CELE
Centro de Estudios en Libertad de Expresión

Artículo de investigación

68

Trust and Safety's Blindspots: A Latin American Perspective

Agustina Del Campo y Ramiro Álvarez Ugarte

Julio de 2025

A. Del Campo y R. Alvarez-Ugarte. (2025). Trust and Safety's Blindspots: A Latin American Perspective. In Maia Levy Daniel, Amanda Menking, Marlyn Thomas Savio, & Jean Claffey, Trust, Safety, and the Internet We Share: Multistakeholder Insights, Abingdon, United Kingdom: Taylor and Francis (forthcoming, late 2025)

Facultad de Derecho

Centro de Estudios en Libertad de Expresión

UP
Universidad de Palermo

Trust and Safety's Blindspots: A Latin American Perspective

Agustina Del Campo
Director
adelcampoiso@gmail.com

Ramiro Álvarez Ugarte
Deputy director
ramiroau@gmail.com

July 2025

Abstract

Trust and safety is slowly but steadily emerging in the United States as an independent field of study and practice. Several initiatives—such as the recent Glossary of Trust and Safety Terms issued by the Digital Trust and Safety Partnership (DTSP)—are attempting to lay the groundwork for a common language among researchers, corporate officials, and practitioners who routinely take part in a set of practices designed to manage the way some internet services are used. A set of normative goals are embedded in the field itself. This paper critically analyzes these processes so far to pose a series of questions regarding the boundaries of trust and safety as a practice and a field of inquiry, particularly in Latin America. Our main argument is that trust and safety favors safety over competing normative goals and principles, potentially altering and shaping human rights online. This trend is problematic from a human rights perspective, a point that we develop by focusing on Latin American human rights standards. We conclude with a proposal to reshape trust and safety in ways that embrace human rights standards as a countervailing commitment that, while difficult to make from a business-centered perspective, may increase the legitimacy of the practice overall.

Keywords trust and safety, internet regulation, self-regulation, internet governance

Introduction

Trust and safety is a field of practice that is slowly striving towards disciplinary recognition, particularly in the United States, through the creation of journals, associations, and a common language that describes a social practice sufficiently unique to deserve a specially hand-crafted disciplinary gaze.[1] It is a normative enterprise: it seeks to achieve goals that are inherently value-driven. But it is crossed by multiple contextual conditions (e.g., for profit corporations, catch-all services, an enabling legal environment) that make the field very unstable.

In this paper we would like to discuss the tension between the field of trust and safety as we currently see it and international human rights law, a corpus of international norms and principles that claims to have a role to play in internet governance and—in particular—content curation. We follow a genealogical approach that seeks to understand where things are and where they come from—in this case, trust and safety as a field of practice and knowledge. This chapter is "a practice of critique in the form of the historical problematization of the present" (Koopman, 2013, p. 2). But it also has—or so we hope—a potential dimension that allows us to imagine how things could have been different or could be different in a yet unwritten future (Lorenzini, 2024, p. 11).

In the first part of this chapter, we discuss trust and safety's origins. We argue that the field came to be in close connection to the law, at first encouraged and now—in Europe, through the DSA—required. The second section shows how the field clashes with standard understandings of international human rights law, especially freedom of expression. We show this highlighting specific standards in the Inter-American system of human rights that would be in tension with decisions and approaches that trust and safety—as we see it—would encourage. The third section offers a way out of this tension by expanding trust and safety goals to include more robust commitments to human rights standards. While this would mean difficult trade-offs, we believe—nevertheless—that it is an enterprise that may help the field build and or

---

[1] *See* e.g. the Digital Trust and Safety Partnership (https://dtspartnership.org/), the TrustCon conference (https://www.trustcon.nety), the *Journal of Trust and Safety*, the *Trust and Safety Professional Association* (https://www.tspa.org/), and so on. *See* also Zuckerman & Rajendra-Nicolucci (2023), 2: "Recent efforts to recognize trust and safety as a profession, with the establishment of the Trust & Safety Professional Association in 2020 and the emergence of a Journal of Online Trust and Safety in 2021 are overdue, as the work of policing online spaces traces back at least to the 1980s, if not earlier."

restore some of its lost legitimacy, amidst political backlash and sharp corporate policy turns.

Trust and Safety's Origin Story

The Digital Trust and Safety Partnership has recently defined trust and safety as the "field and practices employed by digital services to manage content and conduct-related risks to users and others, mitigate online or other forms of technology-facilitated abuse, advocate for user rights, and protect brand safety" (Glossary of Trust & Safety Terms - Digital Trust & Safety Partnership, 2023). It encompasses "a variety of cross-disciplinary elements including defining policies, content moderation, rules enforcement and appeals, incident investigations, law enforcement responses, community management, and product support," that has "developed into a distinct profession in its own right, with several professional organizations (such as DTSP and the Trust & Safety Professional Association) focusing on Trust & Safety functions emerging since 2020" (Glossary of Trust & Safety Terms - Digital Trust & Safety Partnership, 2023; Zuckerman & Rajendra-Nicolucci, 2023). These three elements are important and seem to define a set of connected social practices: goals, things to be done to achieve them, and a burgeoning cadre of professionals in charge of carrying them forward.

It is a charged definition, laid out in front of a background that anyone familiar with the history of internet regulation will plainly see. But user generated content (UGC) is the central piece of the puzzle. It is a way of naming what the internet became famous for: the lowering of the barriers to enter the public conversation that is so relevant for the functioning of democratic societies (Barlow, 1996). As a decentralized network, the internet created millions of publishers who could post their opinions online, share information, explore niche hobbies, or obsessively discuss the minutiae of Grateful Dead songs (Curran, 2012; Lessig, 2006). It was a dream that has been widely fulfilled. Although we have collectively left the late 1990s techno-optimism behind (Morozov, 2011), the democratization of the conversation did happen. That is the double curse and blessing of technology, and thirty years later we remain trapped in this contradiction.

Trust and safety can then be thought of as a tool to reap the benefits of horizontal conversations while dealing with their widely known and undesirable side-effects—harassment and hate speech "have become facts of life for users of many online systems..." (Zuckerman & Rajendra-Nicolucci, 2023, p. 1). Trust and safety is quintessentially a corporate practice, developed within the tech industry to keep the

horizontal nature of the conversation in check, with the double purpose of maximizing the value of the generally unpaid UGC while keeping the service within a certain threshold of civility—at least to the extent that incivility may bring the overall value of the service down and hurt profits (Klonick, 2018, p. 1627). Looking at the practice under its best possible light, trust and safety seeks to keep services usable and prevent the process of decline that has been seen in the industry so many times (Doctorow, n.d.). This is why popular internet services that are not for profit also consider trust and safety to be an essential part of the way they manage what happens on the platforms under their stewardship (Wikimedia Foundation, 2025).

Trust and safety is an inherently normative enterprise. It encompasses goals, practices, procedures, and professionals who strive to promote certain conduct and avoid others. Amidst broader disagreements regarding how the internet should be governed, we believe that trust and safety clearly picks sides on the ongoing debate, and leans towards safety and against freedom—a dichotomy we pose as a way to simplify the argument. To understand why, it is important to lay out the environmental conditions that have shaped the field of trust and safety, the role the law played in this development, and the structure of corporate incentives that explain its rise.

Trust and safety's normative commitments derive in part from the way it was shaped by the law: two distinct fields, both imbued with different forms of normativity, that are however interconnected. This odd relationship went through three historical stages, that—overall—explain why trust and safety's goals are what they are.

In the first stage, nation states created a legal environment that produced the conditions for the development of trust and safety as an autonomous field to govern online content in ways that exceeded, by far, the limited possibilities of the legal field in this area. The adoption of Section 230 of the Communication Decency Act of 1996 in the United States marks its beginning. The law famously granted internet companies immunity for their content moderation practices, so they could avoid liability for UGC. We call it an enabling legal environment because the law encouraged corporations to develop their own practices, under the expectation that they would collaborate with public officials to fight clearly illegal forms of expression or acts (Klonick, 2018, pt. I; Kosseff, 2019). This law was later adopted by the European Union (Directive 2000/31/CE, 2000) and was interpreted by courts as a legal and constitutional principle in many jurisdictions, under freedom of expression arguments and rationales (Álvarez Ugarte & Vitaliani, 2022).

In a second stage, corporations developed their own systems, practices, normative principles, and organizational skills and structures to deal with UGC (Zuckerman & Rajendra-Nicolucci, 2023). These were developed against the corporations' own criteria as to what was to be deemed acceptable in their platforms and what not, through terms of services or community guidelines that defined the scope of the normative world of these services (Cover, 1983). These rules grew in complexity and reach and responded—quite expectedly—to corporations' different interests, business models, and successful campaigns of advocacy towards them. As Kate Klonick puts it, even though they enjoyed the immunity granted by Section 230 they nevertheless moderated and curated UGC because of "an underlying belief in free speech norms;... a sense of corporate responsibility; and... the necessity of meeting users' norms for economic viability" (Klonick, 2018, p. 1618). The law played an indirect but essential function, enabling trust and safety to emerge without expressly demanding it. But the authority of the law does not depend exclusively on the threat of punishment (Raz, 2009). Hence, law-abiding corporations internalized many legal requirements in their own rules (terms of services, community guidelines, etc). The law and the field of trust and safety became, as a consequence, inherently interwoven.

Corporations, however, prohibited much more speech than required by law, a move that made sense from a business perspective. First, an over-reaching approach to the speech that was to be deemed unacceptable would reduce the risk of legal liability, especially in jurisdictions where the immunity granted to them was less than absolute (Zuckerman & Rajendra-Nicolucci, 2023, p. 6). It would also facilitate the development of global enforcement structures to moderate content at scale. (This would explain, for example, the case for rules prohibiting the exhibition of female nipples on social media.) Second, their ad driven business models push most corporations to maximize engagement, casting the biggest possible net to catch the largest amount of potential users. Therefore platforms seeking to capture a large user base would develop services designed to attract the mainstream demographic (Klonick, 2018, pp. 1627–1630). The famous campaign "stop hate for profit" may be cited among the most salient examples of the relationship between advertisers and UGC dependent internet services. [2] This naturally leaves fringe opinions and ideas vulnerable to strong forms of moderation, especially if they are perceived as toxic, in violation of community guidelines and terms of services, and potentially alienating of advertisers concerned with brand association and reputational harm. What constitutes

[2] See e.g. https://www.stophateforprofit.org/.

fringe is in itself controversial, likely defined by shifting ideological currents or corporate political alignments, as well as outside lobby and government pressure (Hendrix, 2025; Zuckerberg, 2024). We thus disagree with Klonick's assessment of the free speech values prevalent in corporations. If they exist, they come—at best—as second order concerns.

During the third and last stage, nation states took trust and safety as the basis for developing regulations that imposed strict procedural demands and duties upon corporations, obligations of transparency and self-assessment, and heavy penalties for failing to comply. The Digital Services Act (DSA) of the European Union marks—in our opinion—the beginning of a new era in platform regulation (Digital Services Act, 2022). The DSA, as well as other statutes that followed this model, encroached trust and safety practices into the law (Keller, 2024). The DSA asks corporations to monitor and assess the risks involved in content defined through broad, over-inclusive, and ambiguous categories. This approach widens the scope of legally mandated action over speech beyond the threshold established by international human rights standards (Del Campo et al., 2025). The recently published first batch of company audited reports and risk assessments show that most companies are reporting their trust and safety programs and practices as legally mandated mitigations for the risks identified by the law. Similarly, the adoption of similar laws in other countries like e.g. the UK Online Safety Act suggest that trust and safety has now become something that companies must do (Online Safety Act 2023, 2023).

When seen together, the three stages of trust and safety development show a field that has been shaped by the law—it emerged independently from but in relation to it. The leeway extended to platforms allowed them to set up their own rules, but in doing so they internalized norms established by states and, often, extended their reach. Their incentives to form a large user base left users without adequate protections (Palumbo, 2024). This is why we pose that trust and safety as a field strives towards safety: it is part of a governance mechanism imbued with expectations as to what companies should do, of business incentives that cut against fringe viewpoints, and it has been embraced as a necessary part of new regulations. This commitment to safety is, then, a commitment towards the attractiveness of platforms for advertisers, towards pleasant experiences for mainstream users, towards services that entice the attention that drives profits. It compromises, however, the value of freedom implied in any system of communication, a point we make in the next section.

Compatibility of Trust and Safety with Human Rights Standards

Trust and safety's goals and normative leanings discussed in the previous subsection clash with the demands made on corporations by international human rights standards. We explore this tension by focusing on business and human rights mandates globally and the Inter-American system of human rights specifically. For that purpose, we first discuss the field of Business and Human Rights and the dominant voluntary framework developed by the United Nations through the UN Guiding Principles on Business and Human Rights (Ruggie, 2008). We then move on to explore different examples of how trust and safety may clash with human rights standards.

International human rights law affirms that corporations have a responsibility to respect human rights (Ruggie, 2008), but falls short of actually affirming that this duty is similar to an obligation under international law (Álvarez-Ugarte & Krauer, 2020; Castañeda & Álvarez-Ugarte, 2020). This was on purpose: the UN Guiding Principles established an ambiguous framework as a way of dealing with the disagreement within the UN regarding the question of what human rights obligations transnational corporations should have (Álvarez-Ugarte & Krauer, 2020). We have criticized this arrangement elsewhere (Álvarez Ugarte, 2024), but—for now—suffice it to say that under the ambiguity of the framework, corporations cannot be held liable but, at the same time, are not immune from human rights-based criticism. Under increasingly evolving standards, corporations are seen as responsible for developing practices and procedures to assess their impact on human rights and mitigate harms when possible (Ruggie, 2007). In Latin America, the nature of corporate responsibility for human rights abuses is rapidly expanding (CIDH, 2019).

The field of business and human rights, weak as it may be, offers a platform based on which trust and safety operations may be assessed. Our claim in this section is that trust and safety's tendency to err on the side of safety goes against the broad responsibility to respect human rights that corporations have. The expansion of trust and safety rules impact rights as varied as political and economic, social and cultural rights. Terms of service and content curation practices affect the potential of any given service to serve as a means and a platform for the exercise and enjoyment of human rights, including freedom of expression, association, religion, education, cultural expression, among many others. We make the point by bringing in a few examples of content moderation practices measured against the Inter-American standards of human rights—the body of law developed by the Inter-American Court (IACtHR) and Commission (IACHR) of Human Rights, both bodies in charge of applying the

1969 American Convention on Human Rights (CAHR, 1969). But we believe the argument could easily be generalized in relation to other human rights and constitutional standards. If—according to the UNGPs—corporations have a responsibility to respect and protect human rights when doing their business; in Latin America, these rights are clearly defined by the way the Inter-American system interprets them.

For instance, the Inter-American system expressly prohibits prior censorship. The prohibition is established in article 13 of the American Convention and developed further in the Inter-American Court's Advisory Opinion 5 of 1985 (La colegiación obligatoria de periodistas, 1985). Advocates for business accountability in the region have posited that companies' efforts within trust and safety may amount to private censorship against the prohibition established in article 13 (Observacom, 2020). The language within Advisory Opinion No. 5 partially supports this claim, but the issue remains unresolved (Bertoni, 2017). The Special Rapporteur's office has—nevertheless—consistently held that State-led content filtering and take down need to be exceptional and subject to judicial control.

Inter-American standards demand—as other human rights standards, in other regions—that restrictions to freedom of expression be established using clear and precise language, that strives away from vagueness and abstraction (CIDH, 2009, par. 72). Criminalized conduct should have a "clear definition" (Caso Usón Ramírez c. Venezuela, 2009, par. 55). The terms of services or community guidelines of most internet companies fail to meet these standards: they change often, rely on vague language, and the value of predictability is—as a consequence—never realized. Users are often left with no idea why their content was taken down. If a Latin American speaker could develop a cause of action against a private corporation that failed to regulate its own speech following some degree of precision, the Inter-American standards—as they stand today—would support this claim (CIDH, 2019). Even more so if trust and safety practices are mandated by law, as in the case of the DSA.

When a corporation decides on a restrictive policy on political speech it may be infringing upon the special protections granted to political speech under article 13 of the American Convention and the constant jurisprudence of the Inter-American Commission and Court (Kimel v. Argentina, 2008, Caso Vélez Restrepo v. Colombia, 2012, Caso Álvarez Ramos v. Venezuela, 2019). While private companies are not necessarily public forums that must respect the rules of the public square in terms of viewpoint neutrality (González Mama, 2024; Klonick, 2018, p. 1611), freedom of expression is affected nevertheless. And in Latin America, many countries have

incorporated human rights in their legal frameworks and accept the principle of horizontality of constitutional rights (Gardbaum, 2003). The combination of both moves bring human rights to bear on the conduct of corporations, not through the indirect means of the UN Guiding Principles but through explicit legal decisions made by some Latin American countries.

The region is also highly restrictive of the grounds under which what is often called "hate speech" can be legitimately restricted. The prevalent interpretation advanced by the Inter-American Commission follows the standards accepted by the United States Supreme Court: inflammatory speech can only be restricted if it is linked to violence or to calls for "imminent lawless action" (CIDH, 2009, par. 58; Brandenburg v. Ohio, 1969, p. 447). This is not the case in Europe, where the grounds for restricting discriminatory speech are more generous (Jersild v. Denmark, 1994, Vejdeland and Others v. Sweden, 2012). Generally, social media platforms follow criteria that are closer to European standards, a source of potential conflict in Latin America for trust and safety.

Counter-arguments could be made against these specific clashing points between trust and safety's reasonable outcomes and Inter-American human rights standards. But the UN Special Rapporteur on Freedom of Opinion and Expression has raised the issue of the importance of corporations' role in governing speech, and the perils involved in failing to do so (Kaye, 2019, par. 41). The human rights responsibility for several online harms has been recalled over and over, and human rights standards many times do not offer clear cut answers to questions that are inherently difficult (Khan, 2021, par. 63). Answering these questions requires a level of nuance and commitment to human rights as law likely not available to corporate human rights officials (trapped in the voluntary framework) and certainly out of the scope of trust and safety practitioners. Our point is narrower: human rights claims against corporations are growing and even though the path towards liability and remedies is far from clear, we believe we have seen judges in Europe holding corporations accountable for human rights violations and we believe we are close to seeing judges in Latin America following those footsteps with an even broader mandate to uphold freedom of speech and expression. The standards developed by the OAS Special Rapporteur on Freedom of Expression have consistently called attention to corporate duties regarding human rights (CIDH, 2013, 2017, 2024). The question is not so much whether these decisions will emerge in the future, but when they will emerge.

Advocates around the world have been complaining about infringements to freedom of expression through private enforcement of trust and safety norms for a long time.

They have also questioned the permeability of companies to both State and non-state pressures to restrict unwanted or disliked legal content under the guise of trust and safety. And although companies originally enacted few actions around UGC, since 2016—but most importantly since the 2020 COVID pandemic—companies have been more active than ever in developing ever more complex rules and processes for content to be admitted. We are only now seeing some of the backlash against this trend amidst global political turmoil. Trust and safety may be entering —at the time of writing— a profound crisis, partially caused by sharp corporate policy turns and new political realignments, that caused some of the biggest players to repudiate past behavior, commitments, and remedial action (Hendrix, 2025; Zuckerberg, 2024). In the United States, pressure to resist the mandates of the DSA is also growing (Carr, 2025). All of which leads us to our final point about human rights as a possible way out of the conundrum trust and safety is currently in.

Expanding Trust and Safety's Frontier

In the previous sections we have shown why—from our perspective—trust and safety picks sides in the ongoing debate on the scope and reach of content moderation. The business incentives, the expectations imbued in the legal framework, and the incorporation of the practice in new regulatory regimes predict over-censorial approaches that are very problematic from freedom of expression standards, among other human rights that depend on speech or expression for their exercise and enjoyment. Even though human rights-based claims can be made regarding the need for more censorship instead of less, the clash with standard freedom of expression standards exists nevertheless.

We believe, however, that there is a way out of this conundrum: trust and safety could explicitly embrace freedom and pluralism as a necessary part of the operation of corporations that provide communication services and facilitate UGC, especially those that strive for maximum engagement and a large user-base—the companies that have shaped, until now, trust and safety as a field. Incorporating a richer approach to this fundamental, often overlooked dimension of the internet's horizontality affordance may help trust and safety achieve greater legitimacy.

A first step in that direction could be acknowledging that the incentives at play currently tilt the field towards different forms of safety that compromises—at least on some occasions, under certain conditions—the values of freedom and pluralism. The conception of "safety" embraced by the field far exceeds the legitimate objectives contemplated to restrict freedom of expression, association and assembly than human

rights treaties do. And although this may be justified, acknowledging and substantiating the difference from a theoretical point of view and developing grounds from a human rights based perspective would provide further strength to the currently unstable and heavily criticized unilateral practice of trust and safety. A more robust commitment to users rights should be instrumental for that purpose.[3]

Trust and safety's cross-disciplinary commitments can provide some leverage in this endeavor. A more robust and structured engagement with human rights policy teams and those in charge of conducting human rights impact assessments could help bridge the gaps and develop a more granular and standard-based conversation. This move means leaning on trust and safety's cross disciplinary pledge, and inviting others' expertise to take a leading role in the development of the discipline. Meta's Oversight Board may be an example of a body that attempts to bridge these kinds of gaps, even if not without its flaws or limits as to their enforceability.

The value of freedom should not be seen as necessarily going against the safety that companies desire, but it would likely mean adopting a more nuanced, less expansive concept of safety. This would mean adopting a concept of freedom that aligns with human rights and a concept of safety that is guided by and contained by human rights. A more nuanced, inter-disciplinary conversation within the trust and safety community may produce different points of equilibrium, with some services maximizing one value over the other, and others still striving to strike a balance. Furthering this conversation within the trust and safety community may strengthen

---

[3] This is not easy to do. Efforts by companies to self-regulate according to human rights standards and for state-led regulation to promote user rights have faced significant difficulties dealing with the differences between the human rights regime, developed specifically for nation states, and companies' duties. This tension is not to be resolved anytime soon, but it must be reckoned with. Brenda Dvoskin, among others, has argued that human rights obligations require certain tweaking to be directly applicable to companies rather than states. *See* Dvoskin (2022). The legitimate objectives that state led restrictions must satisfy are expressed in articles 19 of the ICCPR, article 13 of the American Convention and Article 10 of the European Convention. And as good as those taxative restrictions have worked to guide States and assess State conduct, they do little to assess company conduct, which is guided by an entirely different rationale. Companies develop products, which target specific audiences, to provide specific services. Each company tailors their products to provide specific services and in doing so, they necessarily apply restrictions on speech, directly through code. *See* Lessig (2006). Examples may include allowing only videos on the platform rather than photographs, or allowing only 140 characters and not more. They may also want to provide services at retail—tailored for children, or for adults that share an activity or passion (pets, community organizers, gardening, etc)—or at wholesale, seeking to capture the largest possible audience.

the field in the face of potential capture, especially if the DSA becomes a model and trust and safety operations are absorbed by legal compliance (Keller, 2024).

In this enterprise, trade-offs will have to be made. Our position is towards expanding freedom, which means adopting a less censorial approach to content curation that is more respectful of the plurality of ideas and opinions that exist in a democratic society, even those that "offend, shock or disturb" (Handyside v. the United Kingdom, 1976, par. 49). It coincides with very prominent corporate positions, even though we are skeptical of this sudden shift in corporate policies—especially of its drivers and motives (Hendrix, 2025). If the field of trust and safety is to survive and grow, increased harms will likely occur—it is a necessary risk (no pun intended) of living in a democratic society. People will be offended; some of them may feel disturbed to the point of self-censorship. There are ways of dealing with these undesirable side-effects of the horizontal conversation afforded by the Internet. Better self-defense tools (such as blocking and silencing particularly obnoxious users) and the hard-work of recovering the value of civility in online conversations (not entirely lost in in person conversations) are some of them. And traditional legal defenses against abuse and harassment should provide coverage for strong corporate punishment for those sorts of behaviors.

As outsiders looking in, making a critique from our expertise as human rights practitioners committed to the study of freedom of expression and to its important role in democratic societies, we believe that understanding the strengths and weaknesses of trust and safety practices as the field grows is crucial to its legitimacy. Any practical discussion of how to expand human rights perspectives into trust and safety practices requires —first and foremost—a shared diagnosis and a commitment to address the challenges raised from a law abiding, human rights centric perspective. The current state of crisis we see the field in is a good opportunity to reassess shared principles and goals.

Conclusion

We have considered the emergence of trust and safety as a discipline, something that captures in broad terms an ongoing conversation over the methods, processes, expertise, and corporate bureaucracies in charge of managing the undesired effects of content produced by users on the internet.

We have argued that the field is inherently tilted towards safety, a decision that can be explained by the incentives at play in corporate decision-making on matters of content

and a legal environment that gives these corporations a lot of leeway to decide how to do this, under fairly standard legal principles. We have identified different stages in the development of the field: the state enabled it through laws that encouraged the practice, corporations developed the practice actively collaborating with governments and other outside actors, and these—in the last step—have embraced the practice as a necessary ingredient of how speech is to be governed online. This evolution suggests a dynamic, strong relationship between trust and safety and the law, one that is controversial for the reasons previously discussed. When judged from the standpoint of the rules governing conduct, trust and safety's commitment to safety clearly makes the practice clash with speech rules that require much more freedom. We showed this through an analysis of the Inter-American legal standards on freedom of expression that adamantly affirms that corporations that intermediate on the free flow of information on the internet have specific human rights duties.

However, a stronger commitment to human rights could help trust and safety restore its somewhat lost legitimacy amidst profound regulatory change (Atlantic Council, 2023). This is a challenging enterprise, first and foremost because it may cut against the incentives of corporations bent on making a profit from the services they offer. As we have shown in the second section of this chapter, the business model is not friendly to the kind of fringe views and opinions that justify the existence of freedom of expression clauses in constitutional democracies. This contradiction is an extremely relevant side of our current predicaments, regarding internet governance and democratic erosion. Charting a way out of these muddy waters exceeds the scope of this chapter, but trust and safety could contribute to this enterprise by expanding its normative commitments in the directions we have suggested.

References

Álvarez Ugarte, R. (2024). *Bad Cover Versions of Law. Inescapable Challenges and Some Opportunities for Measuring Human Rights Impacts of Corporate Conduct in the ICT Sector* [Preprint, under review].

Álvarez Ugarte, R., & Vitaliani, E. (2022). Redes de influencia: análisis de la jurisprudencia civil argentina en materia de responsabilidad de intermediarios. *Revista Chilena de Derecho y Tecnología*, *11*(2), 147–182. https://doi.org/10.5354/0719-2584.2021.65368

Álvarez-Ugarte, R., & Krauer, L. (2020). *ICT and Human Rights: Towards a Conceptual Framework of Human Rights Impact Assessments* [Artículo de investigación]. CELE Research Paper Series. https://doi.org/10.2139/ssrn.5152205

Atlantic Council. (2023). *Scaling trust on the web* [Comprehensive Report of the Task Force for a Trust Worthy Future Web]. Atlantic Council.

Barlow, J. P. (1996). *Declaration of the Independence of Cybserspace*. EFF. https://www.eff.org/cyberspace-independence

Bertoni, E. (2017). OC- 5/85: su vigencia en la era digital. In I. Álvarez, E. Bertoni, C. Botero, & E. Lanza (Eds.), *Libertad de expresión: A 30 años de la Opinión Consultiva sobre la colegiación obligatoria de periodistas* (1a ed., pp. 33–46). Comisión Interamericana de Derechos Humanos.

Carr, B. (2025, February 26). *Brendan Carr Letter to internet Companies* [Letter]. https://www.fcc.gov/sites/default/files/Chairman-Letter-to-Big-Tech-on-Digital-Services-Act.pdf

Castañeda, A., & Álvarez-Ugarte, R. (2020). *Human Rights Impact Assessments: Trends, Challenges, And Opportunities for ICT Sector Adoption* [Artículo de investigación]. CELE Research Papers Series. https://doi.org/10.2139/ssrn.5152219

CIDH. (2009). *Marco jurídico interamericano del Derecho a la Libertad de Expresión* (OEA/Ser.L/V/II CIDH/RELE/INF. 2/09). Relatoría Especial para la Libertad de Expresión de la Comisión Interamericana de Derechos Humanos.

CIDH. (2013). *Libertad de expresión e internet*. Comisión Interamericana de Derechos Humanos.

CIDH. (2017). *Estándares para una internet libre, abierta e incluyente* (Informe Temático INF.17/17; OEA/Ser.L/V/II CIDH/RELE). Relatoría Especial para la Libertad de Expresión de la CIDH.

CIDH. (2019). *Empresas y derechos humanos: Estándares interamericanos* (OEA/Ser.L/V/II CIDH/REDESCA/INF.1/19; p. 211). Comisión Interamericana de Derechos Humanos.

CIDH. (2024). *Inclusión digital y gobernanza de contenidos en internet* (OEA/Ser.L/V/II CIDH/RELE/INF. 28/24). Relatoría Especial para la Libertad de Expresión.

Convención Americana de Derechos Humanos, Pub. L. No. 1144 U.N.T.S. 123, U.N.T.S. (1969).

La colegiación obligatoria de periodistas, 5/85 Serie A ___ (Corte Interamericana de Derechos Humanos 1985).

Kimel v. Argentina, Serie C ___ (Inter-American Court of Human Rights 2008).

Caso Usón Ramírez c. Venezuela, Serie C ___ (Inter-American Court of Human Rights 2009).

Caso Vélez Restrepo v. Colombia, Serie C ___ (Corte Interamericana de Derechos Humanos 2012).

Caso Álvarez Ramos v. Venezuela, 308 Serie C ___ (Corte Interamericana de Derechos Humanos 2019). http://www.corteidh.or.cr/docs/casos/articulos/seriec_380_esp.pdf

Cover, R. (1983). Foreword: Nomos and Narrative. *Harvard Law Review*, *97*, 4–68.

Curran, J. (2012). Rethinking internet history. In J. Curran, N. Fenton, & D. Freedman (Eds.), *Misunderstanding the internet* (1st ed., pp. 34–66). Routledge.

Del Campo, A., Zara, N., & Álvarez-Ugarte, R. (2025). *Are Risks the New Rights? The Perils of Risk-based Approaches to Speech Regulation* (63). Centro de Estudios en Libertad de Expresión (CELE). https://doi.org/10.2139/ssrn.5161173

Doctorow, C. (n.d.). The "Enshittification" of TikTok. *Wired*. Retrieved January 24, 2025, from https://www.wired.com/story/tiktok-platforms-cory-doctorow/

Dvoskin, B. (2022). *Expert Governance of Online Speech* (SSRN Scholarly Paper 4175035). https://papers.ssrn.com/abstract=4175035

Handyside v. the United Kingdom, 57499/17, 74536/17, 80215/17, 9323/18, 16128/18, 25920/18 (ECtHR December 7, 1976). https://hudoc.echr.coe.int/eng?i=001-57499

Jersild v. Denmark, 57891/17 (ECtHR [GC] September 23, 1994). https://hudoc.echr.coe.int/eng?i=001-57891

Vejdeland and Others v. Sweden, 1813/07 (ECtHR February 9, 2012). https://hudoc.echr.coe.int/eng?i=001-109046

Digital Services Act, Pub. L. No. 2022/2065, OJ L 277, 27.10.2022 1 (2022). https://eur-lex.europa.eu/eli/reg/2022/2065/oj

Directive 2000/31/CE, DOCE L 178/11 (2000). https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX%3A32000L0031

Gardbaum, S. (2003). The "Horizontal Effect" of Constitutional Rights. *Michigan Law Review*, *102*(3), 387–459. https://repository.law.umich.edu/mlr/vol102/iss3/2

Glossary of Trust & Safety Terms - Digital Trust & Safety Partnership. (2023, January 27). https://dtspartnership.org/glossary/

González Mama, M. (2024). *A propósito de la metáfora de las redes sociales como foro público en los Estados Unidos* (Artículo de investigación 60). Centro de Estudios en Libertad de Expresión.

Hendrix, J. (2025, January 7). *Transcript: Mark Zuckerberg Announces Major Changes to Meta's Content Moderation Policies and Operations | TechPolicy.Press*. Tech Policy Press. https://techpolicy.press/transcript-mark-zuckerberg-announces-major-changes-to-metas-content-moderation-policies-and-operations

Kaye, D. (2019). *Promotion and protection of the right to freedom of opinion and expression* (A/74/486). Office of the Special Rapporteur on Freedom of Opinion and Expression.

Keller, D. (2024). *The Rise of the Compliant Speech Platform*. The Lawfare Institute & Brookings Institution. https://www.lawfaremedia.org/article/the-rise-of-the-compliant-speech-platform

Khan, I. (2021). *Disinformation and freedom of opinion and expression* (A/HRC/47/25). Human Rights Council.

Klonick, K. (2018). The New Governors: The People, Rules, And Processes Governing Online Speech. *Harvard Law Review*, *131*, 73.

Koopman, C. (2013). *Genealogy as Critique: Foucault and the Problems of Modernity*. Indiana University Press.

Kosseff, J. (2019). *The twenty-six words that created the internet*. Cornell University Press.

Lessig, L. (2006). *Code* (Version 2.0). Basic Books.

Lorenzini, D. (2024). On possibilising genealogy. *Inquiry*, *67*(7), 2175–2196. https://doi.org/10.1080/0020174X.2020.1712227

Morozov, E. (2011). *The net delusion: the dark side of internet freedom* (1st ed). Public Affairs.

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

Palumbo, A. (2024). A Medley of Public and Private Power in DSA Content Moderation for Harmful but Legal Content: An Account of Transparency, Accountability and Redress Challenges. *JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law*, *15*(3), Article 3. https://www.jipitec.eu/jipitec/article/view/412

Online Safety Act 2023, 2023 c. 50 (2023). https://www.legislation.gov.uk/ukpga/2023/50

Raz, J. (2009). *The Authority of Law: Essays on Law and Morality* (2nd edition). Oxford University Press, USA.

Ruggie, J. (2007). *Human rights impact assessments --- resolving key methodological questions* (A/HRC/4/74). Human Rights Council. Report of the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises.

Ruggie, J. (2008). *Protect, Respect and Remedy: a Framework for Business and Human Rights* (A/HRC/8/5). Human Rights Council. eport of the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises.

Brandenburg v. Ohio, 395 U.S. 444 (Supreme Court of the United States 1969).

Wikimedia Foundation. (2025, January 24). *Meta: WMF Trust and Safety - Meta*. Wikimedia Meta-Wiki. https://meta.wikimedia.org/wiki/Meta:WMF_Trust_and_Safety

Zuckerberg, M. (2024, August 26). *Meta's letter to the Committee of the Judiciary* [Letter]. https://x.com/juanof9/status/1828245345635311990

Zuckerman, E., & Rajendra-Nicolucci, C. (2023). From Community Governance to Customer Service and Back Again: Re-Examining Pre-Web Models of Online Governance to Address Platforms' Crisis of Legitimacy. *Social Media + Society*, *9*(3), 20563051231196864. https://doi.org/10.1177/20563051231196864