both on the side of the person who reports and on the side of the person who has the post moderated. I think those would be the instruments.** (emphasis added)

Finally, Bárbara Silveira, a member of the business sector, says that inspection mechanisms exist, but they can be improved. In this sense, she points out that one of them is transparency, which guarantees accountability on the part of companies and governments and, therefore, liability. Still, she adds that one area for improvement is in ensuring "laws governing information provide adequate flexibility for valuable disclosures, for example, the provision of data to academics and researchers."

Finally, the arguments presented in favor of the existence and efficiency of means of monitoring in the face of content moderation practices focus on mechanisms of an extrajudicial and judicial nature in the context of protecting internet users rights. In this sense, Luiz Porto, a technical-scientific community member, states:

> (...) and then looking specifically at whether we observe the possibility that there is **collective defense.** I am thinking of class actions, because it is through collective redress that civil society seeks forms of protection. Either individually, through the **public civil actions, through the intervention of third-parties, or by participating in public civil actions with the state, federal Public Prosecutor's Office**, depending on the competences that are brought into consideration. (...) Or judicial administrative, by virtue of the jurisdiction of the state, or moral, when there is, for example, **collective boycott of the use of certain services or certain products, thinking about the condition of user, consumer, customer, citizen,** yeah... so, it would be another shape within a **monitoring consequence**, but today I think it is important to realize that there are more or less strong mechanisms in the field of collective redress, and that depends, because it is often not in the interest of the Public Prosecutor's Office to follow these issues. (emphasis added)

It is thus observed that Luiz Porto interprets the means of moderation from a different perspective than that taken into account by the other interviewees.

4.2.3. On private instances of content moderation review

With regard to private instances of content moderation review, it was considered private initiative bodies that propose to act as an external reviewer in the face of decisions taken by platforms, mainly. Given this, when questioning respondents about their positions regarding such instances, the Oversight Board (OB), a body created by Meta, was used as an example to illustrate the question. In this sense, the vast majority of respondents **adopted the OB as the main reference when discussing the topic (14x).** In addition, it is worth mentioning that **some of the interviewees made reservations at the beginning of their speeches (5x), indicating**: either they do not closely follow OB's decisions; or they have in mind that few decisions have been taken and, therefore, there would not be much to conclude at that moment; or they do not know much about the body.

Thus, when questioned about their position regarding private instances of content moderation review, most of the interviewees stated that they have a **predominantly positive performance (12x),** against a minority that defended the **predominantly negative performance (5x)**, and one respondent who did not take a position. In a complementary way, it was also questioned whether the performance of such instances would guarantee a **greater legitimacy in the face of content moderation activities, to which part of the interviewees said yes (6x)**, some were in the **opposite way (3x)**, a portion opined for **undetermined legitimacy (4x)** and the final installment **did not comment on (5x).**

Starting with the analysis of the responses, regarding the main arguments presented to defend the positive performance of private instances of content moderation review, they were: the initiative in favor powers deconcentration at the time of decision-making and the guarantee of more transparency in content moderation procedures.

In this sense, Theo Freitas, a private sector member, argues:

> I evaluate positively. I think it was a step... I think it was an important step. It is a **certain experimentalism to deal with it**. Undoubtedly, there is a discussion about the extent to which you are outsourcing responsibility, but I don't think this is necessarily a bad thing, because at the limit of the limit, you are not outsourcing responsibility for industrial activity, you are **kind of trying to help in the most difficult cases,** gains transparency... and is a **gain of transparency (...)**It is a gain in transparency in the justification of decisions, not in the decisions themselves. In this sense, I think it is a gain. (emphasis added)

In addition, Lucas Oliveira, a third sector member , argues in favor of the Oversight Board, emphasizing the legitimacy it ensures to Meta's decisions by decentralizing the decision-making procedure:

> **I think that, therefore, it is an initiative designed precisely to try to bring more legitimacy to the content moderation process. What you had before was basically a concentration of powers, so, senior executives in the company, they had the 3 powers in 1: they were legislators, judges and administrators of the platforms at the same time, they performed these 3 functions.** (...) **And Facebook's Oversight Board is one of those separation of powers initiatives, because you create an arm that ultimately reviews content moderation decisions that is independent of the company,** so it's like if you take away a power that was previously concentrated in the top executives of the company and now you give it to a body, an independent committee of experts that will ultimately make that judgment and **the company is obliged to respect the final decision**. This gives

> more legitimacy to the content moderation process precisely because you start to decentralize this decision-making a little. (emphasis added)

Stephany das Neves, a technical-scientific community member, presents similar arguments in defense of the guarantee of greater legitimacy ensured by private review bodies. In that sense, she says:

> **It is natural that, for the decision to be more legitimate, it should not be taken only by someone from the company, as if it were the CEO or the director,** legal director or something. But that actually has more legitimacy, especially the **legitimacy given by the procedure itself**. Perhaps the legitimacy given by **representativeness of this council**, because it is external and the procedure itself, I think it can bring legitimacy. So it certainly has a positive impact. (emphasis added)

With regard to ensuring greater transparency from the actions of the Oversight Board, Igor Peixoto, a member of the private sector, argues that, although the recommendations published by the body are not binding, such guidelines give greater visibility to content moderation activities. Thus, if Meta fails to apply them, there will be no determination of sanctions by the body; however, there may be pressure from affected groups.

With regard to those interviewed who consider the performance of such instances predominantly negative, the most recurrent arguments were **autonomy or performance limitations (12x)**, **little transparency in decision-making and creation of guidelines (3x)**, **low impact of their activities (5x)** and **high cost for implementation and operation (2x)**. Such arguments, in turn, were also presented by some interviewees who positioned themselves in a favorable way to the performance of such instances, being indicated as points for improvement or simply negative criticisms.

On the limitations of autonomy and action, the interviewees spoke about the creation of private instances being initiatives of the platforms themselves, which, in turn, would prevent impartial decision-making. In this sense, they also pointed out the employees themselves are chosen by the platforms and, recurrently, such bodies, despite being considered external, are linked to the companies in socio-political and reputational issues (preventing independent action).

In this area, it is possible to observe that, even among the interviewees who are in favor of the performance of these instances, such criticisms are present. This can be illustrated by the speech of Fernanda Rezende, member of the third sector, when she says that, despite being positive, they are insufficient agents, for they "are strictly **related and organized and directed by the company itself"** (emphasis added).

Likewise, all respondents who argued in favor of the negative performance of decision-making review bodies also believe that such bodies do not guarantee greater legitimacy in the decision-making process in the content moderation field. In this sense, the speech of Samuel Cardoso, a third sector member, illustrates these issues well:

> I think that the affectation of legitimacy, so to say, **there is a risk that these bodies work only to clean up the image, in the sense of the company doing something as a facade and not actually being implemented**, and maybe also **due to the status of composing a body, some members, in prestige to the company, end up not imposing their decisions**. I think there is an eventual chance that this result will be positive, but I think there are not exactly guarantees, so I think that even though there were spaces for a more efficient decision, it is very possible that it will not happen like that. I think that eventually among companies, even when they open up these spaces, they may be giving legitimacy to hope that in fact they sell some more honest positioning like that, more sincere, you know, **but I think that the structure and dynamics of formation of these groups end up limiting the legitimacy they have to can**... propose, suggest measures more drastic or proportionate to the seriousness of the situation. (emphasis added)

Samuel's comments, in this key, find common elements in the speech of Antonio Cavalcanti, a government sector member, when he says:

> These private instances – however they are a demonstration of good faith in addition to playing an important role in the perception that platforms are 'doing something', so to speak – **do not enjoy sufficient transparency and legitimacy to gain the full trust of users**. (emphasis added).

A recurring issue commented both by the interviewees who consider the performance of such instances positive, and those who consider it negative, was the low impact achieved by their activities. Thus, given the volume of demand that the large platforms have, external bodies would not be able to "solve the problem as a whole", as stated by Ana Carvalho, a third sector member. In addition, the interviewee considers that, in relation to the Oversight Board, despite the body being a good initiative, only the paradigmatic cases would be considered by the body, ignoring a large volume of commonplace cases and maintaining controversial measures and possible offense to freedom of expression.

Along the same lines, Guilherme Araújo, a private sector member, opines:

> **(...) this is better than nothing**, we... but I think that **we are still not in a position to assess the effectiveness of this body, either because it judged a few things, or because in fact we have little production**, oh... on how it is working – not in retail, we know that some measures were reversed, others were maintained, but in wholesale, I don't think we are still in a position to make an assessment on this model. (emphasis added)

In this context, it is worth mentioning two interviewees who commented on this bodies' features arguing in a different sense. In their view, there would be a misunderstanding of people about the proposal presented, so that the objective would not be to re-analyze all cases on a large scale, but rather to draw great guidelines precisely from the analysis of more difficult or paradigmatic cases. In defending this point of view, Theo Freitas states that there is "a **dissonance on what people want from a body like this [Oversight Board] and what these bodies really want to offer**" – which, in his view, would be more clarity and transparency to the platform.

In addition, Igor Peixoto, a private sector member, comments on the impact of the Oversight Board and its criticisms, seeking to explain the issue of (non) binding with regard to published recommendations and decisions:

> And I also think that another aspect that aroused some distrust in relation to the real impact of the Board is the fact that, the recommendations, they are not binding. And that means that **the company is not compelled to accept these recommendations. Decisions to remove content or not are binding, but recommendations are not.** However, despite this, the company, Meta, needs to publicly reveal which recommendations it will implement and also which ones it will not implement, and the reasons behind the decision to implement a decision or not to implement it. And I understand that only in this exchange, like, **the amount of information disclosed about the internal functioning of the moderation activity as, this is something positive, from the point of view of transparency, which I would say is a unique means of this relationship between the Board and Meta, so far, there is nothing similar on other platforms.** (emphasis added)

Finally, a criticism of the Oversight Board appeared only once, but it is directly related to other issues scored. It deals with how the engagement between the OB and external actors takes place, an issue commented on by Lucas Oliveira. For the interviewee, the body has faced difficulties in "talking" to the public comments made by the interested public (which can be an institution or any individual in their personal capacity). He also comments that this problem is not exclusive to the Board, but recurrent in higher courts around the world, such as the Federal Supreme Court, by not taking into account the manifestations made by amicus curiae in public hearings. Overcoming this issue, according to Lucas, would give greater legitimacy to the body, while decisions would be constructed collectively.

4.2.4. On the role of Judiciary as a possible instance for judicial review of decisions involving content moderation

In this topic, there are the perceptions of the interviewees about the role of the Judiciary as a possible instance of judicial review for content moderation decisions. This questioning was made due to the possibility of users seeking access to justice to reverse an eventual content removal or account suspension, for example.

In this sense, Marco Civil da Internet was cited as one of those responsible for making the Judiciary an **essential party when it comes to content moderation (3x**), due to what is foreseen in article 19, which establishes the liability of intermediaries for content published by third-parties only in the case of noncompliance with a court order, except for specific legal exceptions. For Sofia Pires, from the third sector, the action for these cases would be fair and necessary, having this Power as a "third-party, capable and proportional enough to decide on these cases".

Guilherme Araújo, from the private sector, pointed out that it is common for judicial orders to be sent to platforms to reverse decisions involving moderation, and that the largest number of cases of dispute on the topic would be in Brazil. This affirmation is in line with arguments raised that the Judiciary acts in the **solution of conflict or restriction of rights (2x)**, analysis of facts with legal repercussions, such as content moderation, in the **repair of damages eventually caused by this activity (2x) and for correction of illegalities (2x)**. In this regard, the position of Stephany das Neves, from the technical-scientific sector, is highlighted:

> So I think it has a role, yes, **which is to identify certain types of excess in moderation, either because it has removed too much content or because it has not removed undue content that it should have removed,** I think the judiciary has a role, yes, to actually correct illegalities.

For Lara Souza, from the third sector, this institution would also play a fundamental role, especially during election periods, in moderating politicians' posts, contributing to a greater legitimacy of the moderation itself. Yet, another argument presented in the interviews was that there would be no way to escape the actions of the Judiciary, due to constitutional guarantees of access to justice. At this point, Emanuella Nogueira, from the third sector, defended the impossibility of limiting this access, such as, for example, conditioning it to the need of a previous request for review to the platform. However, she points out that, in a standard scenario, with more regular situations, moderation should be able to function without the need to take provocations to the Judiciary, which could act only in the most serious cases.

This concern is even more relevant given the recognition that forwarding a large number of content moderation claims for judicial review **could worsen the overload that courts have already been facing in the country (3x).** Moreover, given the volume of information available on the internet, as well as the scale of moderation, it would be unfeasible to take all disputes to the courts, considering also that it would escape their competence to act as a reviewing instance of this activity. The comment by Lucas Oliveira, from the third sector, illustrates this perception well:

> (...) **the Judiciary cannot act as an instance of review of the content moderation process itself, it is not within its competence**. It cannot give tips on how content moderation works, what can be improved and what cannot. So, it works for some specific cases, when you have this collision of rights, to avoid even greater damage or to repair damage that has already happened, **but to improve the content moderation process as a system, then the judiciary can't do much.**

Another negative position was that the Judiciary is a slower institution that needs to be provoked in order to act in specific cases, which is why it has a limited role and cannot be the only alternative. Furthermore, in order to provide a better response to this type of litigation, **more dynamism would be required (x),** such as through the creation of special courts or specialized courts. This last alternative could also help mitigate another point raised by the interviewees, which was the **risk of granting inappropriate or not very technical judicial decisions (2x)**.

For Samuel Cardoso, from the technical-scientific sector, it is necessary that decisions on moderation have "another degree of caution because of the possible consequences of allowing or denying, granting or denying a certain request, even considering the possibility of the judge making a partial concession". Likewise, it was pointed out that it would be necessary for the judiciary to recognize the value of the content moderation activity in its decisions and understand that not all moderation needs to be avoided.

Still, it was mentioned that the Judiciary, despite having an important role, would be **one more actor in the content moderation equation, and not necessarily the main**

one (2x). In this sense, Augusto da Cruz, from the governmental sector, argued that he doesn't understand as a prerequisite for the removal of online content the need to resort to the Judiciary, but that as long as it is well grounded in the legislation, it would be more interesting to guarantee the user's rights or the remedy, by judicial means, when that is the case.

Furthermore, seeking access to justice would involve some costs to users, such as **financial resources and time to file a lawsuit (2x)**. The lack of a specific rule and prior debate would also hinder the performance of the judiciary, since it would increase the risk of decisions not having the expected effects.

For Caio Fernandes, from the governmental sector, it is also necessary that the creation of rules be based on a "local concept of justice and local culture of what can and cannot. On this point, Samuel Cardoso defends the importance of the legislative process for a better performance of the judicial power:

> (...) I think that it will be **more efficient if the legislative process of defining rules is better,** if they have already debated what the principles are, what the value that will be defended is, for each *[inaudible audio]* of moderation, for each situation, what must be restrained, what should be valued, because then, even if the decision is wrong, it will have to address this aspect, it will have to say exactly why it is not applying that designation, what is the factor that it is understanding that is distinctive and that authorizes it not to go in a direction that is determined by the legislation.

In addition, other arguments raised presented more options that would contribute to the role played by this actor: clearer policies and greater transparency about content moderation cases; insisting on mechanisms provided in PL 2.630/2020 ("Fake News Bill"), such as for users to dispute specific moderation decisions in which they felt aggrieved; and investing in independent review bodies, such as the Oversight Board.

However, Ana Carvalho, from the third sector, pointed out that it is important to call the Judiciary to solve problems involving moderation, since it would not be possible to "just accept that the company makes a decision more favorable to it, infringing fundamental rights such as freedom of expression and privacy". For Maria Teixeira, the importance of the judicial route is justified by the legal constraint, capable of making platforms act when necessary. So, its role would be acting together with the Legislative Branch, to enforce approved bills and update existing codes.

4.2.5. On the participation of the third sector and academia in regulatory processes for online content

This topic aims to gather the opinion of the interviewees regarding third sector and academia (or technical-scientific community) performance and participation in online content regulatory processes. Thus, both groups activities concerning content moderation through different actions were contemplated, such as political advocacy, academic production, scientific dissemination, participation in regulatory frameworks, among others.

In view of this, almost all the interviewees gave their opinion as to the **predominantly positive participation of the third sector and academia (16x)** – referring to both sectors together, one interviewee gave his opinion as to the predominantly negative participation of both, and one interviewee argued in favor of a positive action of the third sector and a negative action of academia.

Regarding the positive comments mentioned by the interviewees, the following were mentioned: **qualification of the public debate (10x)**, significant participation in the **legislative construction of regulatory milestones (6x)** – such as Marco Civil da Internet, Lei Geral de Proteção de Dados, and PL 2.630/20, putting **pressure towards platforms and government regarding the fulfillment of guidelines (4x),** and **intermediation** posture **between the sectors (1x)**.

Thus, in the sense of qualifying the public debate and mediating the sectors of Internet Governance, Igor Peixoto, a member of the private sector, points out:

> From the point of view of the regulatory discourse, in the most public sense of the word, it also continues to be super important, because it **qualifies the public debate**. We know that many times there **is a lack of technical knowledge on the part of parliamentarians, wanting to pull solutions out of the hat, things that we know don't work in practice, and that's why I think that this interaction with civil society and academia helps a lot to qualify the debate**, to take away some of these perceptions that some problems are too easy to solve. (emphasis added)

In this context, many interviewees commented on the role of the third sector, especially in legislative processes. In this sense, Fernanda Rezende, a third sector member, comments on the development process of Marco Civil da Internet, LGPD and the emergence of the Rights in Network Coalition:[35]

[35] "The Rights in Network Coalition is a network of entities that brings together more than 50 academic and civil society organizations in defense of digital rights, with the following main themes of action: access, freedom of expression, protection of personal data and privacy on the Internet"

> (...) the participation of civil society was quite large, **there was a very large mobilization of various civil society entities, and that even gave rise to the Coalition**. The Coalition emerged from this collective that was formed during the debates on Marco Civil da Internet. And the contribution of civil society was decisive so that, for example, Marco Civil da Internet did not come with a criminal bias, focused on criminal issues, with a punitive bias. Who transformed the initial project of Marco Civil da Internet – of a law that would regulate the rights on the internet – which initially was a project with a very penal bias, **we, from civil society, organized civil society, were the ones who contributed in a decisive way for us to arrive at Marco Civil da Internet, a framework that would establish principles, general rights, for the use of the internet**. And we have been evolving in this sense; **the Lei Geral de Proteção de Dados: the contribution of civil society was fundamental, not only from consumer defense entities, from communication rights, but for example, in the discussions of the LGPD, in the end, there was a great articulation of the entities from these sectors with business entities to reach a consensus.** I think we matured in this debate, all with the **contribution of the Brazilian Internet Steering Committee**. (emphasis added)

In complement, Antonio Cavalcanti, from the governmental sector, also comments on such regulatory frameworks, when he points out:

> The participation of the third sector and academia are fundamental. Just look at the contribution given by them throughout the history of the Brazilian Internet Steering Committee – CGI.br, in the construction of Marco Civil da Internet, of Lei Geral de Proteção de Dados – LGPD and, more recently, in the legislative process of Fake News Bill (PL 2.630/2020).

Guilherme Araujo, from the private sector, on his turn, agrees that PL 2.630/20 was a mobilization born from the civil society with a main focus on content moderation, although it agglutinated other themes along the time.

As for the position of the third sector and academia as actors who exert pressure on the private and governmental sectors, Sthephany das Neves, an academia member, says that they are dedicated to monitoring what is done by platforms in the content moderation field.

(information collected from the website https://direitosnarede.org.br/quem-somos/).

In this sense, Theo Freitas, a third sector member, makes considerations about the role of pressure exercised by such sectors and also of assistance to the private sector:

> In academia and civil society, what is the role, is... **I think it is twofold.** There is a role of **evaluating and pressuring the companies, so that they become better at this activity, so that they are more transparent in relation to this activity,** so that rights are not, you know... the academy, and civil society have... to be the talking cricket – let's say, conscience – that will indicate the contradictions between business models and solutions adopted, between rights at stake and business models, making this **triangulation**. So there is this place of talking cricket, but I also think that there is a place of **helping the private sector to inform society about** this activity, because society is very badly informed about this activity. Yeah... people analyze content moderation based on individual cases of removal... (emphasis added)

Despite the fact that almost all of those interviewed gave their opinion in favor of the positive performance of the third sector and academia, several criticisms were presented to the activities of both, with emphasis on the third sector. In this sense, the main ones addressed to this agent were: limited performance in terms of scarcity of resources, difficult access to confidential information of the platforms, low qualification in relation to specific themes and presentation of biased positions – since many third sector organizations are funded by large platforms. Regarding the academy, a specific criticism was presented in view of its isolated and not very proactive posture.

A criticism also presented by some interviewees was regarding organizations or people who, due to the flow of information and structure of the current internet, consider themselves momentary experts on hot topics, when, in fact, they do **not have the necessary expertise to talk about such content (4x)**. In this sense, Theo Freitas points out:

> The problem is when the third sector participates without, without... studying, let's say so. But I don't think that this is the rule, then it is not even possible to say that it is a third sector, then it is, I don't know, someone's guess. I don't think that there are organized groups that talk about content moderation today that don't know absolutely what they are talking about as a rule, this is the exception.

Caio Fernandes, a government sector member, agrees with this thought. Thus,on the topic, he comments that the possibility to express oneself was an achievement of the Internet, which is a collective and collaborative initiative - inherent elements of the multistakeholder structure of Internet Governance. However, he observes that, many times, groups of people participate in virtual space debate unproductively, generating noise and "being dazzled by the power of voice". Thus, they create damages to the virtual environment and to society as a whole. He demarcates, in turn, that in this speech he is not referring to specialized organizations dedicated to the study of specific themes, but to a general public observed in some spaces.

Regarding the limitations faced by the third sector and academia, one of the positions presented was that of Lucas Oliveira, a third sector member, who comments on how the difficult access to some information from the private sector can compromise the performance of the third sector and academia, since it becomes difficult to "decipher the content moderation black box". So a problematic scenario is configured, in which it is not possible to propose structural changes, since the procedures adopted are not published. An example given in this sense is the criteria used to define when moderation is done in an automated way and when it is done manually.

In this sense, Ana Carvalho, a third sector member, covers a series of limitations in its speech, when she says that:

> The technical and scientific community **has the most to offer in these debates, but it is the least heard,** along with civil society. They are the **weakest parties, they are the parties that have the least influence, they are the parties that have the least money,** in order to consequently have more power of influence... and then... at the same time, civil society is the party that is going to be most affected by this regulation process.

One limitation presented in a specific way by an interviewee was regarding the biased nature of the third sector's actions. In this sense, Vicente Moura, a government sector member, points out the vast majority of the organizations that compose civil society are funded by platforms. So, in his opinion, even if the research does not present an explicit direction, there is a kind of inhibitory effect (chilling effect) as to publications and research objects. This movement would then be an indirect and "extremely effective" lobby, which would also occur in attracting members of academia and organized civil society with direct funding for research, thus preventing a tougher stance on legislative processes and even practices adopted for content moderation that lack transparency. In light of this, Vicente opines that civil society should seek more independent means for conducting research, although he does not see this occurring in the near future.

Another topic contemplated by the interviewees was how the third sector and the technical-scientific community mobilize in the face of themes of interest. In this sense, some stated that these sectors have a reactive character, **mobilizing only in the face of focal episodes (7x)** which, in turn, would compromise the quality of their actions in a systemic way (since some themes would be over-considered and others disregarded). In this sense, Augusto da Cruz, a government sector member, links this movement to the functioning of social networks, which confer greater visibility and engagement to agendas on the rise. Therefore, he states, "Twitter itself will say, at some point, 'this thing will be a trend topic' and at some point everyone understands something should have been blocked, or someone else understands it shouldn't." This path, even if organizations adopt more vigilant attitudes towards other topics, the population as a whole will be interested in and give visibility to a few topics.

Theo Freitas, likewise, presents a similar position, defending the following:

> the agendas gain traction upon supporters, in social networks... you know, **things move forward when there are emblematic cases, which is a... sadness, because I think civil community organizations make an effort to take this issue outside the logic of, let's say, only discuss an individual case.** But the point is that they gain attention and mobilize when this **is on the agenda**.

Samuel Cardoso, a third sector member, comments on this limitation of the third sector as follows:

> (...) in the third sector, we see a positive performance, but it also ends up being limited because of the **profusion of themes**. So we end up having a **lot of things to take care of, in general, and few organizations, at the limit, right, in each national context, we have a number of issues that exceeds the capacity of action of these third sector organizations, which end up prioritizing what is on the agenda of the moment,** the problems that are being created by the technology companies or by the decision makers themselves, and there is a failure in the proposition of positive agendas, right? Agendas that proactively focus on these issues of moderation, right? So, to say a failure, for example, in this aspect, **there are some rules in Marco Civil da Internet that could have already been improved or complexified, but that in general, in these eight years, almost ten years, the third sector was not very proactive in raising these issues and promoting the debates and suggesting the improvements that would be necessary**. It ends up, in this point of view, being negative, acting very much in the **wake of the situation and** little **with the definition of a previous agenda of what needs to be done.**

In this scenario, Vicente Moura, who characterizes civil society as "reactive", comments that he misses a more proactive incidence in the legislative processes, "effectively proposing solutions, even if pragmatic, for the elaboration of provisions". He also adds that he considers scientific research, reports, and critical analyses about eventual regulation processes to be impractical.

Finally, Maria Teixeira praises the third sector in its actions and comments on the issue of mobilization from a different point of view. For her, the focus on major issues ensures a greater concentration of forces and, therefore, a more cohesive mobilization, bringing together more groups that are part of the organized civil society. Furthermore, she commented positively on the capacity of the third sector to insert itself and have a voice in institutional spaces.

4.3. On the evaluation of the different regulatory models

In this topic, the respondents presented their opinion about the different moderation regimes, as asked in questions 8 to 12 and 14.

4.3.1. On self-regulation by platforms in content moderation

With regard to self-regulation exercised by the platforms, the content moderation rules created by them were considered. Thus, when asked about the evaluation of such guidelines in view of the adequacy to Human Rights, the interviewees were divided into the following opinions: they considered the application **predominantly positive (7x)**, **positive and necessary (1x)**, **predominantly necessary (7x), negative (2x)** and one interviewee classified them as having both positive and negative elements. Among these, 4 interviewees pointed out difficulties in the application of the Human Rights parameters, specifically, and 5 interviewees pointed out general difficulties.

Thus, in view of the arguments presented in favor of the positive performance of the rules applied by the platforms themselves, these were: guarantee of **freedom and autonomy to the platforms(4x),** possibility of adaptation to eventual **new demands(2x),** maintenance of security to the online environment and guarantee of commitment by the platforms. In this sense, Igor Peixoto, a member of the business sector, defends the existence of such rules as follows:

> I think it makes sense for platforms to create their own rules, to start with, because they have **different natures, they have different proposals,** the big ones are global in a certain way, but

> if we think about platforms in general they may have different **geographic environments,** they also have different **enforcement systems,** so they have common practices that **combine the use of artificial intelligence with the performance of content moderators,** but these systems, they end up being different in detail, because they are different companies. (emphasis added)

In addition, Guilherme Araújo, a member of the private sector, points out that it is essential that platforms have their own rules for the maintenance of security in virtual environments, stressing that they should always be updated. Moreover, for the interviewee, self-regulation is a way to adapt to new demands. As an example, he mentions that a few years ago no platform was concerned about developing policies regarding misinformation, which could even be seen as "wanting to arbitrate what is true or not". However, nowadays, facing recent issues related to public health and electoral contexts, platforms have incorporated this concern, developing specific guidelines for the subject.

In a similar vein, Sofia Pires, a third sector member, comments on the quick responses that platforms usually offer in events of an urgent or emergency nature, citing as an example the coronavirus pandemic and the U.S. Capitol Invasion.

As for the general difficulties pointed out by respondents regarding self-regulation of platforms, the following stand out: **opacity (4x), selectivity in the application (3x),** difficulties in the **implementation and monitoring of rules by companies (3x)**, regulatory adequacy or **"legality control" (3x),** and adequacy to the public interest.

Thus, with regard to the demand for more transparency, Samuel Cardoso, a third sector member, argues that the creation of rules by the platforms themselves is a necessity; however, these need to be more transparent and better disclosed. Therefore, he complements by saying that "in our Brazilian view of consumer law, they should be **provided through adequate information, evident and intelligible to the general population, which is not the case** (emphasis added)". In agreement, Ana Carvalho, from the third sector, also makes reference to a legal mechanism to exemplify how self-regulation could be more transparent:

> but I think they should be **more explicit and more transparent**, in the sense that I was talking about, it should be like a Penal Code, let's say, like... **you did this... your sanction is going to be this. If you repeat this conduct, your sanction will be like this.**... You know? Because I think that is what is missing in the content moderation activities that we see today, it is... this is not clear. (emphasis added)

Regarding the issue of selectivity in the application of rules by platforms, Sofia Pires, a third sector member, argues that there would be two types of application: "(...) a type of application of rules for **democracies or spaces that matter,** and a type of application of rules for **democracies and spaces that matter less** (emphasis added)".

In this sense, he comments that the revelations of activist Frances Haugen, a former Facebook employee, in the year 2021, were important to illustrate how platforms "pay attention at different levels to different places in the world." Thus, this would impact how content moderation happens in Global South *versus* Global North spaces, depending on the level of priority given by each platform.

Maria Teixeira, a technical-scientific community member, presents a different position from Sofia when she classifies self-regulation by platforms as negative. However, the two interviewees agree with regard to the selectivity in the application of the rules and consequent absence of defined parameters. In this sense, she says:

> (...) And I think they **are not effective**. For this creation of moderation rules. That's it, **I don't believe in this self-regulation of platforms because to me everything looks like double standards.** It is not that they are going to treat one case of racism the same as another, they are going to **particularize it**. If a political agent expresses racism on social networks, will he be punished, will he be moderated, will he be banned immediately for that content? I have seen several times that no (emphasis added).

Complementing the interviewee's speech, Emanuella Nogueira, a third sector member, defends the application of such rules based on international human rights standards, which would be a possible way to create greater control over the poor establishment of regulation mechanisms observed by the platforms.

As for the difficulties pointed out by the interviewees regarding the identification of Human Rights parameters, those who manifested themselves in this sense defend that, due to cultural and socio-political issues, defining what would actually be a rule that acts in consonance with such principles is a challenge. Given this, the platforms need to pay attention to this diversity, considering their sometimes global action. In this sense, Vicente Moura, a government sector member , comments:

> So, if you ask specifically about **effectiveness**, I think that, as much as there is a literature, an international consensus about some human rights, there is a **need for particularization of each cultural, national, eventually even regional context, regarding these forms of content moderation, around specific**

> **fundamental rights.** And the platforms have capillarity, interest has profited a lot with these regionalized, territorialized users, but when it comes time to particularize their business model to the needs of these people, they refuse, they say it's too costly or laborious, so I think it's time that this excuse is no longer accepted (emphasis added).

In complement, Igor Peixoto, a member of the private sector, presents a position in agreement, when he says that, when designing their own content moderation rules, the platforms think about the "**needs they have to make the space minimally healthy,** that thing of moderation as a product, but this compatibility with international human rights law, it has **never been a priority of the platforms** (emphasis added)."

4.3.2. On state regulation in content moderation

State regulation is understood here as the definition and enforcement of rules by state authorities. Thus, when asked about their positions regarding rules issued by state authorities for the adequacy of content moderation to human rights, most respondents were predominantly in **favor of this type of regulation(11x)**, part were **predominantly against(4x)**, two respondents did not position themselves and one presented arguments both for and against.

Thus, the most recurrent point used to defend the positive action of state regulation was the **establishment of minimum parameters(5x)** for the definition of content moderation rules, thus tracing "minimum and maximums" that do not currently exist in the Brazilian scenario. Together, the following were also mentioned: the establishment of parameters in accordance with Human Rights, the adequacy/ pressure to the activity of the platforms, and the importance of establishing specific guidelines. Moreover, all interviewees who positioned themselves in favor of state regulation did so defending its positive character based on necessity. In this sense, Samuel Cardoso, a third sector member, comments

> Yes, I think that legislation, it is necessary to serve as a parameter for these various internal rules of the platforms, to establish the minimum **so that precisely this defense of rights is not established as a competitive advantage or as an additional resource for those who pay for the service or for those who hire a service specifically,** right? So, of course we have to, I think the legislation has to establish these **minimum levels from which the companies can promote their services**, right? (...) (emphasis added).

In addition, Ana Carvalho, a third sector member, considers it necessary to have a state regulation that "says the minimum, that says exactly that, that the devices, the companies' rules have to be... clear, they have to be explicit, they have to be transparent, they have to be written down saying what is and what is not". This is in agreement with Maria Teixeira, a technical-scientific community member, who states that such norms are fundamental and should be "written, it has to be on paper, it has to be approved by the National Congress. The Judiciary being able to judge things according to this legislation."

Luiz Porto, in turn, comments on the need for state action through regulation when he says:

> What is missing is this, in the platform policies, an **harmonization by force of law,** since the platforms by themselves, despite the important principles, the important Santa Clara principles in this field, from the perspective of self-regulation, from the industry perspective, **are not enough to create within the regulatory environment, and then yes, private, corporate, of transnational private regulation, guidelines and uniform rules regarding content moderation** (emphasis added).

It can thus be observed that the manifestations in view of the establishment of minimum parameters aim precisely at the fields in which there is a "regulatory vacuum". In this sense, the interviewees that defend this point of view see in the state agencies the competence and legitimacy necessary for the definition of these standards that currently do not exist in the content moderation scenario. However, the fact is that such legitimacy was not unanimous among the interviewees, since those who positioned themselves against state regulation presented a greater number of criticisms than those who positioned themselves in favor presented as positive points.

In this context, the arguments pointed out in favor of the negative character of the rules created by state actors were: lack of technical knowledge on the part of State agents, lack of dialogue with other sectors in the establishment of rules, non-application of international parameters (but only internal), application of international parameters focused on removal as the only solution, possibilities of censorship, possibility of abuses in the application of rules, possibility of affecting competition, flexibility and autonomy of companies, and the possibility of imposing legal mechanisms of an imposing character.

Seeing this, Bárbara Silveira, a member of the business sector, defends her position:

> We support a forward-looking approach to regulation, considering the long-term impact on the broader digital ecosystem that protects the Open Internet and universal access. We therefore **believe that any attempt at legislation that only strengthens the dominant position of larger companies will irreparably damage the Open Internet, innovation, and consumer choice.** In short, we consider the **ability for robust competition and the assurance of fair playing field between actors to be** essential (emphasis added).

In this sense, Guilherme Araújo, a member of the private sector, is also wary of such state regulation, saying that "care must be taken" so that the norms in question do not result in a form of state censorship or state pressure for private censorship.

With regard to the challenges of regulation by state actors, a point of view was expressed by three interviewees in a similar manner. It is about the difficulty of regulating the area of technology, given the "speed of technological innovations" and the "risk of rapid obsolescence" to legislation, when too detailed, according to Antonio Cavalcanti, member of the government sector. In agreement, Vicente Moura, also a government sector member, says:

> There is a great difficulty in regulating technology, so, the legal work around this activity, **it must be flexible enough to last over time,** that is, to think of a certain **vitality of this statute**, of this law or of these regulations over time, **to adapt to new technologies or eventually be changed to the extent that these technologies also present new challenges and this is very difficult.** The legislative process is difficult, it takes time; the regulatory process, to be a good regulatory process, involves **impact reports, it involves input from society, public consultation, all these are time consuming processes,** so I think it is an insurmountable challenge that the forms of state regulation will always be behind the challenges posed by these technologies (emphasis added).

In view of the interviewee's comment, it is possible to reflect on how the responses and resources available from the Law are often slow, not very effective and even inefficient. In contrast to the speed of the flow of information on the web and the number of conflicting issues that may come up, the possibility of regulating needs to go through a series of steps, involving games of interest and various challenges. Given this, in the Internet Governance scenario, the regulation of platforms by state bodies is a topic that encounters distinct opinions.

In this scenario, Lucas Oliveira, from the third sector, comments that a state regulation, when of an imposing character, can affect the ecosystem of platforms in a harmful way regarding the rules on content moderation, since platforms should have their freedom and autonomy to determine how content will be moderated. On this topic, he states:

> (...) it [the internet] is plural and it is good that it is so, so each platform has the ability to attract certain content and disengage from other content to create a certain community focused on a certain subject and theme.

In addition, the interviewee criticized state regulation through the creation of mechanisms with exhaustive lists of what could or could not be moderated, citing Provisional Measure 1.068 of 2021[36] as close to this logic. In this case, the referred rule, among other predictions, brought a list of contents that would authorize their exclusion, suspension or blocking by the platform.[37] Nevertheless, Lucas defends the application of state rules that implement greater transparency to content moderation practices, protection of public interest in the digital sphere and the requirement of respect for due process.

4.3.3. On recommendations without legal effect (soft Law)

This topic addresses the opinions of the interviewees regarding the impact of recommendations without legal effect (the so-called soft Law) on the adequacy of content moderation practices to human rights. Thus, as an illustrative example, the Santa Clara Principles were mentioned in the question, which establish transparency and accountability standards for content moderation practices.

When they positioned themselves, a little more than half of the interviewees classified these mechanisms as **positive and efficient (9x)**, and 3 of these interviewees presented criticisms - which did not put the character of efficiency in question. Then, the classifications were: **positive but inefficient (4x)**, **only inefficient (2x)**, **negative (1x)** and **no position (2x)** - one of the interviewees preferred to omit, since he did not consider himself in the condition to answer the question because he knew little about the subject.

[36] BRAZIL. Provisional Measure No. 1.068, of September 6, 2021. **Amends Law No. 12,965 of April 23, 2014, and Law No. 9,610 of February 19, 1998, to dispose on the use of social networks.** DF: Presidency of the Republic, [2021]. Available at: https://www.in.gov.br/en/web/dou/-/medida-provisoria-n-1.068-de-6-de-setembro-de-2021-343277275. Accessed: 02 Dec. 2022.

[37] A few days after the publication of the norm, the STF's minister-rapporteur Rosa Weber, granted requests for injunctive relief in 7 direct actions of unconstitutionality against MP 1.068/21, to suspend its effects. See more in: CARNEIRO, Luiz Orlando. Rosa Weber's arguments to suspend the MP that hindered content removal. **JOTA, September 15, 2021.** Available at: https://www.jota.info/stf/do-supremo/os-argumentos-de-rosa-weber-para-suspender-a-mp-que-dificultava-remocao-de-conteudo-15092021. Accessed: 04 Jan. 2022.

Thus, the main arguments presented in favor of the positive and/or efficient character of such mechanisms were: help in the construction of **parameters for the creation of content moderation rules (5x)**, collaboration for the **creation of regulatory guidelines (3x),** putting **pressure upon platforms (4x)**, creation of debate spaces between various sectors, and construction of relevant information sources.

Thus, regarding parameter creation, the respondents used different expressions with similar meanings, such as: creation of frames, indication of paths and help in guiding the work. In this sense, Guilherme Araújo, a member of the business sector, states that "(...) all good construction starts from the establishment of principles that sometimes come from academia, sometimes from civil society, sometimes from companies.... but in a **range of shared principles**" (emphasis added). Such opinion finds agreement in the speech of Antonio Cavalcanti, a government sector member, when he affirmed "instruments of this kind help mold minds and hearts and **prepare the political ground for legally grounded and protective measures of human rights further ahead**" (emphasis added). For Igor Peixoto, a member of the private sector, these recommendations work as "beacons of expectations for sectors impacted by the platforms," thus allowing the private sector a better understanding by other agents.

Therefore, even though soft law mechanisms do not have normative force, they are seen by some interviewees as important instruments in the regulatory frameworks creation, as they become inspirations in the legislative processes. Then, by determining basic principles and ideal scenarios, they ground the intended normative base in accordance with human rights and ethics. In this sense, Luiz Porto, a technical-scientific community member, considers them to be "more open to transformations," also stating that "it is easier to update and modernize principles of recommendations and guidelines today that capture much more the transformations essence". In this way, the interviewee considers soft law mechanisms as positive and efficient, seeing them as motivating judicial and business decisions. As an illustration, he comments on the impact that the guidelines and recommendations of the Organization for Economic Cooperation and Development (OECD) have achieved in the Brazilian scenario.

On the pressure legal effectless recommendations put on companies, Theo Freitas, a third sector member, comments:

> Companies are conforming in this sense, let's think the Board is an example. Does this mean that companies will implement exactly what these regulations want? No. Let's take, for example, Facebook's latest transparency report. It's... comments they made about errors, the statements about the change in removal numbers by category, and so on. **It creates these efforts in that**

> **sense. It's not something that is provided by regulation, and it's also not something that platforms are just doing because they think it's cool, they're doing it because of pressure. And the pressure comes from these instruments**, so... I think that, to some extent, it does have an effect. (emphasis added)

In agreement, Antonio Cavalcanti points out that, although such instruments do not have coercive power, they play a relevant role by causing direct and indirect effects on the platforms' self-regulation mechanisms, since they "help create an environment that induces and pressures for more humane practices in the use of the most varied technologies."

As for the arguments presented in justification of the negative and/or inefficient character of the instruments under discussion, all of them, directly or indirectly, were related to the limited character that the instruments have in face of the commercial interests of the platforms, considering the current business model. Thus, Vicente Moura, a government sector member, comments in this sense

> I think they are positive, but I'm kind of cynical, so, I think that the **economic** interest **supervening any kind of agreement, declaration of intent or even practices, sector codes of conduct that these companies will develop**, they will not do something that goes against, for example, their economic interests. And content moderation is problematic precisely because **one of the main economic interests of these companies is to gain user engagement and engagement usually occurs through controversial content, fake news, unmoderated content,** so I don't see how **non-binding principles or norms focused on a company's ethics** would be able to propose measures effectively able to regulate content moderation, **so I think they are not enough** (emphasis added).

In a similar context, Lucas Oliveira, a third sector member, considers the recommendations without legal effect positive, but "with moderate effect, not to say null in the moderation of content on digital platforms". What leads to this, in his opinion, is the lack of interest shown by the private sector in actually considering such instruments in their actions. As an example, he cites the lack of reference to the great charters by the Oversight Board in its decisions.

Given the limitations placed by the interviewees who consider the inefficiency of soft law mechanisms, a question was raised about the need for legal force so that such recommendations are in fact followed. This position is illustrated by Fernanda Rezende, member of the third sector, in her comment:

> (...) if the company is watching the dissemination of unequivocally illegal content on its network, of practices that put - abusive and discriminatory use of personal data - and the company does nothing, we need to have accountability mechanisms. Because without this they will not change their conduct. **We need a whole system that integrates even these soft laws and the rules of the companies and the actions of the state. Built based on the multisectoral mechanism, as is proposed in PL 2.630** (emphasis added).

As for the interviewee who considers regulations without legal effect as negative, his main argument was that such instruments would be more directed to promoting a business model than to promoting human rights and fundamental guarantees. In this sense, Samuel Cardoso, a third sector member, states:

> For the promotion of human rights in terms of efficiency, I think they fall short, you know, they could be an instrument for the promotion of human rights, but they end up falling short of what they could and maybe I could evaluate that in general **they are also more focused on promoting a business model than exactly promoting human rights and fundamental guarantees.** So, we have several examples of situations in which, to the extent that the violation of human rights or the doubt in relation to the violation of human rights promotes attention and this attention moves the economy of attention, the platforms end up being more permissive for this type of violation than exactly coercive or limiting this type of violation (emphasis added).

In this way, we have in view that the platforms' business model, fed by the search for engagement and profit, is configured as a determining factor at the moment of (non) application of the principles determined by such instruments, in the interviewee's perspective.

4.3.4. On state rules with procedural guidelines

Regarding the creation of state norms with specific guidelines for content moderation, **there was a broad positive manifestation (14x),** highlighting the need for **minimum bases for an ideal procedure (5x).** In this sense, it was defended the creation of a **due process for moderation (3x)** and the establishment of minimum procedural guarantees for users, such as, for example: **what are the hypotheses and what will be done in moderation activities, reasons and rules used for a decision (3x), possibility of appeal (5x),** deadlines for review of certain content, and rules for notification of users who had content moderated. The mention of these situations demonstrates a concern to make the content moderation procedure more understandable to the people who use the platforms, also thinking of ways to ensure the contestation of decisions.

For Luiz Porto, member of the technical-scientific sector, it is necessary

> make these [moderation] decisions subject to **judicial controls and procedural guarantees for users,** something that is not very clear in the current system of removal decisions by some companies and platforms, since if we look at the community policies, the community rules or the platform policies there is no provision for what procedure the user has to adopt, for example, if he needs to, if he wants to challenge a removal decision.
>
> [...] And there yes, establish, for example, that in making moderation decisions according to the policies of the platforms adopted, the **best guarantees or the guarantee of protection of the rights and individual liberties of Internet users are adopted, as well as the observance of procedural guarantees that are assured to them by force of the Constitution, the Code of Procedure and also the treaties and conventions to which Brazil is a party** [...] (emphasis added).

In addition, other examples of possible guidelines were also listed, such as the determination of: mechanisms to be observed in content removal; creation of specific channels to offer explanations about removals; making content governance policies available in accessible language; requirement for **transparency in moderation rules (3x)**, such as through the **elaboration of transparency reports (2x)** and the type of information they should contain; **disclosure of decision-making parameters and criteria for moderation (2x),** including for research teams from accredited and certified universities and research institutes; incentives for alternative actions to removal, such as flagging or reducing the scope of content and fact-checking with a network of hired experts; information about the use of artificial intelligence; evaluating the alignment of the procedure with international human rights parameters; respect for fundamental rights, such as freedom of expression; and performance evaluation of platforms in the application of their terms of service, "in order to provide flexibility and reduce the incentives for excessively moderate content" along with the incentive to invest in technological solutions.

However, among the favorable positions, some reservations were made about the creation of this type of law. It was pointed out that it is necessary to avoid making the moderation procedure rigid in order to allow its evolution, as well as to preserve a certain degree of autonomy to the platforms to elaborate their terms of use, due to their particularities. Another point was that the creation of a specific guideline on what and how to moderate could create only one way of doing things and that it would change quickly.

On the other hand, the recognition of particularities in the operation of each platform served as an argument for one of the positions contrary to this type of rule. In this case, it was understood that it would be up to the platform to establish its rules, and the State to demand transparency about them. Another unfavorable position recognized the possibility of guidelines to determine cases in which the right of reply or contestation is possible, but argued that it would not be up to the law to give a deadline for content removal.

4.3.5. On state rules providing for the removal of specific content

This debate was closely linked to the considerations made by the interviewees about the viability of the law establishing specific cases of content that should be removed by the platform, in view of what happened with the edition of Provisional Measure 1.068/2021,[38] which brought a list of hypotheses in which removal would be authorized. The question asked for a score, from 0 to 10, about how much the participant considered himself favorable to the creation of this type of rule, **and the average obtained from 14 people who decided to assign a value was 4.8**. This score, close to 5, was well represented in the interviewees' arguments, where the majority presented both favorable and unfavorable considerations.

Among the unfavorable arguments, it was mentioned that establishing a specific list of cases for content removal could generate **incentives for censorship (2x)** and **excessive removal by platforms (3x)**, as well as in the case of applying significant administrative penalties and stipulating short deadlines for removal. On this point, small companies and new services could be more affected because they have fewer resources to litigate or pay eventual fines. Also, the creation of such a list could result in **censorship by the state itself (2x),** as in the case of authoritarian states.

In fact, the distrust in relation to the way in which the government would edit a norm in this sense appeared more than once. Besides the **lack of technical knowledge of legislators (2x)** and a **greater discussion in Brazil about the types of content that should be subject to moderation (2x)**, the debate would lack **more specific definitions around concepts considered "gray areas" (2x)**, as in the case of freedom of speech, and some content, depending on the context in which it was published,

[38] BRAZIL. Provisional Measure No. 1.068, of September 6, 2021. **Amends Law No. 12,965 of April 23, 2014, and Law No. 9,610 of February 19, 1998, to dispose on the use of social networks.** DF: Presidency of the Republic, [2021]. Available at: https://www.in.gov.br/en/web/dou/-/medida-provisoria-n-1.068-de-6-de-setembro-de-2021-343277275. Accessed: 02 Dec. 2022.

could have different meanings, as is the case of nudity content. The lack of certainty of a uniform understanding by the Judiciary, which is possibly responsible for overseeing these rules, about topics such as hate speech, racism, homophobia and others also proved to be a point of difficulty.

In this direction, there was a concern with the creation of categories of content to be removed based on excessively open legal concepts, as in the case of disinformation and hate speech, which could generate legal insecurity and impacts to freedom of expression. Thus, it would be more interesting to focus on the definition of contents that are already considered illegal, **as exceptions provided by law or conducts already considered criminal (9x)**. Some examples cited: child pornography, pedophilia, *revenge porn,* copyright, crimes against life and physical integrity, apology of Nazism, racism, among others. It should be noted, however, that one of the interviewees pointed out that crimes against honor would not be included in the list of contents to be removed, due to the difficulty that would be attributed to the platform in identifying them.

On the other hand, misinformation, as well as again racism, homophobia, transphobia, and violence against women were observed by some of the people who defended the **application of this type of rule for specific or punctual situations (6x)**, such as through judicial determination. However, there were also positions in favor of a state regulation only in general lines, that is, a parameterization on unacceptable points from the social point of view.

The creation of exhaustive lists of content by the State could result in a stifling of the moderation process and impact the plurality of platforms, with the application of a single rule to all of them. Thus, the legislation should not present a closed list of hypotheses. Moreover, by focusing on content, a law like this would also hold **the platform responsible for the actions of third-parties (2x),** threatening, again, rights such as freedom of expression.

For Bárbara Silveira, a member of the business sector, a regulation must

> establish clear standards for the types of content they seek to address, with **substantive definitions and limits consistent with human rights standards**. Where the content in question is legal, but a government believes it is necessary to intervene, the regulatory framework should make a clear distinction between these types of content**. Government requests for the removal of specific pieces of content on the basis of illegality must be based on a legal process and provide transparency about how these powers are used**. It is a fundamental matter of due process of law that a **competent government body,** not a private actor, is

> responsible for determining the illegality of a piece of content. **Companies should be free to let people know that this was the basis for the action to be taken** (emphasis added).

The mention of a government agency was also made by another interviewee, who mentioned the need for greater enforcement of this type of rule by the platforms, something that could be done by an agency similar to the one created for data protection purposes (ANPD), but focused on content moderation cases.

In this vein, another interviewee highlighted the need to avoid delegating state activities to platforms, such as the application and enforcement of the law, since it would constitute a kind of privatization of jurisdiction. It would be possible, in this case, to conduct regulatory impact exercises, in order to "coherently face the forms and procedures of restrictions of these rights, but also, at the same time, demonstrate what are the objectives and the social welfare achieved by the regulatory choice of intervening in the content moderation issue".

In addition to obligations, however, two interviewees advocated the creation of incentives for platforms to conduct content moderation, such as by rewarding those with good work in this area. Yet, another argument highlighted the need to think on a basis around transparency rules and due process before a law that determines the removal of specific content.

Finally, one of the interviewees pointed out that content removal would not be appropriate in the case of private conversations, in which case filing a lawsuit against the offending individual would be the most appropriate way.

4.3.6. On the regime of regulated self-regulation

As already mentioned, the script of questions was prepared based on previous research carried out from a systematic review of foreign and national literature about the different content moderation regimes. In this sense, the question about regulated self-regulation was asked after the identification of the concept in the analyzed sample, so that we sought to better understand if the interviewees already knew the term, what they understood as regulated self-regulation, and how they evaluated this model of regime.

For **two interviewees, regulated self-regulation would actually be a kind of co-regulation**. Lucas Oliveira, a third sector member, stated that "what exists is co-regulation, preserving the sphere of regulation, of self-regulation of digital platforms, and inserting the State in what is its responsibility, that is why co-regulation and not regulated self-regulation".

On the other hand, for Theo Freitas, also a third sector member, regulated self-regulation would differ from co-regulation in that, in the former, the regulator would only elaborate guidelines for action and carry out the evaluation and validation of what will be done by the regulated. Co-regulation, on the other hand, would require the joint elaboration between company and regulator of the "path to be taken" and the details, demanding a more cooperative work.

The difficulty in defining regulated self-regulation as a genus or species was also found in the speech of Caio Fernandes, who stated that it is difficult to define where self-regulation ends and co-regulation begins, since, in the case of regulated self-regulation, "you nod to self-regulation at the same time that you threaten co-regulation".

Despite these impasses, there was a certain consensus that regulated self-regulation would be the possibility of the platforms themselves establishing their own rules, but **based on guidelines provided by the State (10x)**. In this case, the **regulator would not establish the content of this self-regulation (2x)**, but there would be room for monitoring and metrics to evaluate the performance of the platforms.

For Theo Freitas, of the third sector, it would also be up to the regulator to validate the companies' proposed actions, based on the established guidelines, and, for Fernanda Rezende, of the same sector, regulated self-regulation would also allow the existence of a control instance. However, for Stephany das Neves, a member of the technical-scientific sector, this regulation model would be the possibility of enhancing the self-regulation mechanisms based on some legal direction or reinforcement by the State.

With respect to perceptions regarding the effectiveness of this type of regime, some respondents expressed the belief that **regulated self-regulation could be more efficient than other regimes (6x)**, with mention that this option would be better than top-down regulation, as is the case with purely state regulation. However, it was emphasized that its effectiveness would depend on the regulator itself, on good enforcement, and on this regime being both multisectoral and having an integrated international perspective so as not to be captured or circumvented.

Other participants considered regulated self-regulation a **viable path (2x)**, emphasizing again that it is better than a *top-down* regulation, but that it should be guided by a diverse participation, both in its subsequent monitoring and in the formulation of guidelines for the platforms, involving not only the government and business sectors. For other interviewees, this regime would actually be the **best solution available at the moment (3x),** because besides following a trend seen in other countries, of seeking co-regulation, it is a form of regulation that allows state action while ensuring autonomy for the platforms, being public and private.

It was also argued that the model of regulated self-regulation would be a **better alternative than leaving it to the platforms alone to control all their processes (2x)**, including moderation, and other instances to monitor their performance. For Antonio Cavalcanti, a government sector member, this is a regime that makes sense given the current scenario in the technology market:

> (...) the mixture of self-regulation with a self-contained regulation, made by the **State or some regulatory agency**, predominantly principle-based and with a more guiding and directive profile, **allows the maintenance of the necessary space for business experimentation for the development of platforms.**
>
> The characteristics of the technology market and the scenario of intense and fast development of new products and services by Internet platforms **make the adoption of this hybrid solution the ideal scenario**. (emphasis added).

The mention of a control body was also found in the speech of Fernanda Rezende, member of the third sector, who mentioned again the importance of this space being multisectoral, in order to have the participation of the various agents of society in the construction of regulation. Two other arguments highlighted, respectively, that regulated self-regulation would be interesting for promoting a place for dialogue between the object of regulation and the regulator, and that co-regulation mechanisms would be the appropriate path when dealing with situations of risk to fundamental rights.

However, there was criticism of how this regime was translated into PL 2.630/2020, the so-called "Fake News Bill" that, despite its name and after a series of modifications throughout the legislative process, was more aligned to a general regulation of platforms. The proposal presents a model of regulated self-regulation that allows the creation, by application providers, of a "self-regulatory institution focused on transparency and responsibility in the use of the Internet", as shown below:

> **Art. 35 Providers may create a self-regulation institution aimed at transparency and responsibility in the use of the Internet, with the following attributions:**
> I - create and manage a digital platform to receive complaints about contents or accounts and make decisions about measures to be implemented by its members, as well as review decisions about contents and accounts, by means of provocation by those directly affected by the decision;
> II - make decisions, in a timely and effective manner, about the complaints and review of measures covered by this law; and

> III - develop, in conjunction with mobile telephony companies, good practices for the suspension of user accounts whose authenticity is questioned or whose inauthenticity is established; (...)(emphasis added)[39]

In addition, the Bill foresees attributions for the Internet Steering Committee, which would have the role, for example, of "conducting studies, opinions and recommendations on internet freedom, responsibility and transparency"[40] . At this point, the choice to make the CGI **responsible for dialoguing with the private sector a parameter for the platforms' terms of conduct (2x)** was criticized by the interviewees, and it was also pointed out that the proposal would start from the assumption that all content moderation hurts freedom of expression.

It was argued that this strategy would lead to the interpretation that regulated self-regulation necessarily depends on a decision review body, when, in fact, norms, principles and general rules of due process would be enough. Because of this, the PL's provision would be ineffective and fragile, with enforcement problems. At this point, the distrust about how the concept of regulated self-regulation would be applied in Brazil was also mentioned by another interviewee, who showed concern about the distortion of the term.

Doubts were also raised about the effectiveness of the creation of individual policies by each platform, without harmonizing the rules for content moderation:

> Now, I am suspicious of the effectiveness of individual and regulatory policies - sorry, individual codes of self-regulation by the platforms without the potential to harmonize and standardize solutions related to content moderation, because the platforms would have to make an effort that they still do timidly today, which is to be able, in specialized forums, to **sit down with experts and policy makers to find harmonized and standardized solutions on the subject of moderation.** Question: does this happen nowadays? No, if it happens, it is the industry segment itself, but **the industry segment itself would then have to give the opportunity for other multi-sector actors to enter and discuss this model, wouldn't it?**
>
> **Luiz Porto, a member of the technical-scientific sector.**

[39] BRAZIL. Bill 2.630 of 2020. **Establishes the Brazilian Law of Freedom, Responsibility and Transparency on the Internet.** ("*Institui a Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet.*") Brasília-DF: Federal Senate, [2020]. Available at: https://www.camara.leg.br/propostas-legislativas/2256735. Accessed: 19 Dec. 2022.

[40] BRAZIL. Bill 2.630 of 2020. **Establishes the Brazilian Law of Freedom, Responsibility and Transparency on the Internet.** ("*Institui a Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet.*") Brasília-DF: Federal Senate, [2020]. Available at: https://www.camara.leg.br/propostas-legislativas/2256735. Accessed: 19 Dec. 2022.

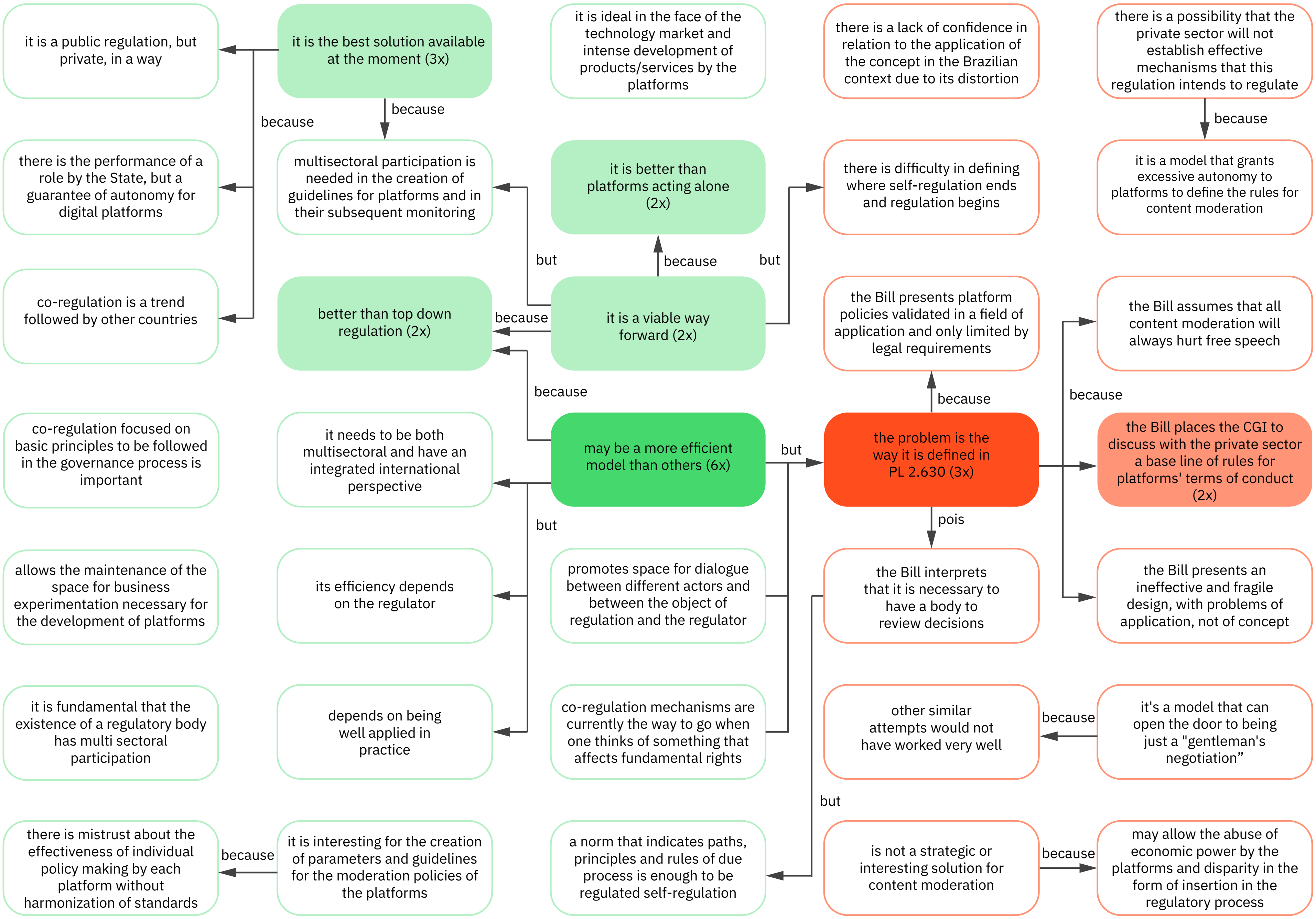
In this sense, the interviewee emphasizes that self-regulation is mistaken in the way it appears in the Bill, since it would appear to represent "a set of platform policies validated within a field or scope of application", that is, as if they could be limited only by strict legal requirements.

In relation to platforms, there was also criticism regarding the power conferred to companies through the adoption of this regime. It was argued that it would not be strategic or interesting for content moderation, given the possibility of "abuse of the economic power of these platforms, the disparity in the way they insert themselves, including in the process of lawmaking today, through lobbying," as well as "the fact that the private sector refuses to establish, in practice, effective mechanisms that this regulation intends to regulate".

In addition, itwas mentioned the fear that regulated self-regulation, through the creation of a regulatory body, might not work properly, as in the case of other agencies, and that its role in the regulation of platforms might occur in a less incisive manner, opening room for a "gentleman's bargain".

Check out the mind map below summarizing the main arguments presented in relation to the regulated self-regulation model:

Figure 3 - Main positive and negative arguments regarding the regulated self-regulation model.



it is a public regulation, but private, in a way

it is the best solution available at the moment (3x)

it is ideal in the face of the technology market and intense development of products/services by the platforms

there is a lack of confidence in relation to the application of the concept in the Brazilian context due to its distortion

there is a possibility that the private sector will not establish effective mechanisms that this regulation intends to regulate

there is the performance of a role by the State, but a guarantee of autonomy for digital platforms

multisectoral participation is needed in the creation of guidelines for platforms and in their subsequent monitoring

it is better than platforms acting alone (2x)

there is difficulty in defining where self-regulation ends and regulation begins

it is a model that grants excessive autonomy to platforms to define the rules for content moderation

co-regulation is a trend followed by other countries

better than top down regulation (2x)

it is a viable way forward (2x)

the Bill presents platform policies validated in a field of application and only limited by legal requirements

the Bill assumes that all content moderation will always hurt free speech

co-regulation focused on basic principles to be followed in the governance process is important

it needs to be both multisectoral and have an integrated international perspective

may be a more efficient model than others (6x)

the problem is the way it is defined in PL 2.630 (3x)

the Bill places the CGI to discuss with the private sector a base line of rules for platforms' terms of conduct (2x)

allows the maintenance of the space for business experimentation necessary for the development of platforms

its efficiency depends on the regulator

promotes space for dialogue between different actors and between the object of regulation and the regulator

the Bill interprets that it is necessary to have a body to review decisions

the Bill presents an ineffective and fragile design, with problems of application, not of concept

it is fundamental that the existence of a regulatory body has multi sectoral participation

depends on being well applied in practice

co-regulation mechanisms are currently the way to go when one thinks of something that affects fundamental rights

other similar attempts would not have worked very well

it's a model that can open the door to being just a "gentleman's negotiation"

there is mistrust about the effectiveness of individual policy making by each platform without harmonization of standards

it is interesting for the creation of parameters and guidelines for the moderation policies of the platforms

a norm that indicates paths, principles and rules of due process is enough to be regulated self-regulation

is not a strategic or interesting solution for content moderation

may allow the abuse of economic power by the platforms and disparity in the form of insertion in the regulatory process

4.4. On the regulation of content moderation in Brazil

This topic addresses the arguments presented by the interviewees about their assessment of the current content moderation regulatory model in Brazil, as well as their perception of the positive and negative points and the impacts of PL 2.630/2020, aimed at regulating platforms in the country.

4.4.1. On the current regulatory model of content moderation in Brazil

In this topic, the interviewees were asked about their perception of the content moderation regulatory model currently in force in Brazil.

For some of them, there wouldn't even be a Brazilian model, that is, **there wouldn't be a specific regulation aimed at content moderation in the country (5x).** The comment by Fernanda Rezende, a third sector member, illustrates this position well:

> Well, I would add that **we don't have a regulatory model for content moderation in Brazil**. Unfortunately, I wish we did, but we don't have one. **We have the self-regulation of the companies, which define their rules**, and we have, incipiently, a draft of this based on the **institutional relationship that we have, which is very early, of the TSE with the platforms,** but we do not have a regulatory model in Brazil. (emphasis added)

At this point, it is noteworthy that most of the interviews were conducted in the period before the 2022 elections, and only one of them was conducted afterwards. Because of this, the Superior Electoral Court's ("*Tribunal Superior Eleitoral*") – TSE more incisive action in moderating online content during the election, especially misinformative, had not yet been observed, but only the attempts at cooperation between the Court and the platforms.

About the inexistence of a Brazilian *sui generis* model, it was also mentioned that Marco Civil da Internet and Lei Geral de Proteção de Dados would be good pieces of legislation, but they would be more focused on internet regulation, and not on content moderation itself. This last theme would be a meta-issue of intermediaries liability, provided by MCI, and its more specific debate **would only have been initiated with PL 2.630/2020 (2x).**

For Augusto da Cruz, a government sector member, the platforms would apply the American model in their actions in Brazil and would follow MCI because it has similar principles to the former, so that the national regulation would be weak and would need to be expanded. On the other hand, some pointed out that the Brazilian model for content moderation would not be non-existent, but rather **insufficient, based on Marco Civil da Internet (9x)**, especially through its article 19, which provides the rule known as notice and take down, that is, which establishes the liability of intermediaries for content published by third-parties on their platforms only when they fail to take the necessary steps after the issuance of a specific court order.

Marco Civil would be a law, therefore, that **responded to the problems of its time (2x),** but its **wording would be outdated (3x)** and short of the complexity of current demands, with **minimal rules on accountability of platforms (2x)** and that would require complementation and improvement. Therefore, in the absence of a more specific law, it was criticized for the **broad power given to companies (2x),** which could abuse it and remove lawful content, without the right to a hearing or the duty to explain to users, generating legal uncertainty. Still, granting more freedom to platforms would not result, as expected, in a more efficient fight against fake news and hate speech. The scrutiny by the Judiciary over their actions, in turn, would also be reduced, since it would be time-consuming and costly, and few people would be able to access a moderation through legal action.

Despite the criticism, it was pointed out that the creation of a more specific content moderation regulatory model could be based on article 19 of MCI, as well as its level of abstraction and more principle-based character. Igor Peixoto, a member of the business sector, highlighted the importance Marco Civil had for internet regulation context in Brazil:

> It [Marco Civil] **defined the issue of responsibility of intermediaries, which was super essential at that moment, that it was not even clear exactly who was responsible for the content posted on social networks** and the regime, then, of article 19 was super important around the court order, which was great, precisely **not to encourage companies on removing content excessively, the side effect of judicial pressure, of State pressure**. And, at the same time, it is the system that left an open **space for platforms to define which are their rules, to remove content based on them**. (emphasis added)

The space for platforms to develop was also listed by Barbara Silveira, from the business sector, who mentioned the possibility given to these actors to take “proactive measures on content that may be legal, but violates your terms of service without fear of litigation.”

Because of its relevance, another argument suggested putting MCI under review, in order to verify how it has been applied and interpreted by courts and enforcement agencies, as well as its effectiveness. For Luiz Porto, from the technical-scientific sector, this proposal is more interesting than creating a piece of legislation that might upset the balance of rights and obligations in Marco Civil.

Still on this topic, the interviewees also commented on how they evaluated the role of the actors involved in the Brazilian regulatory model, as well as if changes would be necessary to reconcile content moderation and human rights, and what would be the mechanisms to ensure that the actors, including intermediaries, would not subvert this model regime.

For Samuel Cardoso, the three State powers are not performing as well as expected, whether due to bad information, concern with secondary agendas, such as party political, lack of thematic specialization, or lack of performance, among others:

> Thinking about the State aspect, I think that, since the **Legislative Branch** is inert – I wouldn't say inert, we have had a lot of debate, but it is badly informed – I think that the Legislative Branch doesn't treat the issue with material seriousness, I think that it has a lot of **political perspective and little technical, technological perspective and little technical-legal perspective, and few effectively social concerns in this matter and many party-political concerns, and many concerns of the demand for power and of meeting limited perspectives**. From the point of view of the **Judiciary**, I think there is a lack of training, updating, and specific instances to be able, in some cases, to deal with these issues in a more specialized way, **but it is the one that has been doing something, although making mistakes in various situations, it is the one who has been, in a certain way, more efficient**. The **Executive Branch** fails in making **public policies**, in **promoting debates,** and the last great measure of interference in terms of moderation was the provisional measure. It was a complete absurdity, a very serious error that fortunately was overthrown by the operation of the checks and balances of the separation of powers in Brazil (...). (emphasis added)

The criticism of the Legislative branch goes along with another perception raised, in the sense that there is fear about the actions of this power, which could result in a bad regulation. According to Vicente Moura, the economic power characteristic of the private sector in this debate would also be able to influence the legislative process, and even the way the Judiciary interprets real cases. In this sense, he highlighted the performance of civil society organizations in the "reaction to something eventually fishy about it that is inserted, either as a legislative proposal or even as novelties in the platform, such as alteration of privacy policies or terms of use of different platforms, as was the case of WhatsApp".

However, one of the perceptions collected was precisely in the sense that the work of the third sector in this scenario has been directed more to a reactive posture, due to the amount of issues to be faced and the lack of time to do so. On the other hand, in relation to the technical-scientific community, it was argued that there is a certain distancing by the technology field, which would not establish a greater dialogue with the other sectors.

Thinking about mechanisms to avoid the subversion of the content moderation regime in effect in Brazil, it would be necessary a greater explanation and exposition of the reasons for the construction of the current level of moderation, which was the result of extensive analysis based on foreign experiences. For Stephany das Neves, from the technical-scientific sector, it is also necessary for the public debate itself to mature around what moderation is and what is expected from a law that will regulate the theme:

> (...) I think we need to have a more mature public debate, we need to be **clear about what are the assumptions of moderation, what are the functions of moderation, if it is a legitimate activity, to what extent, what are its limits.** I would say that we need a serious public debate, supported by facts, so that we can in fact be **clear about what the function of the law is**. This is perhaps the main problem today, we don't have clarity: **what is the purpose of the law? What is the purpose of moderation and, therefore, what is the purpose of the law?**

For the interviewee, only with these answers it would be possible to understand whether the regime is being subverted or not. Another position also defended the need for more debate, but among the players involved in the regulatory scenario, with special attention to the platforms and what they would effectively be able to accomplish in relation to the application of rules and request time. Similarly, fines and sanctions, both administrative and applied by the judiciary, could act as mechanisms to strengthen compliance with legal provisions.

For Lucas Oliveira, from the third sector, however, there would already be a great incentive for intermediaries to respect regulations in this sense, since there would be a predisposition for companies to outsource the problems related to content moderation.

4.4.2. On the positive and negative aspects, as well as the impacts expected from the approval of PL 2.630 ("Fake News Bill")

In this topic, the interviewees were asked about their evaluation of PL 2.630, or "Fake News Bill", which, despite its name, has in its most recent version an attempt to regulate social platforms in Brazil. As already mentioned, this Bill went through intense discussions, mainly aiming to be approved before the 2022 elections, so that it could already be applied during that period.

Along these lines, it should again be noted that most of the interviews were conducted before the election, so this was the context of the interviewees' analysis of the legislative proposal. Thus, they highlighted, in addition to positive and negative points, the possible impacts that could result from the approval of the project.

In relation to the positive aspects of PL 2.630, two topics were among the most cited: **guarantee of greater transparency (11x)**, such as through the obligation to publish transparency reports, and even in relation to advertisements; and the creation of **rules for a due process of content moderation (6x)**, but without stifling the possibility of evolution of this system. Regarding transparency, it was also highlighted that it could help increase the legitimacy of the moderation process and the efficiency in directing civil society pressure on this agenda. However, there would be a certain exaggeration in the amount of items required in the transparency reports, such as specific data about the moderator teams (their configuration, nationality, demographic data), which could lead to risks to privacy.

The availability of **challenge mechanisms for the user (2x)** who does not agree with a certain removal was mentioned as a positive point, as it could also help bring legitimacy and greater transparency to the procedure. Likewise, the **rationalization of eventual fines and sanctions to platforms (2x)**, even if in some cases some excessiveness could be seen; the prohibition of public agents, who use social networks in an institutional way, to block users; the availability of terms of use in accessible language; advances in the encryption agenda in private messaging applications, with restraints to eventual monitoring of users; and providing of a multisectoral space to help define the code of conduct of platforms were other positive arguments raised.

Regarding the latter, however, it was also presented as a negative point, since a regulated self-regulatory entity would be interesting only for some platforms, not for others. In fact, one of the problems would be exactly the search for a global solution that would not face the specificities of each platform.

Even so, it was recognized that the PL 2.630 would be a pragmatic law, which would have concessions for different actors, as was the case of the Lei Geral de Proteção de Dados and Marco Civil da Internet. In this scenario, it would be interesting to **have a minimum regulation for the platforms (3x),** considering that the regulatory void would mean greater vulnerability for all and the proximity to the elections, at least at the time of the interviews.

Moreover, this Bill would already be contributing to the **evolution of the Brazilian public debate on content moderation (2x)** and the importance of reporting platform abuses. In addition, it would seek to improve the quality of content received by users and would be an opportunity to "deepen the rights of Internet users on social networks", based on what is already provided in Marco Civil da Internet.

For Sofia Pires, a third sector member, the differentiation made for large-scale platforms would also be positive by highlighting and protecting the rest of the Internet.

> (...) and this differentiation that the Bill makes between large-scale platforms, with X number of more than 3 or 4 million users, something like 10 million users, I think, and also to the point of being able to **safeguard and continue protecting the rest of the Internet, the rest of the platforms that work with content and do not have a very incisive action or that close them almost immediately**, you know?

On the contrary, Samuel Cardoso pointed out that the metric for differentiating large platforms, i.e. through the number of users, would present major reliability challenges, such as due to fake accounts and bots.

> (...) at the same time, there are some criteria about which I am very, very skeptical, for example, to determine that the platforms have an X amount of users. **We do not have a reliable metric to establish how many Twitter users there are in Brazil, for example** – to give the example of a relevant network and that raises real doubts about its size in the country or in the world as a whole. How many are bots, how many are fake profiles, how many different people use the same amount, the same account or whatever. (emphasis added)

Among the negative points of the PL, two situations emerged as the most mentioned: the **remuneration of news vehicles (12x)** and the provision of **parliamentary immunity in social networks (7x)**.[41] The reasons for the first point would be: a) the agenda complexity, which would demand a more specific debate, and not only in a piece of legislation; b) the fact that it would not be related to content moderation in

[41] "(...) paragraph 8 of article [22], in the current version of the substitute [of PL 2.630/2020], extends the material parliamentary immunity to social networks, preventing deputies and senators from being held liable, civilly and criminally, for opinions and words spoken in the digital environment. (...).) Despite the argument that the text only reinforces what is already in the Federal Constitution, its presence in a specific law represents an extension of parliamentary immunity, allowing the understanding that these actors would be above the content moderation rules of digital platforms" (our translation for the original: "*No entanto, o parágrafo 8º do artigo [22], na atual versão do substitutivo [do PL 2.630/2020], estende a imunidade parlamentar material às redes sociais, impedindo que deputados e senadores sejam responsabilizados, civil e penalmente, por opiniões e palavras proferidas em ambiente digital.* (...) *Apesar do argumento de que o texto apenas reforça o que já está na Constituição Federal, sua presença em lei específica representa um alargamento da imunidade parlamentar, permitindo o entendimento de que esses atores estariam acima das regras de moderação de conteúdo de plataformas digitais*"). RIGHTS IN NETWORK COALITION ("*COALIZÃO DIREITOS NA REDE*"). Alert of civil society organizations on PL 2.630/2020. ("*Alerta de organizações da sociedade civil sobre o Projeto de Lei 2630/2020*"). 6 Apr. 2022. Available at: https://direitosnarede.org.br/2022/04/06/alerta-de-organizacoes-da-sociedade-civil-sobre-o-projeto-de-lei-2630-2020/. Accessed: 04 Jan. 2022.

order to be in the Bill; and c) it would be a sensitive theme, without consolidation elsewhere, and without provisions on how the procedure would actually take place. For Guilherme Araújo, from the private sector, this would not be a topic within PL 2.630:

> Civil society itself has called attention to journalistic remuneration, **we need to debate this from another angle, this is an important debate, I think it's a debate that touches more... one or another platform and not social networks or even messaging apps**, this is more of an issue related to search engines, and this is one of the serious problems of the piece not being a technologically neutral regulation, as is for example the LGPD. (emphasis added)

Among the interviewees, there was also criticism that the legislative proposal would be more focused on this theme, based on the idea that greater investment in journalism would be the solution to eliminate misinformation, and that it would have little result on content moderation. However, a reservation was made in the sense that monetizing journalism might not be all bad, but it would not be enough as a solution to problems involving eventual discrediting of this area, "in the sense of losing readers".

As for parliamentary immunity, it was argued that it would be the right of people to see what public officials share, without limiting access to this content. For Lucas Oliveira, of the third sector, the provision of parliamentary immunity in the networks would not be contemplated in the Federal Constitution and would also go against research that shows the role that political actors have played in the spread of disinformation online:

> So, for example, that attempt to put in the Bill a parliamentary immunity in social networks. Man, for me this is unthinkable, it's foot shooting... it's absurd, right? Because, obviously, parliaments have parliamentary immunity, I can't even discuss this, this is one of the pillars of our democracy, it is there in the Constitution, **but where it is written that this parliamentary immunity is also reflected, is also reflected in the digital environment,** right? And this even **contradicts a series of empirical researches that conclude that, today, the main actors of disinformation on the Internet are those who are running for elections, who hold elective positions,** these people have a very great influence and what they post on the platforms ends up having an impact, so much so that when Twitter removed Donald Trump from the platforms, from Twitter specifically, a research came out afterwards from a study center in the United States

> concluding that disinformation on the elections, on electoral fraud in the United States, fell by 70% just with the removal of Trump's account, you know? So, **focusing on these specific actors is important, even to combat misinformation on digital platforms**. (emphasis added)

Another negative point raised was the **equation of social networks to communication media (2x)**, which would be contrary to what is already provided for in Marco Civil da Internet, which was concerned with differentiating internet application providers that "are not actors dedicated to this activity-end communication". The perception of a confusion brought by PL 2.630 meets the criticism presented about the insertion of issues involving the profiling of user data, when a more specific law, such as LGPD, is already in force, as well as the fact that it is a proposal that tries to address many issues at the same time, being called a Frankenstein law, even more when considering that it started with the goal of fighting disinformation and became a platform governance PL.

Moreover, the lack of a more in-depth public debate about which content should be moderated would also impact the performance of the referred project. For Stephany das Neves, a member of the technical-scientific sector, the very wording of the proposal would start from an assumption that content moderation always violates freedom of expression, being necessary to see that this activity is legitimate, and from then on establish its limits. According to her, there was no creation of incentives for moderation to occur "in an adequate manner, reducing misinformation, expanding the plurality of debates".

In this regard, Bárbara Silveira, from the business sector, pointed out that, after amendments, the Bill has dealt little with disinformation, so it would be more interesting to return to its original intention and combat this issue, especially in the political context, "excluding issues that contribute nothing to tackling the issue such as the aforementioned provision" and bringing "more provisions aimed at combating the financial incentives of agents that create and disseminate disinformation in a coordinated manner". If approved as it is, the interviewee pointed out that

> (...) the text about to be voted at the House of Representatives will **restrict people's access to diverse and plural sources of information; discourage platforms from taking steps to maintain a healthy online environment**; and **negatively impact businesses that seek to connect with their consumers through digital ads and services**. (emphasis added)

Moreover, **creating very strict requirements and obligations on platforms (2x)** could result in social and institutional damage. The incentive that the Bill would bring for certain content removal, for example, could result in collateral censorship. Provisions related to advertising and marketing were also mentioned, and it was suggested that they should be streamlined, as they currently serve commercial interests rather than society.

Finally, negative issues were raised that appeared even in the initial version of PL 2.630 in the Federal Senate, such as: the creation of very open concepts, as was the case of misinformation and inauthentic accounts, for example, which again would generate stimulus for excessive removal; and issues involving the **traceability of instant messages (4x),** with the expansion of user data collection and storage; and data access for investigation purposes.

When questioned about what mechanisms would be interesting to guarantee the effectiveness of a law such as PL 2.630, in the case of its approval, it was mentioned that the approval procedure itself in the legislature would be an instrument. In this sense, the debate about the proposal and the construction of a wording that is attentive to the complexity of the problems that present themselves should be added to a subsequent broad dissemination of the new norm, in order to explain and institutionally capacitate the Public Prosecutor's Office as well as the Judiciary, the Executive, schools, and the academic and scientific community.

Still, the possibility of creating a specific regulatory agency, aimed at monitoring the performance and published transparency reports and requesting information, was pointed out, but considering the need to avoid subversion based on the political interests of the members. The Judiciary itself, through the application of appropriate legal measures in case of legal non-compliance, was also raised as an important mechanism to **enforce the provisions in Bill 2.630 (2x).** The initiative of regulated self-regulation could also help in this scenario.

For Emanuella Nogueira, however, it should be noted that certain obligations provisioned in the legislative proposal would depend not only on the platforms, but also on other actors, such as political agents.

> And also some obligations that are there are not only issues that can be charged from the platforms, but also, for example, from public agents that use the social media platforms, depending on some of the rules there have consequences for other actors. Let's say, that even from the point of view of enforcement by the judiciary alone, I **think they can be very effective depending on how they are used, by political actors or even by civil society**, considering they are there to deal with political or state agents and how they use the public power communication channels. (emphasis added)

Likewise, an instance of regulated self-regulation could make a contribution, but it would be necessary to evaluate how this would actually work.

5. Discussion and final considerations

The objective of this report was to map the beliefs, perceptions and arguments of specialists about the role played by the different sectors that act in the regulation of content moderation, as well as about the possible regime models and the discussion around a Brazilian model construction. The results showed that, despite some consensus, there are still issues that need to be better worked on and further discussed in the public debate, in order to arrive at a regulation that is consistent with the national context.

Initially, regarding the conceptions about the functions and actions constituting content moderation practices, the main idea presented by the interviewees is that moderating the virtual space basically consists of making it "healthy" and, thus, subject to use and interaction by users. To this end, interventional actions would be applied with the main objective of removing harmful, illegal and/or offensive content. After overcoming this basic first step, it is understood that moderation includes actions with objectives beyond simple sanitization, contemplating issues of content curation and behavior modulation. This point, in turn, is related to the business model of the large platforms, since, in order to deepen the users engagement, the contents are moderated to increase network engagement. In this sense, actions such as content ranking, labeling and incentives/disincentives are applied to posts. Consequently, there is a deeper network polarization, which is related to different phenomena, such as bubble filters, surveillance acts, data extraction and algorithmic modulations.

Regarding the use of automated mechanisms for content moderation, the discussions demonstrated recognition of their role importance, even more so considering the large amount of information available on the network every second, as well as their efficiency for specific topics, such as child sexual abuse and exploration material, and copyright. However, the failure in analyzing context or language and the possibility of biases in these systems were some of the important caveats made in this topic.

Indeed, different studies have already shown that automated moderation mechanisms can present errors when analyzing content published by minority groups, for example, drag queens[42] and black people.[43] Thus, it is necessary to pay attention to this type of tool risks of, instead of helping to sanitize the digital platform environment, ultimately silencing people who are historically marginalized in society.

In this sense, interviewees mentioned alternatives such as moderation by AI that is auditable, more transparent, anti-discriminatory and sensitive to context and regional specificities, as well as the possibility of appealing their decisions and the need to guarantee human review. As for the latter, it was pointed out it should be used in a complementary way to automated moderation, as a kind of filter for flagged cases or cases that went through the AI itself.

However, it was possible to verify an important concern with the psychological condition of human moderators, who would be exposed to large amounts of harmful content in a short period of time – a theme that has also been explored in the literature.[44] The delegation of this type of work to developing countries and the limited information about moderation teams were mentioned as aggravating factors in this scenario.

In face of agents' responsibility in terms of content moderation practices, it was possible to observe that the private sector, represented by large platforms and technology conglomerates, is understood as the sector that holds the greatest powers and, therefore, duties in the field. In this scenario, the absence of express regulatory guidelines and specific legal frameworks for content moderation actions generates an ambiguous situation, in which there are no determined obligations, but neither are there express limits. Therefore, due to the privileged situation platforms have in terms of resources availability and the nature of their activities, they are attributed the attribution to moderate.

Although the private sector was chosen as the main responsible for moderating activities, the interviewees were mostly skeptical in relation to the resources

[42] OLIVA, Thiago Dias; ANTONIALLI, Dennys Marcelo; GOMES, Alessandra. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. **Sexuality & Culture.** v. 25. n. 2, p. 700–732 abr. 2021. Available at: https://link.springer.com/article/10.1007/s12119-020-09790-w Accessed: 20 dez.. 2022.

[43] HAIMSON, Oliver L.; DELMONACO, Daniel; NIE, Peipei; WEGNER, Andrea. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. In: ACM CONFERENCE ON HUMAN-COMPUTER INTERACTION, 2021. **Annals [...]** Association for Computing Machinery, 2021. Available at: https://dl.acm.org/doi/10.1145/3479610. Accessed: 19 dez. 2022.

[44] STEIGER, Miriah et al. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In: CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 2021, Yokohama. **Annals [...]** Yokohama: Association for Computing Machinery, 2021. Available at: https://dl.acm.org/doi/pdf/10.1145/3411764.3445092. Accessed: 20 dez. 2022.

available for society to monitor such practices. In this scenario, the question of the absence of express guidelines to conduct content moderation emerges again. While there are no specific rules for publishing transparency reports, for example, there would be no reason for large companies to actually commit to the transparency and accountability principles. This issue, in turn, is also related to the platforms' business model prioritization, since absolute transparency practice could lead to commercial losses or to fulfilling conditions not so interesting to the private sector with regard to the economic interests at stake.

In this context, the discussion on private instances of content moderation decision review is inserted, with the Oversight Board being the main example considered. The unprecedented and still new character of this initiative was a point taken into consideration, leading to reservations on giving an opinion on and/or positions in appreciation of the initiative, and expectations regarding its maturation. Thus, it was observed that, despite the majority of respondents defending the body's positive performance, several criticisms were presented. The main focus of such considerations was on the performance and limited autonomy of private instances, since they start from an initiative of platforms themselves and, thus, their impartiality would be compromised. In relation to the Board, specifically, it was observed that the functioning of the body raises doubts and criticism. Since it intends to analyze a few cases in order to publish recommendations of a systemic nature based on them, the main criticisms were addressed precisely to the fact that more cases – beyond the paradigmatic ones – are not evaluated, in order to generate a "greater impact" performance. Furthermore, criticisms of the impact were also linked to the opacity of the body, given the absence of parameters pointed out by the interviewees for conducting decisions.

As for the possibility of the Judiciary acting as a possible instance of judicial review of moderation decisions, some interviewees pointed out that Marco Civil da Internet had qualified it as an essential part of this activity. This is because, by providing, in its article 19, that intermediaries could only be held responsible for content published by third-parties on their platforms in the event of non-compliance with a specific court order, the courts were empowered as an important figure in this ecosystem. It should be noted, however, that this argument conflicts with another one presented, in the sense that it would be beyond the competence of the Judiciary to act as a reviewing instance of this activity.

On the other hand, besides the fact that there is no way to escape the possibility of seeking the Judiciary due to the guarantee of access to justice for every citizen, it would also be fundamental to act in conflicts solution or rights restriction, in the same way as in illegalities correction and in any damage – caused by the moderation activity – repairing. However, taking a very large number of such claims to court could mean contributing to the overload already observed in courts. Still, the mismatch between the evolution speed of the internet and the Judiciary, as well as the lack of specialized courts on the theme, would hinder a more assertive performance of this branch of power, and there would also be a risk of granting inappropriate decisions.

Regarding the third sector and academia performance in the regulatory processes of online content, the interviewees' positions were mostly in favor of its positive nature; however, criticisms were present in almost all of the speeches. Thus, it is concluded that the most recurrent point was that these sectors make significant contributions to qualifying the public debate, emphasizing participation in legislative processes and pressuring on private and government sectors practices. However, they encounter several limitations in their activities, mainly in view of resource scarcity and, therefore, greater difficulties in obtaining significant influence in the field.

With regard to the creation of rules by platforms themselves, there is a majority understanding in agreement with the need for them to have this autonomy – since internal rules existence is necessary for determining accepted behaviors and, therefore, companies liability. However, there were criticisms to how these rules are developed, focusing on the opaque feature (or even non-existence) of their parameters and guidelines, and on the difficulties of implementation and enforcement, which, by not presenting well-established parameters, may occur selectively and impartially.

In this scenario, another question was raised regarding the difficulty in developing rules, from self-regulation, that are actually conducted in compliance with Human Rights principles. It was pointed out that drawing these parameters in a universal approach is a delicate activity, involving the consideration of diversity issues – even more so when the platforms have a global operation. However, it is up to the platforms to pay attention to these issues when building and enforcing their rules, which most of the interviewees indicated as something that does not happen.

As for state regulation, it was observed that most respondents understand the enforcement of rules by state agents as necessary in the content moderation field. This position, in turn, dialogues with issues already raised, such as frustration with the absence of specific regulatory guidelines regarding content moderation, the absence or inefficiency of monitoring mechanisms regarding the performance of the private sector, and a privileged position of the platforms in managing content moderation. Thus, it was highlighted that state regulation would adopt the main role of establishing minimum standards in the field, absent today – therefore creating transparency obligations, for example. However, criticisms were also presented in view of the possibilities of abuse, censorship and the difficulties of regulating the technology area, given the strong possibility of legislation obsolescence before the speed of processes and transformations.

With regard to the so-called regulated self-regulation regime, it was possible to verify a certain consensus that its definition would be the possibility of the platforms themselves establishing their rules, but based on guidelines provided by the State and without the regulator establishing the content of this self-regulation. However, there were discussions in the sense that this regime would actually be a kind of co-regulation and therefore, the possibility of demanding or not the existence of a control body.

Even so, it was pointed out as a model that could be more efficient than the others, besides being the best option at the moment, following a trend seen in other national initiatives. This position can also be verified in the specialized literature, such as Hartmann and Sarlet, who argue regulated self-regulation as the best current alternative, being "acceptable that platforms have some autonomy in managing the speech", but without ruling out an external control delimited in its characteristics and conditions.[45]

The way this model was introduced in the text of Bill 2.630/2020, also known as "Fake News Bill", was heavily criticized, as for allowing the interpretation that the creation of a moderation decision review body was a condition for this regime's existence. In addition, the stipulation of the Brazilian Internet Steering Committee as the body responsible for dialog with platforms on the elaboration of codes of conduct was criticized. More generally, negative aspects about regulated self-regulation were also related to fearing economic abuse by platforms and that the performance of a regulatory body might not work properly in the Brazilian scenario.

With regard to soft law mechanisms in the face of content moderation practices adequacy to human rights, the absence of legal force and the freedom of platforms to define how their application will actually be were issues criticized by respondents. Thus, the theme finds corresponding elements in previous topics, such as the responsibility for managing content moderation and the skepticism of respondents towards monitoring mechanism implementation. It is thus observed that, once again, the discretionary feature of platforms when it comes to defining content moderation rules and the absence of express guidelines arising from legal mechanisms means that only elements in accordance with the business models are considered. However, it is worth mentioning that many interviewees cited positive points in the legal effectless recommendations, such as the conduction of parameters for creating platforms rules and regulatory frameworks, the execution of pressure towards the platforms, and the creation of spaces for debate.

[45] HARTMANN, Ivar Alberto Martins; SARLET, Ingo Wolfgang. Fundamental rights and private law: the protection of freedom of expression in social media.**Public Law Magazine,** v. 16, n. 90, p. 85-108, dez. 2019. Available at: https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/3755. Accessed: 19 dez. 2022.

In the case of creating state norms establishing procedural guidelines for content moderation, there was a broad positive response, with emphasis on the creation of minimum bases for an ideal procedure and a due process for this activity, establishing minimum procedural guarantees for users. Among them, it was mentioned: what are the hypotheses and what will be done in moderation activities, reasons and rules used for a decision, possibility of appeal, deadlines for reviewing certain contents, and rules for notifying users who have content moderated.

Other examples of guidelines were also presented and are worth highlighting: mechanisms to be observed when removing content; specific channels for offering explanations on removals; content governance policies in accessible language; preparation of transparency reports and the type of information they should contain; decision-making parameters and criteria for moderation; incentives for alternative actions to removal, such as signaling or reducing the range of content and fact-checking with a network of hired experts; information about the use of artificial intelligence; evaluation of the procedure's alignment with international human rights parameters and of the platforms performance in enforcing their terms of service; and respect for fundamental rights, such as freedom of expression.

These conclusions were fundamental, as they demonstrate the due process provisions set forth in PL 2.630/2020 are in line with the position of field experts. However, it is worth mentioning that caveats such as the need to avoid a rigidity in the moderation procedure and respect for the particularities of different platforms are on the radar to delimit the scope of this type of norm.

Still regarding state legislation, the arguments of the interviewees were quite divided when it came to rules aimed at determining the removal of specific content. With the edition of Medida Provisória nº 1.068/2021, by the Bolsonaro government, a warning sign was lit for the possibility that norms in this sense would be misused. Among those who positioned themselves in favor of this law type, there was mention of specific contents that could be subject to it, such as racism, homophobia and transphobia. However, there was concern about the absence of more specific definitions around concepts considered "gray areas", as would be the case of free expression, and a uniform judicial understanding on topics such as hate speech.

Focusing on content that is already considered illegal in the current legal system, i.e., exceptions provided by law or behavior already criminalized, would be one of the main alternatives to avoid possible negative consequences. In this context, it is worth mentioning the fear that laws requiring specific content removal could create more incentives for censorship (both by State and platforms themselves) and for excessive removal by intermediaries, in an attempt to avoid liability.

Regarding content moderation regulation in Brazil, it was possible to verify a general need to advance specific provisions for this area. The opinions of the interviewees were divided, mainly, in the sense that either the current Brazilian model would be non-existent or insufficient, in which case only the aforementioned article 19 of Marco Civil da Internet would be the most tangible norm in this regard. However, the wording of the MCI would be outdated and below the complexity of issues currently in evidence.

Even so, drafting a regulation with ballast in Marco Civil, observing its level of abstraction and principle-based nature, or putting it under review, given its years of validity, were positions that highlighted the importance of this law for the context of Internet regulation in Brazil. These issues are fundamentally important in the face of the debate promoted by Bill 2.630/2020, which faced specific stipulations for content moderation.

In this scenario, the question regarding the lack of regulation aimed at the field was mentioned again, arguing that this absence would open up room for companies to have broad power over moderation, which would facilitate abuses, such as the removal of lawful content without right to contradictory or further explanations to users. Due to the costs and length of time, relying on the Judiciary Power's assessment of their performance would be a reduced possibility, as few people would be able to access the judicial process.

In a more targeted way, regarding the evaluation of positive and negative aspects of Bill 2.630, some issues gained greater relevance in the discussion. Regarding favorable points, the most mentioned topics were the guarantee of greater transparency and the creation of rules for a proper moderation process through the proposal. Next, other positive points raised were the availability of contestation mechanisms for the user; rationalization of possible fines and sanctions for platforms; the blocking users prohibition for public agents who use social networks institutionally; the provision of accessible language terms of use; advances in the agenda of encryption in private messaging applications, with restraints on eventual monitoring of users; and the provision of a multisectoral space to help define the platforms' code of conduct.

It is interesting to note that some of these points had already been assessed as positive in previous questions not addressed to Bill 2.630, such as the creation of a due process and the establishment of mechanisms such as transparency reports. In this direction, the project is in line with those who had already spoken in favor of state regulations that provide minimum procedural guidelines for content moderation.

On the other hand, regarding the negative aspects, the emphasis was on providing remuneration for news vehicles, which should be the subject of a specific regulation for this purpose, and parliamentary immunity in social networks, which would go against the researchers pointing out the role of political actors in the propagation of online disinformation. In addition, the following were also mentioned: equating social networks to mass media; lack of an in-depth public debate on what content should be moderated; reducing predictions around misinformation; creation of very severe requirements and obligations for platforms, which could result in increased censorship, among others.

These examples demonstrate that, despite the progress made in discussions around PL 2.630, many negative points could still be improved for a final version. After the rush to seek its approval before the 2022 election, the next few years will allow the debate on platform regulation, as well as content moderation, to be more in-depth and studied among different stakeholders. Thus, bearing in mind these and other points for improvement presented in this work can be a good starting point in order not to repeat any mistakes in an eventual new legislation attempt.

Anyhow, what was demonstrated by the study carried out is that marked disagreements persist regarding various aspects of online content governance. Several interests are at stake, economic as well as social and political, in addition to the dispute to protect basic fundamental rights, as is the case of freedom of expression and demonstration. The ever-increasing place that large social platforms have been gaining in people's lives, however, leaves no doubt that this debate is necessary, so that abuses are not committed and so that users can feel safe in virtual spaces.

In this sense, many criticisms were pointed out, demonstrating dissatisfaction with the asymmetry of powers that make up the scenario of online content governance today, especially in Brazil, where there is still no specific regulation to delimit the broad field of action of platforms. However, solutions that intend to be quick and instantaneous were not proposed, also because there is an understanding among the actors that, in order to solve the main issues, it is necessary to balance interests, in respect of the multistakeholderism principle.

Between the government sector, the civil society, the technical-scientific community and the digital companies, the regulation of content moderation requires an open, transparent and realistic dialogue, which can guarantee the creation of a normative framework that makes sense in the context and towards the specific demands of the Brazilian territory.

6. Bibliographic references

ACCESS NOW. **26 recommendations of content governance**: a guide for lawmakers, regulators, and company policy makers. Available at: https://www.accessnow.org/cms/assets/uploads/2020/03/Recommendations-On-Content-Governance-digital.pdf. Accessed: 19 dez. 2022.

ARCHEGAS, João Victor; GODOY, Miguel Gualano de. Os limites da jurisdição do Facebook Oversight Board: de Marbury v Madison para Facebook v Trump. **Jota Info,** 02 de fevereiro de 2021. Available at: https://www.jota.info/opiniao-e-analise/artigos/os-limites-da-jurisdicao-do-facebook-oversight-board-02022021. Accessed: 20 dez. 2022.

COALIZÃO DIREITOS NA REDE. **Nota técnica sobre o relatório de 26 de junho de 2020 ao Projeto Lei nº 2.630/2020.** 28 de junho de 2020. Available at: https://direitosnarede.org.br/2020/06/28/nota-tecnica-sobre-o-relatorio-de-26-de-junho-de-2020-ao-projeto-lei-no-2-630-2020/. Accessed: 19 dez. 2022.

COMITÊ DE SUPERVISÃO. **Comitê de Supervisão.** 2022. Available at: https://oversightboard.com/. Accessed: 30 ago. 2022.

DENARDIS, Laura. **The Global War for Internet Governance.** Yale University Press, 2014.

DOURADO, Tatiana Maria Silva Galvão. **Fake news na eleição presidencial de 2018 no Brasil.** 308 f. Tese (Doutorado) – Programa de Pós-Graduação em Comunicação e Culturas Contemporâneas, Universidade Federal da Bahia, Salvador, 2020.

EFF – Electronic Frontier Foundation et al. **Princípios de Manila sobre responsabilidade civil de intermediários**. Available at: https://manilaprinciples.org/pt-br/principles.html. Acesso 20 dez. 2022.

EFF – Electronic Frontier Foundation et al. **Santa Clara Principles on transparency and accountability in content moderation**. Available at: https://santaclaraprinciples.org/pt/cfp/. Acesso 20 dez. 2022.

ESTARQUE, Marina; ARCHEGAS, João Victor. **Redes sociais e moderação de conteúdo:** criando regras para o debate público a partir da esfera privada, Instituto de Tecnologia e Sociedade (ITS): Rio de Janeiro, 2021, p. 20. Available at: https://itsrio.org/pt/publicacoes/redes-sociais-e-moderacao-de-conteudo/ Accessed: 19 dez. 2022.

GAMA, Sophia. Guerra de desinformação: as fake news nas eleições de 2018. **Câmara Municipal de Curitiba,** 15 de julho 2022. Available at: https://www.curitiba.pr.leg.br/

informacao/noticias/guerra-de-desinformacao-as-fake-news-nas-eleicoes-de-2018. Accessed: 19 dez. 2022.

GORWA, Robert. The platform governance triangle: conceptualising the informal regulation of online content. **Internet Policy Review,** v. 8, n. 2, jun./2019. Available at: https://policyreview.info/articles/analysis/platform-governance-triangle-conceptualising-informal-regulation-online-content. Acesso em 30 ago. 2022.

GORWA, Robert. What is platform governance? **Information, Communication & Society,** v. 22, n. 6, pp 854-871, 2019. Available at: https://www.tandfonline.com/doi/full/10.1080/1369118X.2019.1573914. Accessed: 24 ago. 2022.

HAIMSON, Oliver L.; DELMONACO, Daniel; NIE, Peipei; WEGNER, Andrea. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. In: ACM CONFERENCE ON HUMAN-COMPUTER INTERACTION, 2021. **Anais [...]** Association for Computing Machinery, 2021. Available at: https://dl.acm.org/doi/10.1145/3479610. Accessed: 19 dez. 2022.

HARTMANN, Ivar; IUNES, Julia. Fake news no contexto de pandemia e emergência social: os deveres e responsabilidades das plataformas de redes sociais na moderação de conteúdo online entre a teoria e as proposições legislativas. **RDP**,

Brasília, v. 17, n. 94, p. 388-414, jul./ago. 2020. Available at: https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/download/4607/Hartmann%3B%20Iunes%2C%202020. Acesso em 19 dez. 2022.

HARTMANN, Ivar Alberto Martins; SARLET, Ingo Wolfgang. Direitos fundamentais e direito privado: a proteção da liberdade de expressão nas mídias sociais. **Revista de Direito Público,** v. 16, n. 90, p. 85-108, dez. 2019. Available at: https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/3755. Accessed: 19 dez. 2022.

HOFMANN, Jeanette. Multi-stakeholderism in Internet governance: putting a fiction into practice. **Journal of Cyber Policy,** v.1, n. 1, pp. 29-49. Available at: https://www.tandfonline.com/doi/pdf/10.1080/23738871.2016.1158303. Accessed: 11 Jan. 2023.

KADRI, Thomas E.; KLONICK, Kate. Facebook V. sullivan: Public figures and newsworthiness in online speech. **Southern California Law Review**, v. 93, n. 37, pp. 37-99, 2019, p. 94. Available at: https://scholarship.law.stjohns.edu/cgi/viewcontent.cgi?article=1292&context=faculty_publications. Accessed: 30 ago. 2022.

KLONICK, Kate. The new governors: the people, rules, and processes governing online speech. **Harvard Law Review,** v. 131, n. 6. p. 1598-1669, abr. 2018. Available at: https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech/ Accessed: 30 ago. 2022.

KURTZ, Lahis Pasquali; DO CARMO, Paloma Rocillo Rolim; VIEIRA, Victor Barbieri Rodrigues. **Transparência na moderação de conteúdo**: tendências regulatórias nacionais. Belo Horizonte: Instituto de Referência em Internet e Sociedade, 2021. Available at: https://bit.ly/3xjAUka. Accessed: fev, 2022.

MARS, Amanda. Como a desinformação influenciou nas eleições presidenciais? **El País,** 25 de fevereiro 2018. Available at: https://brasil.elpais.com/brasil/2018/02/24/internacional/1519484655_450950.html. Accessed: 19 dez. 2022.

OLIVA, Thiago Dias; ANTONIALLI, Dennys Marcelo; GOMES, Alessandra. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. **Sexuality & Culture.** v. 25. n. 2, p. 700–732 abr. 2021. Available at: https://link.springer.com/article/10.1007/s12119-020-09790-w Accessed: 20 dez. 2022.

RODRIGUES, Gustavo Ramos. "A função de uma lei de proteção de dados é proteger a todos, inclusive aquele que coleta dados": o debate sobre a Lei Geral de Proteção de Dados e a ideologia da harmonia no Fórum da Internet no Brasil"). In: VII ENCONTRO NACIONAL DE ANTROPOLOGIA DO DIREITO, 2022, São Paulo. **Anais [...]**. São Paulo: NADIR/USP, 2021. v. 1. p. 12. Available at: bit.ly/3GBlZYx. Accessed: 11 Jan. 2023.

SILVA JUNIOR, LA; LEAO, MBC O software Atlas.ti como recurso para a análise de conteúdo: analisando a robótica no Ensino de Ciências em teses brasileiras. **Ciência & Educação,** Bauru , v. 24, n. 3, p. 715-728, set. 2018. Available at: https://www.scielo.br/j/ciedu/a/yBwC9L74v4vD3s4PwVXggsk/?lang=pt. Accessed: 21 dez. 2022.

STEIGER, Miriah et al. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In: CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 2021, Yokohama. **Anais [...]** Yokohama: Association for Computing Machinery, 2021. Available at: https://dl.acm.org/doi/pdf/10.1145/3411764.3445092. Accessed: 20 dez. 2022.

SARLET, Ingo Wolfgang; SIQUEIRA, Andressa de Bittencourt. Liberdade de expressão e seus limites numa democracia: o caso das assim chamadas "faew news" nas redes sociais em período eleitoral no Brasil. **Revista Estudos Institucionais,** v. 6, n. 2, pp. 534-578, maio/ago. 2020. Available at: https://www.estudosinstitucionais.com/REI/article/view/522/511. Accessed: 19 dez. 2022.

TWITTER. **Conselho de Trust & Safety.** 2022. Available at: https://about.twitter.com/pt/our-priorities/healthy-conversations/trust-and-safety-council. Acesso em 30 ago. 2022.

7. Appendix I - Interview script

Block I - Presentation and general topics

I. What is your area of training and main activity?

II. Have you ever acted, in research, in practice, in regulation, in relation to content moderation?
Explore: in what activities related to online content moderation, whether in practice, regulation or research, have you already acted?

Block II - Content moderation and interaction between the different actors involved

III. In your opinion, what are the main functions of content moderation? Explore: What activities do you consider to be content moderation? Why?

IV. Among all internet governance actors, is there any actor with greater responsibility for content moderation management?
Explore: why? What are the responsibilities? How can other actors cooperate with them?

V. In your perception, are there mechanisms to ensure society's monitoring of content moderation practices?
Explore: what would they be? Are they effective?

VI. Nowadays, there are private instances for reviewing content moderation decisions, like Facebook's Oversight Board, for example. In general, how do you evaluate the performance of these instances?
Explore: what's positive? And negative? How do you think this impacts the legitimacy of the procedure? And what is the role of the Judiciary as a possible instance of judicial review of content moderation?

VII. How do you evaluate the performance of third sector subjects and technical-scientific communities in regulatory processes for online content?
Explore: do you rate it as positive or negative? In what situations

is civil society most mobilized in the context of Internet Governance? Which organizations are mobilized?
If the evaluation is negative: do you think there is marginalization? Do you see any solution for this?
In case it's interesting: how do you assess the public debate about content moderation on social platforms today?

Block III - Regulatory Models

VIII. Thinking about the adequacy of moderation to human rights, how do you see the effectiveness of creating content rules by the platforms themselves?

IX. There are also recommendations without legal effect, such as the Santa Clara Principles, for example, which establish transparency standards for content moderation practices. Still thinking about adequacy to human rights, how do you evaluate the impact of this type of standards and recommendations without legal effect?

X. How do you evaluate the role of norms issued by state authorities for the adequacy of content moderation to human rights?

XI. On a scale of 0 to 10, how much are you in favor of creating laws that force platforms to remove certain types of content on their own? Why?
Explore: if you can't give a grade: what difficulties do you see in giving a grade in this case?
If the answer is intermediate: why not 0/10? If it depends on the case, what are they? What types of content? Would it be possible to reconcile this with the guarantee of user rights?

XII. Do you think law should establish specific guidelines for content moderation procedures?
Explore: What mechanisms could guarantee the effectiveness of this norm? What are these guidelines supposed to ensure?
How to consider platforms that have voluntary or smaller scale moderation?

XIII. How do you rate the use of automatic mechanisms for removing or restricting content? Why?

Explore: are there specific risks associated with them? If yes, which ones? How do you rate human review in relation to these mechanisms?

XIV. Have you ever heard the expression "regulated self-regulation"?
If yes: what sources? What do you mean by that? How do you rate this proposal? In your opinion, is it more or less effective than the others?
If not: there are authors who mention this term, based on German works, as a way of establishing standards for the internal regulation of content moderation procedures; How do you rate this proposal? In your opinion, is it more or less effective than the others?

XV. In your perception, how do you see the regulatory model of content moderation in force in Brazil today?
Explore: how do you evaluate the role of the actors involved in the Brazilian regulatory model? In your perception, would changes be necessary to reconcile content moderation and human rights? Which ones?
If it is interesting depending on the person: what would be the mechanisms to ensure that actors, including intermediaries, do not subvert this regime model?

XVI. Are you following the debate on PL 2.630 ("PL das Fake News")?
Explore: in the case of implementation of PL 2.630, pending in the Brazilian Legislative, how do you assess the effects on content moderation and user rights in Brazil? What are the necessary mechanisms for the effectiveness of this regulation? What are the pros and cons of the PL?
Alternative, if the person does not know about the PL: what are the necessary mechanisms to guarantee the effectiveness of the regulation on transparency in content moderation? What would be the pros and cons of these mechanisms?

XVII. Would you like to add any comments?

XVIII. Do you have any questions that were not asked that you think would be important?

8. Appendix II - List of codes used

1. Academia Sector

1. Government Sector

1. Private Sector

1. Third sector

2. MOD action ::

2. MOD action :: curation (modulate behavior/content ranking)

2. MOD action :: intervention (content removal/reduction/deletion)

2. MOD function

2. MOD function :: healthy environment (virtual space/receptive environment sanitization)

2. MOD function :: product

3. R1 - platforms :: resource (expertise/economic power)

3. R2 - State :: jurisdictional power (supervision/regulation/legislation)

3. Responsible for management – other

3. Responsible for management – government

3. Responsible for management – platform

4 MF – non-existent ::

4. MF – insufficient ::

4. MF – enough ::

5. IP – no legitimacy

5. IP – legitimacy not determined

5. IP – legitimacy yes

5. IP – negative :: limited autonomy

5. IP – positive :: other

5. IP – negative ::

5. IP – negative :: lack of transparency

5. IP – positive ::

5. IP – positive :: transparency

5. PJ – collision of rights

5. PJ – negative ::

5. PJ – negative :: competence

5. PJ – negative :: scale

5. PC – neutral

5 . PJ – positive

6. 3rd sector and academia – negative performance

6. 3S and academia – intermediation

6. 3S and academia – limit. ::

6. 3S and academia – limit. :: limited autonomy (resources/financing)

6. 3S and academia – mob ::

6. 3S and academia – mob :: focal episodes

6. 3S and academia – particularities (impact on platforms/pressure)

6. 3S and academia – qualification of the public debate

6. 3S and academia – proactivity

7. AR – difficulties

7. AR – negative

7. AR – positive :: need

8. RE – negative :: censorship

8. RE – negative :: competence

8. RE – positive

9. SL – negative

9. SL – negative :: limitations

9. SL – positive

9. SL – positive :: impact/pressure

10. Procedural guidelines – negative

10. Procedural guidelines – positive

11. Regulable content – negative

11. Regulable Content – Note

11. Regulable content – positive

12. AI – negative

12. AI – negative :: contextual error

12. AI – positive

12. AI – positive:: need

13. Human review – positive :: review errors

13. Human review – negative

13. Human review – positive

14. AR-R – conception

14. AR-R – negative

14. AR-R – positive

15. Current BR – non-existent

15. Current BR – relevant info

15. Current BR – insufficient

15. Current BR – negative

15. Current BR – positive

15. Current BR – sufficient

16. PL 2.630 – impacts

16. PL 2.630 – negative

16. PL 2.630 – negative :: parliamentary immunity

16. PL 2.630 – negative :: press remuneration

16. PL 2.630 – positive

16. PL 2.630 – positive :: due process

16. PL 2.630 – positive :: transparency

iris

INSTITUTE
FOR RESEARCH
ON INTERNET
AND SOCIETY