Scientific Research Publishing

Beijing Law Review, 2025, 16(2), 991-1010
https://www.scirp.org/journal/blr
ISSN Online: 2159-4635
ISSN Print: 2159-4627

Civil Liability of Social Media for User Statements and Moderation

André Felipe Santos de Souza, Flávia Moreira Guimarães Pessoa, Taiane Danusa Gusmão Barroso Sande

Law Department, Federal University of Sergipe (UFS), Aracaju, Brazil
Email: andrefelipe@academico.ufs.br, profaflaviapessoa@gmail.com, taianebarroso@gmail.com

How to cite this paper: Souza, A. F. S. d., Pessoa, F. M. G., & Sande, T. D. G. B. (2025). Civil Liability of Social Media for User Statements and Moderation. *Beijing Law Review, 16,* 991-1010.
https://doi.org/10.4236/blr.2025.162050

Received: March 5, 2025
Accepted: June 17, 2025
Published: June 20, 2025

Abstract

The article examines the current regulatory context of social media, focusing on civil liability and the moderation powers of platforms. It begins with a discussion on the constitutionalization of transnational corporations in light of Teubner's theory, highlighting how technology companies have become a kind of "meta-state" with their own normative frameworks. It then presents a comparative analysis of regulatory models in the United States and the European Union, followed by an analysis of the regulatory landscape established in Brazil through the Marco Civil da Internet (Law 12.965/2014), which sets a judicial notification rule for content removal. The article also explores international principles on moderation, including the Manila and Santa Clara Principles, which advocate for moderation based on transparency and accountability. The study concludes by emphasizing the complexity of the issue and the need for a multisectoral debate that enables a balance between freedom of expression and protection against harmful content, suggesting regulated self-regulation as a potential pathway for standardization within the Brazilian context.

Keywords

Social Media, Civil Liability, Content Moderation, Regulation, Freedom of Expression, Comparative Law

1. Introduction

As Capanema (2024: p. 19) aptly highlighted, "it is no longer possible to study Law without analyzing the profound effects of the Internet and Information Technologies on human life over the past 25 years." The internet has revolutionized access to information and the way people interact, bringing immense agility in disseminating information and facilitating communication between individuals across

different parts of the world. From the emergence of ARPANET, which connected a few American universities for data exchange, to the current concept of the internet as a means of communication among all users, permeated by social networks whose primary purpose is social interaction, a long journey has taken place. However, these advancements come with inherent challenges in this new way of acting in the virtual world, where opinions can gain immeasurable reach in a matter of seconds.

Inspired by Stanley Milgram's Six Degrees of Separation Theory, which states that individuals are six connections or friendships away from any other individual, the first social network, SixDegrees, emerged in 1997. The interest in the possibility of profiting from the internet attracted significant attention, driving the purchase of shares in ".com" companies. In 2000, the .com bubble burst. However, this did not kill the Web. On the contrary, "it gave rise to websites and services that moved users from the role of mere spectators to that of protagonists, allowing interactions on platforms" (Capanema, 2024: p. 33). The development of technology-enabled the emergence of "GAFAM", a group consisting of the largest technology companies—Google/Alphabet, Apple, Facebook/Meta, Amazon, and Microsoft—whose initials form its name. These companies generated immense wealth by monetizing their users' personal data (Capanema, 2024: p. 37). On the other hand, rumors and controversies involving cases of user rights violations are becoming increasingly common. In this context, the need for the regulation of social networks arises, along with the issue of civil liability for the improper use of user data, as well as for potential user statements due to a lack of moderation.

This research aims to analyze the current regulatory framework of social networks concerning civil liability and moderation. Initially, it will address the phenomenon described by Teubner as the constitutionalization of transnational companies through the creation of rules and terms of use adopted by content platforms. Next, the evolution of civil liability rules for content platforms regarding user statements in the United States and Europe will be examined. Following this, principles guiding the regulation of moderation powers of social networks will be analyzed. Finally, the topic will be discussed from the perspective of the laws currently in force in Brazil.

Given the contemporaneity of the topic, whose discussion extends across the academic community, civil society, and various countries, encompassing a wide range of perspectives, this study does not aim to predict the best proposal for regulating the liability of social networks. The objective is to critically analyze the models that have been adopted, especially in Brazil, with the aim of fostering a more mature debate on the issue. This is a pressing topic whose ideal solution is far from being reached, particularly due to the rapid pace of change. Thus, the subject will be addressed from the perspective that the discussion on the regulation of freedom of expression is complex, making it essential to promote a multistakeholder debate aimed at developing plausible, effective solutions that, above all, safeguard human rights in their broadest sense.

Bill 2.630/2020, although relevant to the subject under study, will not be addressed in this research, as the focus will be on models that are already in use. Moreover, the legal nuances surrounding the aforementioned bill could be the subject of a separate study. The breadth of the issues to be analyzed would not allow for its examination here with the depth that the matter requires. Thus, the analysis will be limited to existing regulations based on a bibliographic study that will employ the deductive method.

2. Constitutionalization of Transnational Companies

Teubner (2011: p. 17) addresses the growing constitutionalization of transnational companies through the interconnection between private and public codes of conduct. The eminent sociologist explores the thesis that these corporations are not only subject to state legal norms but also develop their own internal normative structures, performing functions typically associated with constitutions. There is, therefore, a juxtaposition between private codes (autonomous, self-imposed) and public codes (external regulations imposed by states or international organizations), ultimately creating a sort of constitutional order specific to these corporations. These codes function as self-regulation mechanisms aimed at legalizing fundamental social principles and establishing boundaries for corporate actions. This reflects a phenomenon of "double reflexivity," in which both codes feed into and transform each other, forming what Teubner calls "binary metacodes."

For Teubner (2011: pp. 18-21), the self-constitutionalization of transnational companies does not imply subordination to a hierarchical system between the public and private spheres but rather the formation of qualitatively different normative networks. According to the author, these codes function as true transnational constitutions, establishing self-limiting rules that are essential for dealing with the challenges of global governance in a world where state sovereignty is often insufficient to regulate the complexity of global corporate operations. Thus, constitutionalization is not a process restricted to the nation-state. Transnational social orders, such as transnational companies, also develop their own constitutions. These constitutions are not merely metaphorical. On the contrary, they perform functions similar to state constitutions in terms of establishing fundamental principles, limits on power, and internal control mechanisms, regulating the central decision-making mechanisms (Teubner, 2011: pp. 25-26).

This is a particularly relevant dynamic in the context of growing global interdependence, where the combination of private and public regulations can promote greater accountability and social responsibility on the part of transnational corporations. In the current scenario of large transnational companies, where the debate on transnational governance and the transformation of the role of the State in the face of the influence of non-state actors is especially pertinent, Teubner offers an innovative perspective on the constitutionalization of the global corporate sphere, challenging traditional concepts of legality and sovereignty. This perspective helps us reflect on the dimensions that the issue of social network ac-

countability imposes upon us.

Teubner (2011: p. 20) highlights that the development of these transnational constitutions is rooted in specific historical contexts. The growth of global trade, financial and environmental crises, and the pressure from NGOs and governments for greater corporate social responsibility are factors that have contributed to the need for transnational companies to adopt complex codes of conduct that perform constitutional functions. Regarding the autonomy and reflexivity of corporations, the concept of "double reflexivity" is essential for understanding the interaction between public and private codes (Teubner, 2011: pp. 24-27). Corporations, by formulating their own codes of conduct, become responsible not only for regulating their internal activities but also for responding to the demands for social and environmental responsibility imposed by public and international actors. This reflects a capacity for legal and moral self-limitation.

There is a continuous tension between self-imposed private norms and public regulations, and although both spheres complement each other, Teubner (2011: pp. 28-30) suggests that the balance between these forces is in constant negotiation. This tension shapes the type of constitutionalism that emerges, where corporations are compelled to adapt their normative structures both to internal expectations and external pressures. Regarding the legal nature and legitimacy of Corporate Codes, Teubner (2011: pp. 31-33) investigates whether these codes can be considered "law" in the traditional sense, arguing that, although corporate codes of conduct are not binding in the same way as state laws, they exert significant normative power over corporations and their stakeholders, generating real effects on their global operations. All these points reinforce Teubner's argument that transnational corporations, by self-constitutionalizing, not only respond to governance demands but also become central actors in the creation of global legal norms.

The topic is of utmost importance, as the moderation of content platforms can turn them into highly powerful political regulators and even the world's largest censors, given the close relationship between network moderation and freedom of expression. While there is no consensus on the extent of this power, its role as a significant facilitator and mediator of public debate is undeniable.

The phenomenon of self-constitutionalization by transnational corporations, as theorized by Teubner, presents significant tensions when these corporate norms interact with the legal systems of the jurisdictions in which such companies operate. These internal "constitutional" frameworks—typically embedded in terms of service, community standards, and algorithmic governance policies—frequently assert normative authority over user behavior in ways that may conflict with local regulations and cultural values.

One prominent area of friction arises in content moderation. For instance, platforms like Facebook or YouTube may remove content that violates their global community standards even when such content is legally protected under national constitutions that strongly safeguard freedom of expression, such as

the First Amendment in the United States or Article 5 of the Brazilian Constitution. Conversely, in jurisdictions with stricter speech regulations—such as Germany's NetzDG or the EU's Digital Services Act—platforms must comply with swift content takedown mandates, sometimes in ways that conflict with their own moderation guidelines or raise concerns about overcompliance and censorship.

These tensions reveal a deeper structural conflict between the universalizing logic of platform governance and the pluralism of state-based legal orders. While platforms strive for global scalability and uniform rules, states demand compliance with their specific legal and cultural frameworks. This creates a regulatory grey zone in which platforms act as quasi-judicial entities, often balancing between legal mandates, their own policies, and international human rights standards. The risk, as Teubner warns, is that corporate constitutional norms may begin to override or displace democratic legal systems, especially in countries with weaker enforcement capabilities. This underscores the need for hybrid governance mechanisms—such as co-regulation and independent oversight bodies—that can reconcile private and public authority in a way that ensures legitimacy, accountability, and respect for local constitutional principles.

3. A Brief Overview of the Civil Liability of Social Networks in the United States

According to Rodrigues (2020), contemporary conflicts over the regulation of content moderation have their roots in the evolution of internet-related discussions that began decades ago, marked by tensions between freedom and control. In the 1990s, the prevailing view was that the internet should be a space of self-governance, free from state regulation, based on ideals of decentralization and individual liberties. This vision was embodied in John Perry Barlow's "Declaration of the Independence of Cyberspace" in 1996, which rejected any state authority over the "digital territory." The declaration was a response to the Telecommunications Act of 1996, a legislative reform in the United States that introduced the Communications Decency Act (CDA), which sought to restrict certain types of content, such as pornography and online threats, in an initial attempt to regulate the internet.

Estarque et al. (2024) recount that in the United States, two decisions from the 1990s shaped the issue of platform liability for content posted by third parties on digital platforms such as Facebook and Twitter. In Cubby v. CompuServe (1991), CompuServe was deemed a distributor, merely providing space for third-party content without editorial intervention, which exempted it from legal responsibility for those contents. Conversely, in Stratton Oakmont v. Prodigy Services (1995), Prodigy, which actively moderated some posts, was considered liable for the content, similar to a publisher, due to its direct involvement in moderation decisions.

These cases created a dilemma: moderating offensive content increased the risk of liability while not intervening provided legal protection. To resolve this tension,

Congress passed Section 230 of the Communications Decency Act of 1996, which determines that digital platforms are not considered publishers and, therefore, cannot be held liable for third-party content. The "Good Samaritan" clause of Section 230 extended immunity even to platforms that choose to moderate offensive content, allowing self-regulation without the risk of lawsuits. Although the CDA was deemed unconstitutional by the U.S. Supreme Court for violating freedom of expression, the only section that survived was Section 230, a legal milestone that established the "safe harbor" for internet service providers, which would not be held responsible for third-party content. This provision removed the duty of surveillance from providers and prevented them from assuming a preventive censorship position, allowing the internet to develop as a relatively free space for expression.

This initial legislation shaped the current moderation framework, as it defined the position of providers as intermediaries rather than censors of online content. Contemporary disputes over content moderation revisit these same principles, confronting the need to regulate the digital environment with the preservation of freedom of expression in a context where challenges have expanded to include disinformation and hate speech. It is worth noting that Section 230 of the Communications Decency Act (CDA) not only shielded internet providers from liability for third-party content but also, through the "Good Samaritan" clause (§230, c, 2), allowed them to voluntarily and in good faith moderate content without incurring legal responsibility. This immunity was interpreted by the U.S. Supreme Court as an incentive for providers to take a proactive stance in removing content deemed harmful or obscene.

This balance, which granted providers the right but not the obligation to moderate content, shaped the internet over the following decades, guiding business decisions on moderation on a global scale and influencing policies in other countries, such as Brazil. Thus, the provision of Section 230 laid the foundation for current moderation practices, allowing the creation of a digital environment where providers have the freedom to act against problematic content but are not legally required to continuously monitor everything that is posted.

Teffé and Souza (2024: pp. 27-29) explain that, in the United States, the discussion on the responsibility of digital platforms is currently governed by the Communications Decency Act of 1996, which prevents liability for content published by third parties as well as for moderation actions that remove posts deemed harmful. However, some landmark cases brought to court have been judged under the Justice Against Sponsors of Terrorism Act (JASTA) of 2016. On that occasion, the U.S. Supreme Court ruled that there was insufficient evidence to classify the platforms' conduct as aiding and abetting terrorist acts, stating that holding them liable would be equivalent to holding telephone companies responsible for the content of phone calls. It was emphasized that algorithms did not favor ISIS content, as it received the same treatment as other content. The connection between ISIS content and certain user profiles did not transform passive assistance into active

support.

It is noteworthy that the U.S. Supreme Court did not address the issue of platform liability, as it focused on determining whether failure to moderate accounts and content could constitute "aiding and abetting" terrorist acts. Thus, the issue of digital platform liability may still be addressed under different allegations, which have the potential to lead to different outcomes.

Bowers and Zittrein (2020), in examining the evolution of digital platform governance, classify it into three periods or "eras": Rights, Public Interest, and Process, proposing a reconfiguration of the approach to online content regulation. In the Era of Rights, Section 230 of the Communications Decency Act (CDA 230), enacted in the United States in 1996, established a regime of immunity for digital intermediaries, excluding them from civil liability for user-generated content (UGC), except in criminal cases. Its purpose was to protect online freedom of expression and foster the development of a public discourse space free from the interference of external entities.

Over the years, with the rise of disinformation and social polarization driven by algorithms, the Era of Public Interest emerged, questioning the responsibility of platforms in the use and amplification of content harmful to social and institutional integrity. In this new context, platforms are expected to adopt a more proactive stance to mitigate the harmful effects of viral content that can interfere with elections, incite violence, and undermine trust in democratic institutions.

Finally, the Era of Process proposes an advancement in content governance with a focus on transparency and the legitimacy of decision-making mechanisms. Instead of opaque models centered on public relations, a governance model is proposed that adopts "procedural accountability" practices, in which platforms would be responsible to users through a fiduciary duty and, eventually, through the delegation of complex decisions to independent entities. This era aims to build a content governance model that balances individual rights with the public interest, considering the ethical and social impact of moderation practices. Thus, the regulation of platforms must go beyond revisions to CDA 230, incorporating a hybrid governance structure that harmonizes freedom of expression with responsibility regarding the broad and collective effects of digital content, thereby preserving the balance between fundamental rights and the protection of the public interest in an era of global digital communication.

In practical terms, the legal immunity granted to platforms under Section 230 of the Communications Decency Act has profoundly influenced both content moderation practices and user behavior on major social media platforms in the United States. Freed from the risk of being held liable for third-party content, platforms such as Facebook, YouTube, Twitter (now X), and TikTok have developed extensive internal moderation systems, often involving automated tools and human review. These systems allow them to proactively remove or deprioritize content deemed harmful without fear of legal repercussions for either action or inaction.

This legal environment has also shaped how users engage with platforms. Facing evolving and often opaque moderation rules, users adapt their speech to avoid triggering automated removal—frequently resorting to euphemisms or coded language. At the same time, others have learned to exploit algorithmic behaviors to amplify divisive or sensational content, taking advantage of engagement-driven recommendation systems. This dynamic, rooted in the CDA's framework, has led to a paradoxical effect: while enabling greater freedom of expression by limiting government interference, it has also concentrated enormous discretionary power in the hands of private companies, which now operate as the de facto moderators of public discourse.

As Bowers and Zittrein (2020) suggest in their classification of governance eras, the current "Process Era" reflects an increasing demand for procedural legitimacy, transparency, and user recourse. Platforms are gradually incorporating appeals mechanisms, transparency reports, and oversight boards, but these initiatives remain uneven and largely voluntary. Consequently, Section 230 has not only underpinned a unique regulatory approach but also fostered a behavioral ecosystem where the boundaries of acceptable expression are set, enforced, and sometimes contested by corporate actors rather than democratic institutions.

4. Panorama in Europe

Regarding Europe, Teffé and Souza (2024: pp. 30-32) report that the Digital Services Act (DSA) of 2022 and the Digital Markets Act have come into effect. The DSA regulates the obligations of digital services with a view to protecting users and fundamental rights, establishing measures to combat illegal online content, and allowing users to report such content for identification and removal. It also provides for the possibility of contesting content moderation decisions and rules on transparency, including better information on the terms and conditions for suggesting content or products to users.

Under the Digital Services Act (DSA), content moderation obligations for platforms—particularly Very Large Online Platforms (VLOPs)—are grounded in the need to assess and mitigate systemic risks related to the dissemination of illegal content. However, the DSA does not establish a universal definition of illegality; instead, it defers to the national laws of EU member states to determine what constitutes illegal speech, creating a multi-layered legal environment in which platforms must operate.

To navigate this complexity, platforms employ a combination of automated detection systems, flagging mechanisms, and human moderation, supported by internal guidelines that map national legal categories onto their global content standards. These guidelines often rely on typologies of content—such as hate speech, terrorist propaganda, copyright infringement, or defamation—and are continuously updated in response to local jurisprudence and regulatory guidance. In this context, ensuring consistency requires the implementation of internal audit systems, localized moderation teams with cultural and legal expertise, and oversight mech-

anisms to prevent both under-enforcement and over-removal.

The DSA mandates transparency and due process as key principles: platforms must explain the rationale for content removals, notify affected users, and offer accessible appeal mechanisms. Furthermore, to promote consistency, the regulation requires platforms to conduct risk assessments and independent audits, submit transparency reports, and engage in structured dialogue with regulatory authorities and civil society. These obligations are designed not only to protect users' rights but also to create a harmonized framework across the EU, where platform practices reflect both the diversity of national legal traditions and the shared commitment to the rule of law and fundamental rights.

Furthermore, it establishes obligations regarding the protection of minors on platforms, including age verification tools and parental controls, as well as mechanisms to assist minors in reporting abuse and seeking support, in addition to additional obligations for very large search engines. It also promoted the adoption of systemic risk detection measures, supervision through independent audits, and risk management measures, as well as cooperation with other service providers by initiating or adhering to codes of conduct and self-regulatory measures, in addition to the development of awareness-raising initiatives.

Obligations were established to mitigate risks related to disinformation, electoral manipulation, violence against women, and harm to children and adolescents, as well as to develop crisis response mechanisms and ensure special transparency in advertising and recommendation systems. Additionally, the use of strategies that mislead users into making choices they would not otherwise make (dark patterns) was prohibited. The European Commission was designated as the primary regulator for very large platforms and search engines, while other services remain under the supervision of the competent authorities of each Member State.

According to Estarque et al. (2024: pp. 19-21), in Germany, the NetzDG (Network Enforcement Act) was approved in 2017 by the Bundestag (German Parliament), requiring platforms to remove "manifestly illegal" content within 24 hours after notification. For "illegal" content, the deadline is up to seven days. In case of non-compliance, fines can reach 50 million euros. The NetzDG mandates that companies themselves interpret German law to determine what qualifies as illegal, creating incentives for the preventive removal of content that might fall into these categories, potentially limiting freedom of expression.

In 2019, France proposed a regulatory framework aimed at striking a balance between punishment and prevention, suggesting that platform content moderation should be more transparent and aligned with the public interest. This report emphasizes that French authorities should play a role in ensuring that the moderation process of social networks is informed by the public interest and not just by the private interests of platforms. The French proposal includes the establishment of an independent body responsible for overseeing transparency obligations and user integrity, fostering self-regulation within minimum parameters defined by the State.

In the same year, the United Kingdom launched the Online Harms White Paper, proposing a regulatory system led by an independent body. This body would be responsible for setting standards to ensure user safety on social networks while also protecting freedom of expression. The British document promotes the creation of a "duty of care" for platforms, encouraging a culture of transparency, trust, and accountability. Among the regulator's responsibilities are the development of codes of best practices, oversight of the implementation of the duty of care, and conducting educational campaigns on the challenges of online freedom of expression.

The German approach is more punitive and focused on local obligations, requiring platforms to interpret national legislation and promptly remove illegal content, which creates a risk of preventive censorship. In contrast, the French and British solutions adopt a procedural approach, seeking to make the moderation process more transparent and accountable in a manner more aligned with the interests of global platforms, which prefer universal guidelines and self-regulatory procedures monitored by independent bodies. While the NetzDG imposes a local and immediate obligation to remove content, the French and British models emphasize a "duty of care" and transparency that can be applied globally, offering flexibility that facilitates the adaptation of technology companies to different legal contexts.

5. International Principles for Content Platform Moderation

In March 2015, international organizations, on the occasion of the RightsCon conference, published the Manila Principles on Internet Liability (2015), a list of six principles aimed at protecting freedom of expression in the digital space. The document takes into account that, in the context of digital communications, the role of intermediaries, including internet providers, social networks, and search engines, is central, as they facilitate the flow of information. It also recognizes that regulatory policies regarding their liability for third-party content have direct effects on fundamental user rights, including freedom of expression, the right to association, and privacy.

The formulation of the principles aims to protect freedom of expression and create a balanced innovation environment that meets the needs of different sectors, both governmental and private. It proposes a framework guided by international human rights principles and best practices to ensure fair accountability, which aligns with international legal instruments such as the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights and provides guidelines for regulators and intermediaries to adopt policies that respect these rights.

The main criticism of the current legislation was recognized, pointing to intermediary liability policies that are often inflexible and coercive, ignoring the principles of necessity and proportionality, contributing to censorship and human rights abuses, limiting free expression, and creating an environment of insecurity that, in addition to impacting individual rights, inhibits progress and innovation

in the digital sphere. In the meantime, to build a liability system that is harmonious and respects user rights, the creation of interoperable and consistent norms across different jurisdictions is proposed. This approach aims to ensure a safer and more innovative digital environment aligned with international principles, including the United Nations Guiding Principles on Business and Human Rights. Therefore, the norms must observe the following principles:

1. Intermediaries must be protected by law from liability for content produced by third parties;
2. Content removal should not be requested without an order from a judicial authority;
3. Requests for content restriction must be clear, unambiguous, and follow due process;
4. Laws, orders, and content restriction practices must comply with the tests of necessity and proportionality;
5. Laws, policies, and content restriction practices must respect due process;
6. Transparency and accountability must be integrated into laws, policies, and content restriction practices.

Already in 2018, when the European Union proclaimed the Code of Practice on Disinformation, establishing the first normative self-regulation standards, academic entities and third-sector organizations from various countries developed the Santa Clara Principles on Transparency and Accountability in Content Moderation (Santa Clara Principles, 2018). This document describes best practices to ensure minimum levels of transparency and accountability in the private sector. Among the fundamental principles, the following were listed: human rights and due process, understandable rules and policies, cultural competence, state involvement in content moderation, integrity, and explainability.

The Human Rights and Due Process Principle requires that companies integrate human rights and due process considerations into moderation stages, ensuring transparency and reliability. Automated moderation methods must be used with high accuracy, allowing users to access clear channels to contest decisions about content and accounts, guaranteeing fairness in the process. The Understandable Rules and Policies Principle establishes that content moderation rules must be disclosed in an accessible, centralized, and easily understandable manner, allowing users to understand the criteria for possible actions on content and promoting transparency and predictability in the use of platforms.

According to the Cultural Competence Principle, moderation decisions must consider the cultural, linguistic, and political diversity of users. It aims to ensure that moderators are trained to understand content in context and that appeal and notification processes are provided in the user's language, preventing discrimination that may arise from cultural or regional barriers. The State Involvement in Content Moderation Principle highlights the risks to user rights when the State interferes in moderation, whether by creating policies or ordering content remov-

als. It underscores the need for caution and oversight regarding state pressures or interests that may compromise neutrality and user rights.

Finally, the Integrity and Explainability Principle requires that effectiveness and impartiality be ensured in both automated and human moderation systems. It also includes the need for regular evaluations and external audits, promoting public accountability through the disclosure of data on system accuracy and a transparent appeal process.

The document also addresses fundamental principles for content moderation, dividing responsibilities between companies and governments to ensure transparency, respect for human rights, and freedom of expression. First, transparency is highlighted as central in the "Numbers" principle, requiring companies to publish data on moderation actions, allowing the public and researchers to understand the reasons behind blocks or removals. The Notification principle emphasizes that every user affected by a disciplinary measure must be informed of the specific reason for the action taken against their content, with clearly defined exceptions (such as spam or phishing), ensuring that everyone understands the reasons for the restrictions applied. Under the Appeal principle, users can contest decisions, with priority given to content removal and account suspension cases, as they have a greater impact on human rights, particularly freedom of expression.

The principles for governments and other governmental actors highlight the responsibility of States to respect the freedom of expression standards established in international instruments, prohibiting them from using moderation systems to censor groups or individuals and preserving the autonomy of platforms. Regarding government transparency, States must report all interventions in moderation decisions in detail, including requests and orders for content removal or account suspension, identifying the legal basis for these actions.

Finally, it is suggested that governments facilitate platform transparency, avoiding obstacles that prevent companies from disclosing information about their moderation decisions and promoting regulatory and non-regulatory initiatives to enhance the clarity of control and governance processes on platforms. These principles, taken together, aim to establish balanced content moderation, where companies and governments have well-defined roles, respecting individual rights and promoting transparency in their actions.

6. Moderation and Civil Liability in Brazil

Currently, in Brazil, liability for damages arising from third-party content is governed by Article 18 of the Marco Civil da Internet (MCI) (Brazil, 2014), which states that "the internet access provider shall not be held civilly liable for damages resulting from content generated by third parties." Article 19, in turn, establishes an exception, providing for the possibility of civil liability in cases where, after a specific court order, the provider fails to take the necessary measures to make the infringing content unavailable within the scope and technical limits of its service and within the specified deadline, unless otherwise provided by law. The Judicial

Notice and Takedown regime was adopted.

The purpose of the norm is to ensure freedom of expression and other guarantees provided in Article 5 of the Federal Constitution, as well as to prevent censorship, assigning the Judiciary the role of assessing whether content should be made unavailable. The court order must clearly and specifically identify the content deemed infringing, allowing for its unequivocal location. In cases of copyright and related rights violations, a specific legal provision is required. For the anticipation of the effects of the requested relief, in addition to the requirements of the plaintiff's claim's verisimilitude and well-founded fear of irreparable harm or harm that is difficult to remedy, it is necessary to assess the existence of unequivocal evidence of the fact and the collective interest in keeping the content available on the internet.

It is observed, therefore, that Brazil, regarding the liability of internet application providers, has established the rule that liability is only configured in cases of non-compliance with a court order for content removal. This norm was the subject of debate within the Superior Court of Justice (STJ), which recognized the normative deficiency of the Marco Civil da Internet in protecting human dignity, as it conditions the liability of providers on the requirement of a specific court order. (Brasil, 2022) The court acknowledged the prevalence of protective norms for children's and adolescents' rights, as reflected in the following ruling summary:

> DIREITO CIVIL, INFANTOJUVENIL E TELEMÁTICO. PROVEDOR DE APLICAÇÃO. REDE SOCIAL. DANOS MORAIS E À IMAGEM. PUBLICAÇÃO OFENSIVA. CONTEÚDO ENVOLVENDO MENOR DE IDADE. RETIRADA. ORDEM JUDICIAL. DESNECESSIDADE. PROTEÇÃO INTEGRAL. DEVER DE TODA A SOCIEDADE. OMISSÃO RELEVANTE. RESPONSABILIDADE CIVIL CONFIGURADA.
>
> 1. O Estatuto da Criança e do Adolescente (art. 18) e a Constituição Federal (art. 227) impõem, como dever de toda a sociedade, zelar pela dignidade da criança e do adolescente, colocando-os a salvo de toda forma de negligência, discriminação, exploração, violência, crueldade e opressão, com a finalidade, inclusive, de evitar qualquer tipo de tratamento vexatório ou constrangedor.
>
> 1.1. As leis protetivas do direito da infância e da adolescência possuem natureza especialíssima, pertencendo à categoria de diploma legal que se propaga por todas as demais normas, com a função de proteger sujeitos específicos, ainda que também estejam sob a tutela de outras leis especiais.
>
> 1.2. Para atender ao princípio da proteção integral consagrado no direito infantojuvenil, é dever do provedor de aplicação na rede mundial de computadores (Internet) proceder à retirada de conteúdo envolvendo menor de idade—relacionado à acusação de que seu genitor havia praticado crimes de natureza sexual—logo após ser formalmente comunicado da publicação ofensiva, independentemente de ordem judicial.
>
> 2. O provedor de aplicação que, após notificado, nega-se a excluir publicação

ofensiva envolvendo menor de idade, deve ser responsabilizado civilmente, cabendo impor-lhe o pagamento de indenização pelos danos morais causados à vítima da ofensa.

2.1. A responsabilidade civil, em tal circunstância, deve ser analisada sob o enfoque da relevante omissão de sua conduta, pois deixou de adotar providências que, indubitavelmente sob seu alcance, minimizariam os efeitos do ato danoso praticado por terceiro, o que era seu dever.

2.2. Nesses termos, afigura-se insuficiente a aplicação isolada do art. 19 da Lei Federal n. 12.965/2014[1], o qual, interpretado à luz do art. 5º, X, da Constituição Federal, não impede a responsabilização do provedor de serviços por outras formas de atos ilícitos, que não se limitam ao descumprimento da ordem judicial a que se refere o dispositivo da lei especial.

3. Recurso especial a que se nega provimento. (REsp n. 1.783.269/MG, relator Ministro Antonio Carlos Ferreira, Quarta Turma, julgado em 14/12/2021, DJe de 18/2/2022)[2].

In this case, it was understood that liability for omission arose from the violation of the provisions established in Article 18 of the Statute of the Child and Adolescent (ECA), which includes the legal obligation to respect the physical, psy-

[1]12.965/2014 (Marco Civil da Internet) que determina a necessidade de prévia e específica ordem judicial de exclusão de conteúdo para a responsabilização civil de provedor de internet, websites e gestores de aplicativos de redes sociais por danos decorrentes de atos ilícitos praticados por terceiros. https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm

[2]Translated as: **CIVIL, CHILD AND ADOLESCENT, AND TELECOMMUNICATION LAW. APPLICATION PROVIDER. SOCIAL NETWORK. MORAL AND IMAGE DAMAGES. OFFENSIVE PUBLICATION. CONTENT INVOLVING A MINOR. REMOVAL. JUDICIAL ORDER. UNNECESSARY. FULL PROTECTION. DUTY OF THE WHOLE SOCIETY. RELEVANT OMISSION. CIVIL LIABILITY ESTABLISHED.**

1. The Child and Adolescent Statute (Article 18) and the Federal Constitution (Article 227) impose, as a duty of the entire society, the obligation to safeguard the dignity of children and adolescents, protecting them from all forms of neglect, discrimination, exploitation, violence, cruelty, and oppression, including preventing any type of humiliating or distressing treatment.

1.1. The protective laws concerning children's and adolescents' rights have an extremely special nature, belonging to a category of legal provisions that extend through all other norms, with the function of protecting specific subjects, even if they are also under the protection of other special laws.

1.2. To fulfill the principle of full protection enshrined in child and adolescent law, it is the duty of an application provider on the global computer network (Internet) to remove content involving a minor—related to allegations that their parent had committed sexual crimes—immediately upon being formally notified of the offensive publication, regardless of a judicial order.

2. An application provider that, after being notified, refuses to remove an offensive publication involving a minor must be held civilly liable, and compensation must be imposed for the moral damages caused to the victim of the offense.

2.1. Civil liability, in this context, must be analyzed from the perspective of the relevant omission in its conduct, as it failed to take measures that were undoubtedly within its reach to minimize the effects of the harmful act committed by a third party, which was its duty.

2.2. In this regard, the isolated application of Article 19 of Federal Law No. 12.965/2014 is insufficient, as it must be interpreted in light of Article 5, X, of the Federal Constitution. This interpretation does not preclude the liability of the service provider for other forms of unlawful acts, which are not limited to noncompliance with the judicial order referred to in the special law.

3. Special appeal denied.

(**REsp No. 1.783.269/MG, Reporting Justice Antonio Carlos Ferreira, Fourth Panel, judged on 12/14/2021, DJe of 02/18/2022**)

chological, and moral integrity of children and adolescents, establishing "a duty to act, directed at all members of society, who become agents in protecting the rights of minors, within reasonable and possible limits". This obligation resonates with Article 222 of the Federal Constitution, which enshrines the principle of the absolute priority of children and adolescents. On that occasion, Minister Marco Buzzi was in the minority, emphasizing that "the legislator expressly chose to grant the Judiciary the power to control the content that should be removed" so that any compensation should fall on the individual responsible for the publication (p. 19).

The issue regarding the normative deficiency of Article 18 of the Marco Civil da Internet (MCI) was included in the list of topics of general repercussion by the Supreme Federal Court (STF) under Theme 987, which discusses the constitutionality of Article 19 of Law No. 12.965/2014 (Marco Civil da Internet). This provision establishes the requirement of a prior and specific court order for content removal as a condition for the civil liability of internet providers, websites, and social media application administrators for damages resulting from unlawful acts committed by third parties. However, the matter has not yet been adjudicated.

Article 21 of the Marco Civil da Internet (MCI) also provides for the subsidiary liability of the internet application provider that makes third-party content available in cases of privacy violations resulting from the unauthorized disclosure of images, videos, or other materials containing scenes of nudity or private sexual acts. However, such liability is only recognized if, after receiving an extrajudicial notification from the participant or their legal representative, the provider fails to diligently take action within the technical limits of its service to make the content unavailable. It is the responsibility of the notifying party to provide, under penalty of nullity, elements that allow for the specific identification of the material allegedly violating the participant's privacy and to verify their legitimacy in submitting the request.

Article 20 of the Marco Civil da Internet (MCI) (Brazil, 2014) regulates content moderation, imposing on the internet application provider the responsibility to inform the user responsible for the content about the reasons and information regarding its removal, ensuring their right to adversarial proceedings and full defense, whenever contact information is available, except in cases of explicit legal provisions or a specific judicial order to the contrary. The user is granted the right to request the replacement of the content that was made unavailable due to the stated reason or the judicial order that justified its removal.

Teffé and Souza (2024: p. 32) highlight the importance of constructing an interpretation of the Marco Civil da Internet that aligns with contemporary reflections on content moderation, international human rights norms, and the General Data Protection Law (LGPD), seeking strategies that respect the duty of transparency, accountability, and the protection of communicative freedoms, free enterprise, and innovation. (Brasil, 2018) It is necessary that regulatory norms are not disproportionate to avoid legal barriers and legal uncertainty, thus fostering competition and enabling new platforms to introduce innovative and disruptive solu-

tions. However, they warn that while moderation rules are important in combating the spread of extremist, hate, or violent speech, a disproportionate increase in platform liability for third-party content could lead to severe censorship, infringing on freedom of expression and the diversity of opinions online, resulting in the removal of controversial content and profiles.

Thus, they identify four possible models and systems for holding providers and internet platforms liable for third-party content. The first is liability only for platforms that actively moderate content. The second would allow for the liability of platforms that know or should have known about the existence of illegal or harmful content. The third would involve adopting an immunity system similar to the "Good Samaritan" clause (a legal principle that protects online platforms and service providers in content moderation, allowing for self-regulation and good faith in their moderation policies—aimed at facilitating the free flow of information and protecting providers from legal liability for user-generated content, as long as they act in good faith when removing inappropriate content). The fourth model would focus on holding platforms accountable in cases where they fail to identify large-scale systemic risks to rights (Teffé & Souza, 2024: p. 33).

It is important to emphasize that the Marco Civil da Internet does not exclude the possibility for application providers to establish their own requirements for content removal through usage policies, and they may be held liable for failing to comply with their own operational rules. The moderation model, therefore, may arise both from the platform's own constitutionalizing norms and from state policies and regulations.

In any case, the logic of harm prevention must go hand in hand with the protection of constitutional freedoms, ensuring that immunities and duties are upheld. Since moderation and liability are two sides of the same issue, it is necessary to consider that the greater the liability for third-party content, the greater the power of moderation and, consequently, the stronger and more powerful the constitutions formulated by major content provider companies through their rules and terms of use. It is essential to reflect on whether the decision on which content violates the limits of freedom of expression and other fundamental rights should increasingly be entrusted to digital platform departments.

7. Perspectives and Possibilities for Regulation

Teffé and Souza (2024: p. 36) foresee that a possible solution in Brazil could be the expansion of legal provisions allowing content removal through extrajudicial notification based on clear and specific parameters, considering this approach more aligned with the Brazilian experience than the importation of foreign models.

In the document titled "Contributions for a Democratic Regulation of Major Platforms to Ensure Freedom of Expression on the Internet," it is proposed that content platforms should not be legally responsible for third-party content, provided that they do not engage in modifications or editing of the content and comply with judicial orders or official authority directives that follow due process. Li-

ability would arise only in cases of action or negligence by platforms in prioritizing or actively promoting content that may harm the rights of third parties. However, even in such cases, liability should not be of an objective nature. In summary, platform liability should be limited to cases of active involvement or significant omissions that constitute deviations from the established principles (Intervozes et al., 2021).

Any model to be adopted must be guided by prudence and caution, carefully balancing the uncertainty regarding content retention, the risk of censorship by large transnational conglomerates, and excessive immunity for platforms. It is essential to reflect on who should be responsible for moderating speech that does not directly target any individual or group but remains controversial.

It is necessary to maintain a balance between innovation and user rights. Liability in content moderation practices by companies can encourage a responsible balance between security, privacy, and freedom of expression. National regulatory approaches should respect the global nature of the internet, promoting interoperability between regulators and norms without imposing national standards on citizens of other countries. Regulators must consider the impact

of their decisions on freedom of expression, ensuring compliance with Article 19 of the International Covenant on Civil and Political Rights (ICCPR).

Understanding technological capabilities and limitations is essential, ensuring that internet companies have the flexibility to innovate, as one approach may not work universally for all platforms or content. Regulators must assess the severity and prevalence of harmful content, its legal status, and the efforts already undertaken to address it, applying proportional and necessary measures. These guidelines aim to build a regulatory framework that balances the protection of rights and the promotion of innovation in a globalized digital environment. It is necessary to keep in mind, as Neto (2024: p. 111) emphasized, that the strict liability of application providers could render the exercise of freedom of expression unfeasible, leading to prior censorship of content and information available on the internet.

For Dutra (2024: p. 595), the best path forward is regulated self-regulation, where the State would not regulate content itself but rather establish general rules for the creation of self-regulation mechanisms by providers, with the participation of civil society, private companies, and the State. State norms could also include the possibility of imposing sanctions in cases of violations of the duties imposed by law, ranging from warnings or fines to the prohibition of business operations in the country.

The ongoing debate around platform liability and content moderation would greatly benefit from the integration of concrete case studies that illustrate how current regulatory models have succeeded—or failed—in addressing real-world challenges. For instance, the enforcement of Germany's NetzDG law has led to the swift removal of thousands of posts deemed illegal, demonstrating regulatory efficiency but also sparking criticism over over-removal and the stifling of legitimate expression. Similarly, the implementation of the Digital Services Act has already

prompted major platforms to enhance their transparency infrastructure and risk assessment protocols, yet questions remain regarding the efficacy and independence of these mechanisms.

In contrast, the United States' reliance on Section 230 of the Communications Decency Act has preserved a robust space for free expression but has also allowed platforms to avoid accountability in cases of widespread misinformation or coordinated abuse. One illustrative example is the controversy surrounding the dissemination of conspiracy theories during the COVID-19 pandemic, where platforms were slow to respond due to ambiguity around responsibility and fear of political backlash. These cases highlight the practical implications of different liability regimes and reinforce the importance of developing evidence-based, adaptive regulatory models. By anchoring theoretical analysis in empirical outcomes, such case studies help clarify which mechanisms effectively balance freedom of expression with harm reduction and democratic accountability.

Despite recent advancements in platform regulation, current legislation still falls short in several key areas when confronted with the evolving dynamics of technology and user behavior. First, many regulatory frameworks are not sufficiently equipped to address the opaque functioning of algorithmic recommendation systems, which can amplify polarizing or harmful content without transparent criteria or accountability. While some regulations require impact assessments and audits, these measures often lack the granularity or enforcement power needed to influence algorithmic design and governance meaningfully.

Second, legislation tends to focus on reactive content moderation rather than proactive structural interventions. Laws typically address the removal of content *after* it has caused harm, overlooking the importance of designing platforms that *prevent* the virality of disinformation or abuse in the first place. Additionally, current models rarely consider the behavioral economics underlying user engagement—such as dopamine-driven interface designs that reward outrage and extremism.

Third, the global nature of platforms creates a regulatory mismatch: content circulated transnationally can escape the reach of national legal orders, especially in jurisdictions with limited enforcement capabilities. As a result, harmful actors may exploit these gaps to spread illicit or harmful content with little risk of accountability. These shortcomings underscore the urgency of developing adaptive legal tools that account for technological complexity, cross-border dynamics, and the psychological architecture of user engagement in the digital sphere.

8. Conclusion

The issue of social network liability deserves special attention and is far from reaching a settled understanding. To determine the best liability model, it is essential to understand the intricacies of how social networks operate, in order to identify which activities can be considered lawful and which violate individual and collective rights, as well as to define the limits of the moderator's role and the

actions of the platforms themselves, and to assess the risks inherent to the mechanisms that may be adopted.

It is necessary for the debate to mature further, thoroughly examining the positive and negative aspects associated with any chosen approach. The priority is for the discussion to contribute to refining the models, allowing for a choice that best meets societal expectations while still ensuring the broadest possible exercise of the fundamental right to freedom of expression without neglecting the special protection against extremist, hate, or violent speech. For this to be achieved, the debate must remain free from the influence of emotional and polarized narratives that often dominate the public discourse on this matter.

Whatever path a society chooses through legislative decisions, it is essential that various social sectors contribute their perspectives and insights on the issue to enable the broadest possible understanding of its nuances. The knowledge and experience of other communities aid in the comprehension and refinement of ideas. This ensures that the regulation of civil liability for social networks takes place within an ethical framework capable of meeting societal expectations regarding both freedom of expression and effective accountability.

As social media continues to evolve over the next decade, several emerging regulatory challenges are likely to test the limits of existing legal frameworks. Among the most pressing is the governance of immersive digital environments, such as the metaverse and extended reality (XR) platforms, where traditional concepts of jurisdiction, identity, and harm become increasingly blurred. These spaces will likely involve complex interactions between avatars, AI agents, and decentralized data structures, complicating the attribution of legal responsibility and the enforcement of user protections.

Another anticipated challenge involves the integration of generative artificial intelligence into social platforms, which may flood digital ecosystems with synthetic content—blurring the line between authentic expression and algorithmically generated speech. The legal and ethical implications of such content, particularly when it comes to defamation, political manipulation, or deepfake technology, remain largely unaddressed.

Lawmakers can begin preparing for these shifts by investing in anticipatory regulation: fostering interdisciplinary research hubs, promoting regulatory sandboxes, and engaging in structured dialogues with technology companies, civil society, and international partners. Flexible legal instruments—such as principles-based legislation and adaptive co-regulation frameworks—will be essential to ensure that future governance mechanisms are both technologically informed and rights-respecting. Proactive engagement today will help build resilient legal infrastructures capable of withstanding the pace and unpredictability of digital transformation.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Access Now et al. (2018). *Santa Clara Principles*. https://santaclaraprinciples.org/pt/

Bowers, J., & Zittrein, J. (2020). Answering Impossible Questions: Content Governance in an Age of Disinformation. *Harvard Kennedy School Misinformation Review, 1,* 1-8. https://misinforeview.hks.harvard.edu/wp-content/uploads/2020/01/bowers_content_governance_disinformation_20200114.pdf

Brasil (2014). Lei nº 12.965, de 23 de abril de 2014. Estabelece princípios, garantias, direitos e deveres para o uso da Internet no Brasil. [Marco Civil da Internet]. Diário Oficial da União: Seção 1, Brasília, DF. https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm

Brasil (2018). Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Diário Oficial da União: Seção 1, Brasília, DF. https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

Brasil (2022). Superior Tribunal de Justiça (STJ). Recurso Especial n. 1.783.269/MG. Relator Ministro Antonio Carlos Ferreira, Quarta Turma, julgado em 14/12/2021, DJe de 18/2/2022. https://scon.stj.jus.br/SCON/GetInteiroTeorDoAcordao?num_registro=201702627555&dt_publicacao=18/02/2022

Capanema, W. A. (2024). *Manual de Direito Digital: Teoria e Prática*. JusPodivm.

Dutra, L. F. (2024). Capítulo 25. Fake News e Regulamentação das Redes Sociais. In H. V. N. Jorge (Ed.), *Tratado de Direito Digital* (pp. 597-606). Editora JusPodivm.

Estarque, M., Archegas, J. V., Bottino, C., & Perrone, C. (2024). *Redes Sociais e Moderação de Conteúdo: Criando Regras para o Debate Público a Partir da Esfera Privada. Instituto de Tecnologia e Sociedade no Rio*. https://itsrio.org/wp-content/uploads/2021/04/Relatorio_RedesSociaisModeracaoDeConteudo.pdf

Intervozes—Coletivo Brasil de Comunicação Social, Observacom—Observatorio Latinoamericano de de Regulación, Medios y Convergencia, Desarrollo Digital, IDEC—Instituto Brasileiro de Defesa do Consumidor (2021). *Contribuições para uma regulação democrática das grandes plataformas que garanta a liberdade de expressão na internet: Uma perspectiva latino-americana para alcançar process's de moderação de conteúdo compatíveis com os padrões internacionais de direitos humanos*. https://www.observacom.org/wp-content/uploads/2019/08/Contribuic%CC%A7o%CC%83es-para-uma-regulac%CC%A7a%CC%83o-democra%CC%81tica-das-grandes-plataformas-que-garanta-a-liberdade-de-expressa%CC%83o-na-internet.pdf

Neto, L. G. d. S. (2024). Capítulo 2. Marco Civil da Internet no Brasil. In H. V. N. Jorge (Ed.), *Tratado de Direito Digital* (pp. 53-70). Editora JusPodivm.

Rodrigues, G. (2020). *Liberdade de Expressão, Moderação de Conteúdo e o PL das Fake News*. https://irisbh.com.br/liberdade-de-expressao-moderacao-de-conteudo-e-o-pl-das-fake-news

Teffé, C. S. d., & Souza, C. A. (2024). Moderação de Conteúdo e Responsabilidade Civil em Plataformas Digitais: Um Olhar Atual. In J. B. d. M. Menezes, & F. N. Barbosa, (Eds.), *A Prioridade da Pessoa Humana do Direito Civil-Constitucional: Estudos em Homenagem a Maria Celina Bodin de Moraes* (pp. 32-41). Editora Foco.

Teubner, G. (2011). Self-Constitutionalizing TNCs? On the Linkage of "Private" and "Public" Corporate Codes of Conduct. *Indiana Journal of Global Legal Studies, 18,* 617-638. https://doi.org/10.2979/indjglolegstu.18.2.617