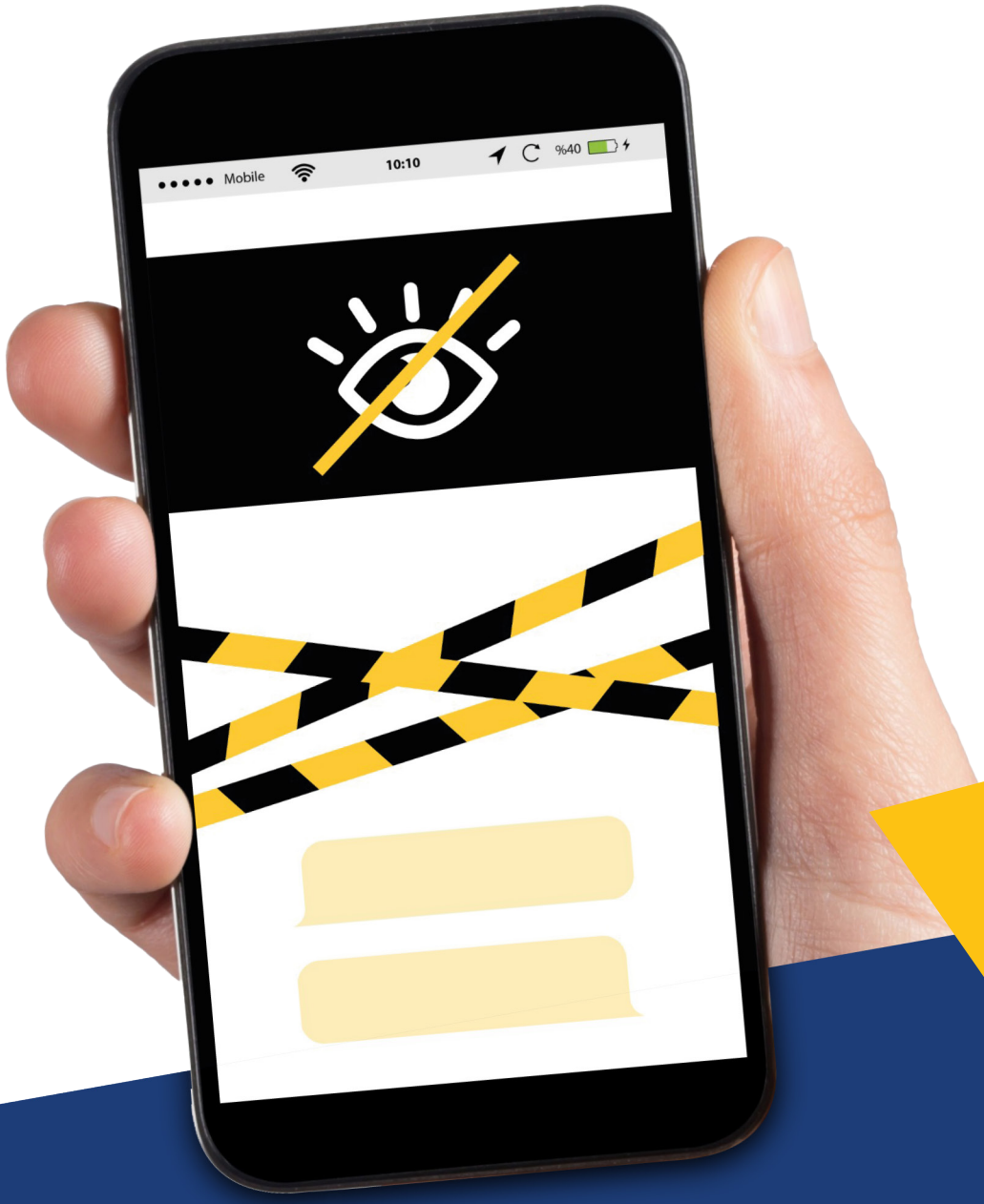

“SHADOW BANNING”:
The Subtle and Covert Censorship of the Major Tech Platforms

OBSERVACOM

“Shadow banning”: The Subtle and Covert Censorship of the Major Tech Platforms

A study on practices of reducing the reach of content and accounts on social media, and their impact on freedom of expression online

Published by
OBSERVACOM
Observatorio Latinoamericano de Regulación, Medios y Convergencia

Av. Libertador 1878 apto. 715
Montevideo, Uruguay
www.observacom.org

Supported by

FID | FORUM INFORMATION DEMOCRACY

and

Digital Action ><

Carolina Martínez Elebi - Author

She holds a degree in Communication Sciences from the Universidad de Buenos Aires (UBA) in Argentina, where she has been a professor since 2011. She is a consultant on issues related to the impact of information and communication technologies on human rights and is the director of the digital media outlet DHyTecno. In 2018, she completed the Internet and Communications Technology Law Program at the Centro de Estudios en Tecnología y Sociedad (CETyS-UdeSA)(CETyS-UdeSA). She is the academic coordinator of the Advanced Diploma in Artificial Intelligence and Society at the Universidad de Tres de Febrero (UNTREF) and is a member of the Observatory on the Social Impact of Artificial Intelligence at the same university.

Vladimir Cortés Roshdestvensky - Author

He is a human rights specialist with over a decade of experience in digital rights, freedom of expression, and democratic governance. He holds a Master's degree in Human Rights from the University of Padua and has led research on AI, the digital divide, and content moderation. He is currently Director of Campaigns and Partnerships at Digital Action, where he coordinates strategies to hold governments and technology companies accountable. He is a recipient of the LACNIC Leaders and Nicola Tonon Fellowship. He has worked at ARTICLE 19 and as an analyst for Freedom House's Freedom on the Net report.

01.

INTRODUCTION

In today's digital ecosystem, platforms play a central role as intermediaries in the circulation of information. As part of that role, they implement content moderation systems that include visible and relatively well-known measures. These can range from removing posts, temporarily or permanently suspending accounts, and other sanctions that users are generally informed about. These decisions are usually accompanied by users' access to appeals mechanisms —at least under the terms set by the companies themselves— and are framed as part of compliance with their community guidelines.

However, in recent years, these traditional forms of moderation have been complemented —and in some cases even replaced— by practices that have just as much an impact but are more subtle, less transparent, and much harder to detect. As it's no longer just about directly removing content, but about vague, behind-the-scenes interventions that affect the circulation of public-interest information and other forms of user-generated content. The Office of the Special Rapporteur for Freedom of Expression (SRFOE) of the Inter-American Commission on Human Rights (IACHR) has warned that tech companies must avoid allowing algorithms and automated systems —especially those that operate without meaningful human oversight— to become a threat to freedom of expression. The risk is especially serious when such decisions result in excessive and disproportionate restrictions, which tend to impact historically marginalized groups more frequently.[1]

Shadow banning refers to these kinds of tactics. While technically keeping a user's content available, this is a practice that drastically reduces its visibility, affecting media outlets, activists, entrepreneurs, and users who, in many cases, aren't even aware that access to their content is being limited. What's particularly concerning is that this visibility reduction acts as a form of silent censorship. Without notifications or clear explanations, diverse voices are effectively excluded from the digital public sphere, undermining informational pluralism and democratic debate.

While shadow banning doesn't completely block users' ability to express themselves, it has the potential to significantly affect four crucial dimensions of discourse: how available really is the content; its true visibility within the digital ecosystem; its accessibility to different audiences, and lastly, its ability to generate impact in public discussions.

This concern is especially relevant in light of the standard established by the Inter-American Court of Human Rights (IACHR), which has enshrined a fundamental principle: expression and its dissemination form an indivisible whole. This interpretation significantly broadens the protective scope of the right to freedom of expression, guaranteeing not only the ability to express ideas and opinions, but also the essential right to

1 Inter-American Commission on Human Rights (IACHR), Office of the Special Rapporteur for Freedom of Expression (2024). Digital inclusion and Internet content governance (OEA/Ser.L/V/II CIDH/ RELE/INF.28/24, p. 61, para. 276). Organization of American States: https://www.oas.org/en/iachr/expression/reports/Digital_inclusion_eng.pdf

use any appropriate channel to ensure that such ideas reach as many people as possible, and that potential recipients are effectively able to access such information. [2]

This alteration in moderation practices raises serious questions about transparency and accountability. When platforms algorithmically reduce the reach of certain content without notifying users, who exactly supervises these decisions? And what criteria are being used to implement them?

This study aims to unravel the impact of these practices on media outlets, journalists, activists, and individual users. It also explores how transparent platforms are about these mechanisms which, despite being concealed, profoundly restructure our information ecosystem and affect people's psycho-emotional well-being.

The purpose of this investigation is to examine the phenomenon of shadow banning that takes place on digital platforms, identifying specific cases and analyzing their impact on the visibility of media outlets, critical voices, and underrepresented sectors. Additionally, it seeks to assess how transparent platforms are regarding these practices, along with the consequences for democratic participation in online public spaces.

This investigation, with a focus on identifying shadow banning practices in Latin America, used a predominantly qualitative method through semi-structured interviews, carried out with a range of actors taking part in the digital ecosystem. It also involved reviewing the terms and conditions of major tech companies such as Meta (Instagram and Facebook) and X (formerly Twitter), along with a review of the pertinent literature. The document analysis of platform policies and terms of service allowed us to compare reported experiences with the platforms' stated rules, revealing significant discrepancies between perceived practices and official policies. In addition, the literature review included both academic research and reports from civil society organizations and think-tanks, allowing us to contextualize the phenomenon within the broader global debate on content moderation and freedom of expression in digital environments.

This method enabled us to document and validate firsthand experiences from activists, journalists, human rights defenders, and digital entrepreneurs who reported facing obscure visibility restrictions on various platforms. Through these testimonies, we were able to identify common patterns in reported experiences along with specific impacts on the exercise of free expression in digital spaces.

It's important to note that shadow banning research can encounter considerable methodological challenges, be they technical or in terms of data collection. At present, major digital platforms have imposed significant restrictions on research access, whether through complex data request schemes, limits on API usage, the shutdown of tools such as CrowdTangle (which previously supported independent research),

2 Idem

or a general decline in transparency around their algorithmic operations. These technical barriers severely hinder efforts to document shadow banning with robust quantitative evidence, which led us to prioritize qualitative documentation of representative cases.

It's important to note that the inherently secretive nature of shadow banning is perhaps the greatest methodological challenge. Unlike other forms of content moderation, where users receive explicit notifications, shadow banning is characterized precisely by the lack of transparency and communication with affected users. This characteristic, combined with the ambiguity in platform policy language —which rarely refers directly to these practices and often uses euphemisms like "reduced distribution" or "visibility adjustments"— creates a scenario in which systematic documentation becomes extraordinarily difficult. This is why our methodology focused on triangulating reported experiences with observable changes in reach and content visibility, while acknowledging the inherent limitations of researching practices that are deliberately designed to be imperceptible. It's also worth noting that academic and technical literature on shadow banning is primarily available only in English, making it a relatively unexplored topic in Spanish-speaking contexts.

02.

SHADOW BANNING: THE COVERT INVISIBILIZATION OF CONTENT AND USERS

In the digital space dominated by big tech platforms —and where millions of people debate, share, and access information— there operates a stealthy form of silencing that affects activists, journalists, and other users, though few manage to detect it in time. This is shadow banning: a set of moderation practices through which platforms quietly reduce the reach and visibility of certain profiles or posts without notifying the affected user. This is why it's often described as a form of "covert" moderation in a double sense: it renders content invisible to other users and hides the sanction from the person subjected to it. Like a ghost moving through the algorithms, shadow banning leaves its victims trapped in a communicative limbo: they keep speaking, but their presence fades away without a trace.

Unlike explicit restrictions —such as content removal or account suspension— shadow banning maintains an appearance of normalcy. Affected accounts can continue posting as usual, but their content is gradually excluded from public conversations: disappearing from search results, becoming invisible in trending hashtags, or no longer appearing in the feeds of their own followers.[3] In many ways, it is just a more sophisticated version of digital silencing.

The manifestations of shadow banning are as diverse as they are subtle: from a sudden drop in engagement to complete disappearance from recommendation systems or search results. A journalist who typically receives hundreds of comments may suddenly find themselves speaking into a void; an activist using hashtags related to human rights might discover their posts never appear in those searches; a sex educator may see their informative content filtered out by algorithms that flag it as "borderline content."

What's most concerning, however, is that these algorithmic restrictions are not applied equitably. Evidence shows they disproportionately affect marginalized communities: i.e., social and political activists, independent journalists, LGBTQIA+ individuals, and practitioners of comprehensive sex education. The core issue is concealment. Without notifications, explanations, or effective appeal mechanisms, affected users are forced to carry out invisible, exhausting work: from formulating hypotheses about how the algorithm works, to modifying their language (algospeak) to avoid being sanctioned, or building collective networks to verify and bypass covert censorship.

This phenomenon is certainly not a technical anomaly. Rather, it is a concrete threat to the fundamental rights of freedom of expression. This is especially serious in contexts like Latin America, where digital visibility can be critical to activism, public denunciation, and democratic participation. It is urgent, therefore, to push for stronger mechanisms of accountability and transparency in the decisions made by those who currently control the gateways to public information.

The most common forms of shadow banning include:

3 Nicholas, G. (2022). Shedding light on *shadow banning*. Center for Democracy & Technology. https://cdt.org/wp-content/uploads/2022/04/remediated-final-shadowbanning-final-050322-upd-ref.pdf

- Reduced reach of a user's posts: This means that the content someone publishes reaches a significantly smaller audience than usual, without being deleted. For example, a post that would typically receive hundreds or thousands of interactions (likes, comments, shares) gets far fewer, because the platform decides not to display it in other users' feeds or doesn't prioritize it enough for people to see it (placing it far down or at the bottom of the feed where no one scrolls to). This drop may be sudden or gradual and shows a disproportionate decline in likes, comments, and views relative to the user's follower count and typical engagement metrics. Some authors have documented that "posts from those who reported being [shadow banned] don't appear in their followers' feeds at all and are apparently deprioritized by the algorithm entirely." [4]

- Restricted visibility of a user in search results: This means that even if someone types the exact name of an account into the search bar, it doesn't appear in the results, isn't suggested,[5] or shows up very far down, making it difficult for others to find. This limits the organic growth of an account and its user's ability to engage in public debate.

- Removal of a user's account from suggestions and recommendations shown to other users: The platform stops including the account in sections such as "People you may know," "Suggested for you," "Recommended accounts," or similar. For example, an account is no longer suggested to users who might be interested in its content, reducing the chances of reaching new audiences.

- Exclusion from hashtags, discovery feeds, and trending topics: Even if a person tags their post with a specific hashtag, the post doesn't appear when others click on that hashtag. For instance, a user posting under #FreePalestine or #MeToo might notice their post isn't listed among the results for that hashtag, preventing their message from joining the public debate. This also refers to posts being excluded from algorithmic discovery pages like Instagram's "Explore" or TikTok's "For You" page.

- Limiting interaction with other users: The user's comments or replies are hidden or downgraded for others, even though they remain visible to the author. For example, a journalist comments on a viral post, but their comment is invisible to others, reducing their ability to participate in public conversations.

- Blocking features: The user becomes unable to use certain functions that allow interaction with other users. For example, they might be unable to like or reply to others' posts, or their post may not be linked to their profile name.[6]

4 Blunt, D., Wolf, A., Coombes, E., & Mullin, S. (2020). Posting into the void: Studying the impact of shadowbanning on sex workers and activists. Hacking//Hustling.
5 Le Merrer, E., Morgan, B., & Trédan, G. (2021). Setting the record straighter on shadow banning. In IEEE INFOCOM 2021-IEEE Conference on Computer Communications (pp. 1-10). IEEE. https://arxiv.org/pdf/2012.05101
6 Blunt, D., Wolf, A., Coombes, E., & Mullin, S. (2020). Posting into the void: Studying the impact of shadowbanning on sex workers and activists. Hacking//Hustling.

All these actions not only reduce the reach and circulation of a user's content but also diminish or block the "discoverability" of their posts, and also their account and presence on the social network, making it difficult or impossible for such a user to grow their audience or follower base.

The problem with this practice, employed by platforms to sanction supposed violations of their community guidelines, is that it restricts the circulation of users' content without the users themselves noticing, unlike an open ban, where an account is suspended or deleted directly and usually comes accompanied by a platform notification. Even if the account isn't blocked in strict terms, by limiting the reach of a user's content and their discoverability to new audiences, the impact is the same: it hinders —or outright excludes— their participation in online public debates.

Shadow banning is applied automatically through algorithms, which are linked to artificial intelligence that "moderate" the circulation of content, so as to control the discourse within a given platform. The lack of transparency of this practice, along with the absence of clear mechanisms to detect or appeal such decisions, make it one of the most controversial forms of corporate intervention in the new public spaces found on the Internet.

Consequently, these practices severely affect freedom of expression and informational pluralism, as media outlets, journalists, and activists may find their impact on public discourse diminished without adequate and timely mechanisms to assert their rights or reverse the measures taken against them.

To properly understand the phenomenon of shadow banning, it is essential to distinguish between two key concepts that operate on digital platforms such as X, Instagram, Facebook, etc. Firstly, content moderation refers to the set of policies, systems, and tools that platforms employ to manage user-generated content, determining what gets published, what is removed, or how it is controlled.[7] This process can be structured through three types of system: (1) "centralized" (as in Twitter/X, Facebook, or YouTube), where the platform enforces rules internally; (2) "distributed" (as in Reddit or Wikipedia), where the communities themselves manage moderation with minimal platform intervention; or (3) "hybrid" (as in Twitch), which combines both approaches. Moderation, in general, unfolds through sequential phases that range from rule-setting to appeals mechanisms and can be applied either before publication (ex ante) or after (ex post).[8]

To define the concept of content moderation on social networks, the Office of the Special Rapporteur for Freedom of Expression of the Inter-American Commission on Human Rights has adopted definitions from the Americas Dialogue process, as well as from documents produced by civil society organizations that specialize in this area.

7 Center for Democracy & Technology. (2021). Outside looking in: Approaches to content moderation in end-to-end encrypted systems. https://cdt.org/insights/outside-looking-in-approaches-to-content-moderation-in-end-to-end-encrypted-systems/

8 Klonick, K. (2018). *The new governors: The people, rules, and processes governing online speech*. St. John's University School of Law.

Thus, in paragraph 187 of the report Digital Inclusion and Internet Content Governance, published in June 2024,[9] "content moderation is defined as the organized practice of screening content generated and viewed by users and posted on digital platforms." The report outlines several types of content moderation: pre-moderation, post-moderation, reactive moderation, distributed moderation, and automated moderation.

The Rapporteur also emphasized in the report that "the moderation process may be carried out either by a person directly or through automated processes based on artificial intelligence tools together with the processing of large amounts of user data." Moderation may involve "taking down content permanently or temporarily, across the entire platform or in relation to certain groups of users in a specific geographic area, or affecting accounts of users under different modalities." Another type of moderation may include actions such as labeling content, providing additional and contextualized information about a post, or de-monetizing their posts, among others.

Content curation, on the other hand, is the process by which digital platforms select, organize, and present content to an audience according to criteria that users are unaware of. This process determines which content will gain greater visibility and which will be relegated in feeds, search results, and personalized user recommendations on the platform.

The Office of the Special Rapporteur for Freedom of Expression of the Inter-American Commission on Human Rights defines content curation as "automated decisions about the reach, ranking, promotion, or visibility of content. Platforms usually curate content based on personalized recommendations for user profiles.[10] To the extent that certain content is favored, curation can end up amplifying or reducing the reach of certain speech" the Rapporteur warns.

In this sense, content curation is not neutral, as it follows criteria defined by each platform, influencing what users are able to see and what remains concealed.

These processes are largely automated and managed by algorithmic and AI systems that analyze user activity to decide which content to promote and which to limit, based on criteria such as "making the platform a safe space for inspiration and expression" — criteria that ultimately align with the big tech companies' business models and commercial interests in capturing users' attention and keeping them engaged. More recently, policy changes at various major digital platforms have confirmed that political considerations also shape these criteria.

Shadow banning occupies a unique position within the spectrum of content governance practices, located at the intersection between content moderation and curation. It does not involve the outright removal of content (traditional moderation), but rather an algorithmic

9 Digital Inclusion and Internet Content Governance (OEA/Ser.L/V/II. CIDH/RELE/INF. 28/24. June 2024). Available at: https://www.oas.org/en/iachr/expression/reports/Digital_inclusion_eng.pdf
10 Idem.

intervention that significantly reduces its visibility or reach (negative curation).

Specifically, shadow banning is primarily located within the realm of content curation, as it directly affects how material is distributed and presented to others without deleting it. However, when reduced visibility is applied as a consequence of perceived violations of community guidelines, it also functions as a form of ex post moderation that is less severe than complete removal. The defining feature of shadow banning —and what makes it particularly problematic from a human rights perspective— is its deliberately secretive nature: unlike other moderation measures where users are notified of actions taken, shadow banning operates intentionally without transparency, leaving users with a sense of uncertainty as to why their content is no longer reaching its usual audience.

03.

STUDIES AND CONCEPTUAL EVOLUTION OF SHADOW BANNING

Shadow banning has undergone significant conceptual changes since its origins in early internet forums. Initially, the term referred specifically to a moderation technique in which the comments and posts of users deemed "problematic," i.e., identified as harassing or trolling others, were hidden from the rest of the community, while the affected user was left under the illusion that their content remained visible.[11] This strategy was aimed primarily at preventing sanctioned users from creating new accounts upon perceiving the restriction.

Savolainen offers a more socio-cultural analytical approach by seeking to understand shadow banning as some form of "algorithmic folklore," that is, as part of a set of "beliefs and narratives about moderation algorithms that are passed on informally and can exist in tension with official narratives."[12] This perspective highlights how the term functions as a discursive anchor for diverse but connected experiences of platform governance, unified by a shared sense of secrecy and uncertainty. Other research has emphasized how these types of restrictions disproportionately affect users from marginalized communities. In particular, some authors suggest that users have developed what they call "folk theories"[13] to try to decipher how platform algorithms work. These strategies include altering hashtag use, modifying images, or even creating secondary accounts to verify whether they have been shadow banned.

Research by Kojah and other authors expands on this approach by defining shadow banning as "a controversial aspect of platform governance characterized by the use of opaque algorithms to reduce or demote content."[14] Their analysis describes the practice as a form of "light, insidious" censorship that impacts multiple dimensions of the user experience, including visibility, earnings, mental health, and interpersonal communication.

11 Cole, S. (2018) 'Where Did the Concept of "*Shadow Banning*" Come From?', VICE, 31 July. Available at: https://www.vice.com/en/article/where-did-shadow-banning-come-from-trump-republicans-shadowbanned/ (Accessed: 10 March 2025).
12 Savolainen, L. (2022). Algorithmic lore and the myths of non-promotion. Information, Communication & Society, 25(8), p.1096
13 Delmonaco et al. (2024)
14 Kojah, J., et al. (2025). "Obviously it affects the business side of things too": Algorithmic invisibility and its impact on marginalized social media content creators. New Media & Society, 27(2), p.1

Intersecting characteristics of *shadow banning*

Regardless of its specific manifestation, studies agree on identifying certain defining features of shadow banning:

1. **Secrecy**: A lack of notification or explanation from the platform about the restriction.
2. **Variable gradation**: Shadow banning is not binary; rather it exists in a continuum of reduced visibility.
3. **Cumulative effects:** Different types of shadow bans can coexist, amplifying their impacts.
4. **Indirect detection**: Users develop informal methods to verify whether they are being shadow banned, such as comparing their visibility to that of others or using third-party tool.

The previously described typology, together with its intersecting features, illustrates the complexity and evolution of algorithmic moderation practices on digital platforms. It also shows how the diversification of the term shadow banning beyond its original definition, reflects how users conceptualize and respond to emerging forms of algorithmic governance marked by secrecy and uncertainty.[15]

Unequal Impacts: affected groups and marginalization dynamics

A consistent finding in the reviewed literature is that shadow banning disproportionately affects already marginalized groups. Some studies indicate that this type of moderation is more commonly applied to content related to sexuality, racial identity, or social protest. For example, Instagram has been criticized for censoring images of female bodies, including posts by activists from the *Free the Nipple* movement.[16]

Other investigations have specifically documented how *borderline* content policies negatively affect vulnerable communities such as sex workers, sex educators, and members of the LGBTQIA+ community. Journalistic reports have suggested similar tendencies with respect to Black people,[17] women,[18] and members of the queer community.[19]

One study based on diaries and interviews with eight marginalized content creators documented how "creators from marginalized communities (women, pole dancers, plus-size individuals and/or LGBTQIA+ members)

15 Savolainen, L. (2022).
16 Are, C. (2021). The *Shadowban* Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram. Feminist Media Studies, 22(8), 2002-2019. p.2002
17 BBC (2020) 'Facebook and instagram to examine racist algorithms', British Broadcasting Corporation. Available at: https://www.bbc.com/news/technology-53498685
18 Cook, J. (2019) 'Instagram's *shadow ban* on vaguely 'inappropriate' content is plainly sexist', Huffington Post. Available at: https://www.huffpost.com/entry/instagram-shadowbansexist_%20n_5cc72935e4b0537911491a4f
19 Joseph, C. (2019) "Instagram's murky '*shadow bans*' just serve to censor marginalised communities', The Guardian. Available at: https://www.theguardian.com/commentisfree/2019/nov/08/instagram-shadow-bans-marginalised-communities-queer-plus-sized-bodies-sexually-suggestive

[...] are disproportionately affected by *shadowbanning*."[20] The participants perceived that "people with marginalized identities [...] were more affected by shadowbanning and other forms of censorship than men who created similar content, and than people who fit conventional beauty standards."[21]

Carolina Are's research offers a particularly detailed analysis of how shadow banning affects pole dancers on Instagram, revealing how the censorship of pole dance-related content reflects biased perceptions of female bodies and expressions of sexuality. Her study highlights how even artistic and athletic activities can be incorrectly categorized as "sexually suggestive"[22] when performed by people with certain types of body or from certain communities.

20 Kojah, J., et al. (2025). p.2
21 Idem p.9
22 Are, C. (2021). p.2004

The invisibilized work of users under *shadow banning*

A particularly valuable contribution from recent literature is the identification of the invisibilized work users must undertake to navigate, mitigate, and adapt to opaque moderation systems. The research makes a significant contribution by identifying three specific categories of invisible work:

1. **Mental and emotional work**: The cognitive and psychological burden of constantly anticipating which content might be restricted. As one study participant explained: "Shifting my focus from creation to solving alignment issues with platform policies is a distraction, which disrupts my consistency in producing content."[23]
2. **Aimless work:** fforts made in the hope of avoiding shadow banning that don't directly contribute to the creation of meaningful content. This includes practices such as posting selfies "for the algorithm" after potentially controversial or high-risk content such as pro-Palestine posts, or using algospeak (altering potentially problematic words).[24]
3. **Community work:** Collaborative work among creators to share strategies and mutually support each other. As one participant who took part in the study described: "I've worked with other creators to help boost engagement during a shadow ban. We promote and interact with each other's content during a suspected suspension".[25]

23 Idem p.13.
24 According to Kojah et al. this term refers to "intentionally misspelling words or blocking images to trick the algorithm and circumvent content moderation and suppression, a practice that requires a lot of time and effort." Kojah, J., et al. (2025). p.14
25 Idem p.16

04.

IMPACT ON THE MEDIA AND USERS

Shadow banning functions as a particularly insidious form of digital censorship: invisible, poorly documented, and overwhelming for those who report experiencing it. In Latin America, activists, sex educators, journalists, and small entrepreneurs describe how their posts — which once were viewed by thousands and often addressed matters of public interest— suddenly reached only a few hundred people, with no explanation or warning. This algorithmic reduction in reach not only affects the visibility of their content but also causes deep psycho-emotional impacts marked by constant anxiety, feelings of helplessness, and pre-emptive self-censorship, all of which lead to the silencing of voices in the public debate.

Affected creators are trapped in a digital limbo where they continue publishing for a ghost audience, investing time and resources into efforts that the platforms have quietly condemned to irrelevance.

The issue becomes even more serious when we consider that the moderation carried out by big tech companies affects content related to sexual health, body diversity, journalism, or political criticism, i.e., precisely the kinds of discourse that receive special protection[26] under the inter-American human rights system.

Moreover, this ongoing uncertainty forces users to deploy various strategies, either individually or collectively, and to invest resources not in improving their content or products, but in continuing an unequal struggle against secretive algorithmic systems that relegate their digital presence to a dusky corner.

Thus, what might be considered a simple "business decision" about content distribution becomes a powerful mechanism of social control. Combined with the inability to predict what might trigger restrictions —from mentioning terms like "marijuana," "sex education," "Free Palestine" to showing non-normative bodies— this creates a chilling effect, where people preemptively censor their own expressions on matters of public interest.

What's more, this "digital ostracism" is particularly serious due to its secrecy. Affected individuals continue producing content that virtually no one sees, experiencing a form of isolation that harms both their professional projects and emotional well-being.

26 The jurisprudence of the Inter-American human rights system has established three categories of especially protected speech, recognizing their fundamental role in strengthening democracy and enabling the full exercise of human rights. Firstly, it protects political speech and speech involving matters of public interest, considering that such expressions are essential for the formation of an informed public opinion and for people's participation in democratic processes and public affairs. Secondly, it provides reinforced protection for speech regarding public officials in the exercise of their duties and candidates for public office, on the understanding that scrutiny of those who hold or seek positions of power is crucial for transparency and accountability. Finally, it grants special protection to speech that constitutes an element of the speaker's identity or personal dignity, thus recognizing the importance of freedom of expression for individual development and personal autonomy. This differentiated protection aims to ensure that these forms of speech, which are fundamental in terms of public debate and personal fulfilment, are not unduly restricted, thereby promoting a more open, pluralistic, and democratic society. See: Special Rapporteur for Freedom of Expression of the Inter American Commission on Human Rights. (2009). Inter-American Legal Framework Regarding the Right to Freedom of Expression (paras. 32–56). Organization of American States. https://www.oas.org/en/iachr/expression/docs/publications/INTER-AMERICAN%20LEGAL%20FRAMEWORK%20OF%20THE%20RIGHT%20TO%20FREEDOM%20OF%20EXPRESSION%20FINAL%20PORTADA.pdf

From the perspective of international human rights law, shadow banning represents a particularly problematic form of restriction on freedom of expression. When analyzed through the lens of Article 19 of the International Covenant on Civil and Political Rights (ICCPR) or Article 13 of the American Convention, it fails to meet the essential requirements of legality, necessity, and proportionality that all legitimate restrictions on this right must satisfy. The principle of legality is therefore undermined because these restrictions are implemented through ambiguous terms of service and obscure algorithms that provide no legal certainty about what kinds of expression may be limited. Nor are the principles of necessity and proportionality met when platforms apply these measures in an automated manner, without assessing the specific context, the impact on public debate, or considering less restrictive alternatives such as warnings or labels.

The Office of the Special Rapporteur for Freedom of Expression (SRFOE) of the Inter-American Commission on Human Rights (IACHR) has clearly stated that any restriction must be specific and applied through reasoned decisions that allow for subsequent accountability. Shadow banning, by definition, clearly violates these principles by implementing restrictions without notification, explanation, or any effective possibility to challenge them. It also relies on vague and confusing terms of service, raising the question of whether shadow banning, beyond violating the principle of legality, constitutes a form of prior censorship and a restriction on the right to freedom of expression through the use of indirect means. The SRFOE has emphasized that this right "may not be subject to "preventive measures" or "prior restraints," but only to the imposition of subsequent liabilities for those who have abused its exercise."[27]

The responsibility of major tech platforms in relation to shadow banning is indisputable. The United Nations Guiding Principles on Business and Human Rights (UNGPs) establish that companies have a responsibility to respect human rights, regardless of the ability or willingness of States to fulfill their obligations.[28] For tech giants such as Meta, TikTok, or X, this entails three specific obligations: (i) to identify and prevent negative impacts on human rights, (ii) to implement appropriate processes according to their scale, and (iii) to provide effective compensation mechanisms for victims.

The SRFOE has been clear in stating that "[t]he exercise of the regulatory power of moderation by internet platforms, especially large platforms, should be aligned with the principles of human rights, the promotion of public debate, and the consolidation of democracy in the Americas. They should not only adhere to the norms of the inter-American system, but also align their power with standards of transparency and accountability, based on equality and nondiscrimination. This is essential to create an online environment that respects

27 Idem. para. 91, p. 31
28 Office of the United Nations High Commissioner for Human Rights (2011). Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework. Principle 11, p.13, Available at: https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf

human rights and is free, open, and inclusive, and that fosters the autonomy and rights of users."[29]

The gap between corporate discourse and reality is concerning. While Instagram's Adam Mosseri claimed in 2021 that "if anything makes your content less visible, you should know about it and be able to appeal," the everyday experiences of Latin American creators, activists, and journalists reflect a systematic lack of notifications and effective review mechanisms when their content is restricted.

This contradiction underscores the urgency of regulatory frameworks that recognize digital platforms as spaces for public debate where contemporary democratic discussions take place. In a region where media concentration has historically limited the plurality of voices, social media initially represented a promise of democratization, one that is now undermined by secretive moderation practices that reproduce, and even amplify, pre-existing exclusions, all under the guise of seemingly neutral and technocratic algorithmic decisions.

What is at stake, beyond metrics and outreach, is the right of democratic societies to a diverse, pluralistic, and accessible information ecosystem, and one in which traditionally marginalized voices can participate on an equal footing in shaping the Latin American public debate.

The paradox becomes even more evident when corporate promises are contrasted with the terms and conditions of the platforms. While X (formerly Twitter) explicitly states in its policies that it "will limit the visibility" of certain content and claims that affected users will be notified and may request a review, the findings of this research suggest that such guarantees rarely occur. Meta, for its part, openly admits to reducing the distribution of "problematic content" without committing to informing the affected individuals or providing a clear appeals mechanism.

What happens when the guarantees promised in the terms and conditions become meaningless in the face of users' actual experience? How can people in regions like Latin America defend their rights when they don't even know they're being restricted? Who ensures that algorithmic decisions don't reproduce biases against historically marginalized voices? These questions remain unanswered, while users continue navigating blindly in a digital ecosystem where the written rules rarely align with day-to-day practices, and where the promise of greater freedom of expression yields to a new form of digital exclusion, as effective as it is invisible.

A robust and pluralistic public debate, essential for the wellbeing of a democratic society, requires both states and major tech companies to confront this invisible form of silencing. People must be able to know when and why their content is being restricted, have access to effective mechanisms for challenging those decisions, and be guaranteed that any limitations on their freedom

29 Inter-American Commission on Human Rights (IACHR), Office of the Special Rapporteur for Freedom of Expression (2024). Digital inclusion and Internet Content Governance (OEA/Ser.L/V/II CIDH/RELE/INF. 28/24, p. 54, para. 246). Organization of American States. https://www.oas.org/en/iachr/expression/reports/Digital_inclusion_eng.pdf

of expression will comply with international human rights standards. Without such safeguards, shadow banning will continue to quietly undermine the vitality of democratic debate in the region, creating the illusion of participation while at the same time exerting sophisticated forms of control over public discourse.

05.

SOME CASE EXAMPLES OF SHADOW BANNING

In order to obtain a deeper understanding of how shadow banning operates on digital platforms, this investigation includes the analysis of a series of representative cases from Latin America and the Global South. These are situations in which journalists, activists, and communication or advocacy projects have reported a significant reduction in the reach, visibility, or "discoverability" of their content and accounts, without receiving clear explanations or notifications from the platforms.

One of the main areas of analysis will focus on how Facebook, Instagram, and X have treated accounts and posts that share information in support of Palestine in the context of Israel's military offensive against the Palestinian population, an offensive widely denounced as a genocide against the Palestinian people by experts, governments, UN agencies, and the International Criminal Court.[30] Various public condemnations and studies have documented the use of opaque moderation mechanisms that limit the reach of such content, posing a threat to informational pluralism and the right of people to freely express themselves in digital environments. As part of this investigation, we interviewed Mona Shtaya, Campaigns and Partnerships Manager (MENA) and Corporate Engagement Lead at Digital Action, who uses her Instagram account to share information about the situation in Palestine and has publicly denounced the restrictions these types of content face on digital platforms.

Another documented case is that of the Mexican news outlet *Chiapas Sin Censura* (Chiapas Uncensored), which suffered a drastic reduction in reach on Facebook after receiving a penalty notice for "fraud" associated with a post some years earlier. The sanction, imposed without any prior notice or an effective appeals process, affected both the outlet's visibility and monetization, and forced the team to limit publication of certain content out of fear of further retributions.

In addition to these cases, other significant incidents have been reported, such as that of the Chilean cannabis activist known for her projects *Santiago Verde* and *Muy Paola,* who reported repeated visibility restrictions on Instagram. And in Peru, the account *Emma y yo*, a sex education space led by Alesia, also reported a sharp drop in the reach of its posts, especially those related to sexual and reproductive rights.

Cases in Argentina will also be analyzed, such as that of photojournalist and Ph.D in Social Sciences Cora Gamarnik, who experienced a drastic drop in reach on Facebook without the platform providing her with a mechanism to appeal the decision, as well as that of journalist Sebastián Lacunza, who denounced on X an unexplained decline in the reach of his posts after publishing content about political and media issues.

These cases, contextualized in various thematic and geographic frameworks, help reveal how shadow banning

30 International Criminal Court. (n.d.). Palestine. https://www.icc-cpi.int/palestine; United Nations. (2024, 26 March). Special Rapporteur accuses Israel of committing genocide in Gaza. https://news.un.org/en/story/2024/03/1147976

impacts the right to freedom of expression and the circulation of diverse voices, particularly when they involve some form of critical discourse or come from historically marginalized sectors.

Shadow Banning in Palestine: algorithmic censorship against a people

Since the beginning of Israel's ground and air assault on Gaza, various users have been reporting that Instagram, Facebook, TikTok, and X are limiting the visibility of their pro-Palestinian posts, even though the content is not deleted.[31] Across different platforms, it's possible to find users who post content in support of Palestine, describing how they are no longer allowed to livestream, or who have noticed an acute reduction in engagement or video views, or are receiving messages from other users saying they're unable to leave comments on their posts.

"Authors, activists, journalists, filmmakers, and users have said that the platforms are hiding posts that contain hashtags like #FreePalestine and #IStandWithPalestine, as well as messages expressing support for Palestinian civilians slaughtered by Israeli forces."

"The company claims there's no bias, and that everyone is treated the same, but that's simply not true. We haven't seen a single Israeli user complain about being shadow banned, not even in terms of trending topics or blocked comments," says Nadim Nashif, founder of 7amleh, also known as the Arab Center for the Advancement of Social Media.[32] In an attempt to bypass shadow banning, some users try to trick the algorithms by replacing terms considered pro-Palestinian with others in their hashtags or posts. In Meta's case, censorship of pro-Palestinian content even prompted nearly 200 of its employees to sign an open letter addressed to Mark Zuckerberg in December 2023.[33]

In November 2024, Palestinian journalist Younis Tirawi, who is known for exposing Israeli war crimes, began receiving messages from various users on X telling him about "glitches" when trying to follow his account on the platform.[34] The media outlet Decensored News demonstrated the glitch through screen recordings and reported that Tirawi "suddenly lost most of his followers," many of whom said they had been unfollowed "involuntarily and that X wouldn't allow them to follow him again." It would appear that after clicking the "follow" button, the page such

31 "Denuncian shadowban de grandes plataformas a contenidos sobre Palestina" (Condemnation of shadow banning of (pro-)Palestinian content by large tech platforms). https://www.observacom.org/denuncian-shadowban-de-grandes-plataformas-a-contenidos-sobre-palestina/
32 "Usuarios propalestinos denuncian 'shadowban' en plataformas de redes sociales" (Pro-Palestinian users denounce 'shadowbans' on social media platforms) https://www.youtube.com/watch?v=xm9llqJjg1A&ab_channel=FRANCE24Espa%C3%B1ol
33 "Dear Mark Zuckerberg and Leadership" https://metastopcensoringpalestine.com/
34 "Platform flaws or hidden censorship?" controversy over restrictions placed on Palestinian journalist on X https://www.observacom.org/fallas-o-censura-oculta-polemica-por-restricciones-a-periodista-palestino-en-x/

users were on would then automatically revert to how it was before they (re)tried to follow him.

Although X claimed it was a "technical issue," it's highly significant that the same thing had happened one month earlier to the account of @Palestinahoy01 on the same platform. For seven days, Elon Musk's platform would not allow users to follow the account, and its follower count fluctuated constantly.

An almost identical situation was described by Mona Shtaya, a Palestinian expert in digital rights and content moderation, currently Campaigns and Partnerships Manager for the Middle East and North Africa and Corporate Engagement Lead at the organization Digital Action. She recounted to us her experience as an Instagram user, and how she was affected by the mechanisms of shadow banning: "My account has more than 20,000 followers and focuses mainly on digital rights and holding big tech accountable. For the past two years, I've been intensely focused on Palestine because of the situation there."[35]

According to Mona when we interviewed her, the first documented cases of shadow banning in Palestine occurred in May 2021, when Israeli forces attempted to forcibly displace Palestinian families from the Sheikh Jarrah and Silwan neighborhoods in Jerusalem. In that same period, a 12-day offensive on the Gaza Strip also took place. In that context, reports began to emerge of visibility restrictions on Palestine-related content across digital platforms.

Regarding her own Instagram account, Mona said the shadow banning began in November 2023, roughly a month after the start of the current Israeli military offensive on Gaza. Until then, she had not experienced any similar restrictions.

Mona realized something was happening with her Instagram account when she began comparing her stats. For example, when she posted a selfie, her stories would receive over 2,600 views, but if she shared "a story criticizing Meta for silencing Palestinians, that story wouldn't even reach 200 views." Or rather, a fall in visibility of over 90%. Mona used various strategies to test her reach and to try and understand what was going on. "I undertook collaborations with big accounts, with more than 10 million followers. That week, my posts reached over a million accounts, but my stories barely hit 200 views. That shows there's a problem. It doesn't make any sense that I'd have a reach of a million users and yet my stories received so few views," she told us.

In some cases, people who tried to access Mona's Instagram account couldn't find it. Others received a warning message when attempting to send her a private message: "Are you sure you want to message this person?" the platform asked when they clicked "send."

Mona Shtaya explained that the shadow banning of her account had a clearly identifiable starting point: "I wasn't being shadow banned at the beginning of the genocide. I then made a video that went viral and had around 200,000 views. It basically criticized Meta's complicity in the genocide. That video went

35 Interview of Mona Shtaya carried out by Observacom and Digital Action (2025), April 3, 2025.

viral, and everything that came after was madness. Everything I posted from then on was heavily *shadow banned*."

Before that moment, Mona had worked with human rights defenders who had faced similar restrictions, but she had never experienced it on her own social media account. "It was the first time it happened to me. And after that, I felt that any content I shared encountered the same problem. Because during the genocide, I basically didn't post anything except content related to the genocide and digital rights. And all of that content clearly had trouble reaching people," she told us.

In her testimony, Mona Shtaya shared both her own observations and those of other users who documented shadow banning cases on Meta's platforms, all linked to specific keywords and symbols associated with the Palestinian cause. While she herself didn't usually use hashtags in her stories, she explained that many people who did experienced a noticeable reduction in reach when they included terms like "Palestine," "Gaza," "genocide," or "apartheid."

One of the most significant pieces of evidence regarding this type of practice was published by The Intercept in October 2024. According to that investigation, Meta had applied visibility restrictions to posts containing the red triangle, a symbol many users began using to represent support for Palestine, but without the platform clearly disclosing this policy.[36] Mona noted that while there had already been suspicions about this conduct, the report served as concrete proof that visibility-reducing measures were indeed being applied.

Another pattern observed involved Instagram comments containing the Palestinian flag, heart emojis in the colors of the flag, or phrases like "Free Palestine." These comments would appear hidden or marked as "hidden comment," with no clear warning or indication as to why. Following an investigation prompted by Mona herself, *The Intercept* reported the issue on its own account. However, when they asked Meta for an explanation, the company responded that it only concealed comments when it detected "hostile speech." Activists interpreted this explanation as a sign that, at least in practice, Meta was classifying terms and symbols related to Palestine as hostile content, even though it did not publicly admit to this or make any transparent justification for it.

According to a 2022 survey conducted by the Center for Democracy and Technology, the platform with the highest percentage of users reporting experiences of shadow banning is Facebook (8.1%), followed by what is now X (4.1%), Instagram (3.8%), and TikTok (3.2%). This opaque form of censorship tends to affect certain social movements more "frequently and harshly." In addition to what is happening to users who post pro-Palestinian content, similar patterns have been observed with the Black community, the Black Lives Matter movement, and the LGBTIQ+ community.

Why should platforms respond to complaints about the censorship they

36 The Intercept, "Facebook and Instagram Restrict the Use of the Red Triangle Emoji Over Hamas Association" https://theintercept.com/2024/10/02/meta-facebook-instagram-red-triangle-emoji/

impose? According to the United Nations Guiding Principles on Business and Human Rights (UNGPs),[37] companies have a responsibility to avoid infringing on human rights, to identify and address the human rights impacts of their operations, and to provide meaningful access to remedy for those people whose rights have been violated.

For social media companies, this includes aligning their content moderation policies with international human rights standards, ensuring that decisions to remove or restrict content are transparent and not overly broad or biased, and applying their policies consistently. Even though Meta allows a significant amount of pro-Palestinian expression, this does not justify the unwarranted restriction of peaceful content supporting Palestine, which runs counter to the universal rights to freedom of expression and access to information.

Sebastián Lacunza: invisible sanctions on X

Between late 2022 and September 2023, Argentine journalist Sebastián Lacunza began noticing a sharp drop in interactions on his account on the social network X (formerly Twitter). The statistics he occasionally checked showed unusually low numbers: a consistent loss of followers and very little engagement, even when cited by other high-profile users.

"There was a very steep drop in all the stats, and I was constantly losing followers. I also noticed that sometimes people with a large follower base would quote me, and even that seemed to have no impact, which was strange,"[38] he told us when we interviewed him for this report.

Lacunza later learned that this phenomenon might be related to a practice known as shadow banning, though at the time, he was unfamiliar with the term. It wasn't until mid-2023, after hearing about it from colleagues and activists, when he began to understand that what he was experiencing might not just be a normal drop in engagement, but rather a covert sanction imposed by the platform itself.

Confused, Lacunza considered paying for the premium subscription, although he wasn't sure it would actually solve anything. He eventually decided to publish a tweet denouncing the situation.[39] "I posted the tweet, and within 48 hours it was lifted; or less than 48 hours." Thus in under 48 hours after he tweeted about the shadow banning, his visibility began to return to normal. "It was immediate," he confirmed. There was no official notification from X, or any automated response explaining what had happened. The interactions simply began to increase, and the stats, which

37 Available here: https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusiness-hr_en.pdf
38 Interview of Sebastián Lacunza carried out by Observacom and Digital Action (2025), March 18, 2025
39 Posted on X by Sebastián Lacunza (15/09/2023) https://x.com/sebalacunza/status/1702735419777880281

had been falling for months, started to go up.

"At first, the recovery was quite fast: I was gaining between 500 and 1,000 followers per month. Then it stabilized, but something clearly changed," he noted. In hindsight, Lacunza acknowledges that the invisible penalty he experienced could only be confirmed through his own observation and by comparing engagement metrics. "During the months I was shadow banned, I was operating at a level where getting 15 likes was a lot. And then, suddenly, after that tweet, I had a post that got 21,000 likes. I'd never had anything like that before," he explained. Despite the improvement, he still doesn't know whether the restrictions on his account were completely lifted or if some kind of partial constraint remains.

Another clear sign of shadow banning was the loss of his account's "discoverability." Lacunza recalled that during the period when he suspected he was under some form of algorithmic sanction, it was notably difficult to find his profile, even when typing his full name into the platform's search bar. "I would search for my name, and my account wouldn't appear. But three or four fake accounts that someone had created using my name and avatar would show up," he explained. This matched one of the indicators often cited by other users in forums or tutorials on how to detect shadow banning: being excluded from search results despite using the exact account name of the user.

This lack of visibility wasn't due to a system error or a general issue with the search engine, rather it affected only his account, while the fake profiles under his name were easily found. "That's an important detail, because it wasn't that there were no results: the fake accounts showed up, but mine didn't," Lacunza noted.

His case clearly illustrates how shadow banning can operate discretely, without users being aware it's happening, and without them having access to any formal appeals or review mechanisms. At the same time, it demonstrates how such forms of sanction can be lifted just as discretely, reinforcing the secretive nature of these practices and the lack of safeguards for those whose visibility is restricted in digital environments, which today function as public spaces for debate and expression.

.

Shadow banning of the independent media outlet Chiapas Sin Censura

The Mexican news outlet *Chiapas Sin Censura* (Chiapas Uncensored), founded in 2012, represents a significant case of *shadow banning*, particularly on Facebook. Its founder and editor-in-chief, José David Morales Gómez, has condemned the severe visibility restrictions on social media that the outlet has been experiencing, especially on Facebook. According to David, the news agency experienced a drastic drop in reach after receiving a sanction for an

alleged violation linked to a post published four years earlier. The platform flagged the content as "fraud" without offering further explanations or any effective way to appeal. The post in question featured a young cancer patient requesting support from boxer Canelo Álvarez, who is known for his charitable actions.

"This post, which was real, was also circulating on other pages without any fuss being made. And then suddenly we were penalized for fraud and completely ostracized," David told us. Up to that point, *Chiapas Sin Censura* had been registering over 100 million monthly views; after the sanction, its reach dropped to 28 million. "Our organic growth stopped, followers began to drop off, we lost monetization, and we got no response to our appeals," he explained.

After receiving the penalization, the outlet attempted to appeal the decision through various channels, but without success. David said he used the appeal option provided by the digital platform at the time of notification: "It said we'd get a response within four business days, but they never answered." He also sent emails, opened manual reports from the admin profile, and subscribed to the Meta Verify service —a paid tool promising personalized support— in hopes of getting a quicker response. The subscription cost around 7,000 Mexican pesos per month.

During that process, he managed to speak once with someone at Meta, who told him that the case had already been passed on and that he would receive a response within 24 hours. That was the only direct interaction he had. After that, there was no further contact. "I tried to submit another report and I was told there was already an ongoing discussion. There was nothing more I could do," he said.

The lack of effective appeal mechanisms was one of the most frustrating aspects of the experience. The sanction remained active for weeks, directly affecting the outlet's reach, monetization, and engagement with its audience. "If it hadn't been for the assistance of an organization that stepped in to help us, I think the penalization could well have lasted a year," David stated, referring to the support the media outlet received.

This case highlights the limitations of Meta's internal appeals process, directly affecting the rights of users to defend themselves. The absence of a clear review mechanism, i.e., one that is accessible, transparent, and with reasonable response times, represents a critical vulnerability for media outlets that rely almost exclusively on digital platforms for their distribution and survival. It also negatively impacts the public's right to receive information from these media outlets.

Furthermore, the impact was not just technical or economic, as it also affected their journalistic work. According to David, the outlet began avoiding the publication of articles referencing sensitive issues, such as cases of violence, vulnerable children, or allegations of organized crime, for fear of further sanctions. "Sometimes we'd say: this story isn't worth risking more sanctions. Even if it's true, we'd best not publish it as it could do us more harm than good."

The blockade also had emotional and editorial consequences. "I went for days without wanting to publish. What for if the work doesn't reach people?" said David. As with many other digital outlets in the region, *Chiapas Sin Censura's* main public channel is on Facebook: "If tomorrow they decide to delete our page, we lose 12 years of work and 10 families lose their income" he reflected.

Although the sanction was eventually lifted, this episode exposes the lack of accessible mechanisms to challenge automated and unilateral decisions. Moreover, it also shows how the threat of invisibility directly affects the coverage of issues of public interest and the sustainability of local and critical media.

@MuyPaola: if you mention cannabis you're made invisible on social media

Sitting in front of her phone during the video call for our interview, Paola Díaz shows us her Instagram account stats: "Your account can't be shown to people who don't follow you,"[40] reads the notification, without further explanation. Clearly frustrated, she points to the brutal difference in reach between seemingly similar posts: a piece of activism content gets 5,300 views, while a festive-themed post reaches 70,000.

The Chilean cannabis activist, creator of the accounts @Stgoverde[41] and @muypaola, tells us about the systematic reduction in her account's reach without her receiving any explicit notification or transparent justification. And it's not that they directly block her account, rather it's simply rendered nearly invisible. "The *shadow ban* starts showing your posts less and less, and only to your followers," explains Paola as she documents how her posts, which once received hundreds of thousands of views, now barely reach her immediate circle. Even her discoverability through searches has been severely restricted: "When people search for me, they have to type my full name and press enter because I don't even appear in the suggestions," she adds, describing a system seemingly designed to make certain profiles disappear from the public radar.

Paola's case illustrates how algorithmic intervention by platforms like Instagram can operate with a form of re-doubled invisibility: it hides the content from other users and, at the same time, conceals the restriction process from the creator themselves. Unlike a direct account suspension, *shadow banning* lacks formal notifications or explanations about which terms or topics are being penalized. Paola has identified specific patterns after years of documentation. For example, mentioning terms like "marijuana," "420," or

40 Interview of Paola Díaz carried out by Observacom and Digital Action (2025), March 12, 2025.
41 At the time of the publication of this report, the @StgoVerde account remained suspended by Instagram, after numerous cycles of restriction and reinstatement. The platform has not only suspended the account, but had previously and systematically removed content of public interest, including journalistic reports on cannabis culture and educational material produced by the activist.

even sharing information about cannabis culture and associated rights immediately triggers invisible filters that drastically reduce her public outreach. "Meta flags words you're not supposed to say because they will then *shadow ban* you," she reveals, underlining a highly specific system of thematic moderation.

This covert system of moderation has had devastating consequences that go far beyond mere online visibility. In September 2024, after years of intermittent *shadow banning*, Instagram finally deactivated one of Paola's main accounts after requesting biometric verification. But astonishingly: "They kept charging me for the subscription [to Meta Verified]," she says indignantly, describing how Meta continued to bill her for a verification service on accounts she could no longer access. Once again, the lack of transparency is compounded by questionable commercial practices, with no effective appeals or redress mechanisms in place.

The emotional and professional toll, as in the case of "*Emma y yo*," has had a profound impact on Paola. "Now, whenever I go on social media, I start to feel anxious," she admits. The constant fear of digital ostracism and being silenced has created a form of preventive self-censorship that frustrates her ability to educate on crucial topics: "I can't talk about harm reduction. There are so many times I'd like to address issues related to drug trafficking [but] I feel I can't warn my community about the real risks." The fear of being expelled from the platform has had an inhibiting effect on her right to participate in the cannabis debate in Chile. The paradox is clear: while platforms justify their moderation policies as protections against harmful activities, they are silencing the very voices that could help prevent the real dangers related to the use and criminalization of cannabis.

Paola's response to this situation has been multidimensional, combining legal and collective visibility strategies. Together with a legal team, she filed a constitutional protection claim in Chile that was rejected and is now under appeal. At the same time, she has reported these practices to the National Consumer Service (SERNAC), arguing that consumer rights are being violated when Meta charges users in Chile without adhering to local jurisdiction. "They're billing people in Chile, so they have to operate in accordance with Chilean jurisdiction," she insists, underlining a regulatory challenge that transcends national borders.

Paola's case illustrates a global issue that affects various communities, from cannabis activists to human rights defenders in Palestine, as well as creators of inclusive fashion content featuring non-normative body types. Paola has sought to build transnational networks with affected individuals from Argentina to Thailand, highlighting the systematic nature of this selective form of censorship. "I've formed one group with many people from different parts of the world, Argentina, Thailand, Spain, Mexico, Uruguay, all cannabis activists, and all of them have been taken down," she explains, describing how creators with hundreds of thousands of followers have experienced the same pattern of gradual invisibilization followed by total removal from their social media networks.

Cora Gamarnik: suspensión y caída drástica del alcance en Facebook

"What I experienced was a drastic reduction in the reach and interactions of my Facebook account, especially after I shared a post related to the *Lago Escondido* scandal, where I posted a screenshot of the judges' group chat," explains Cora Gamarnik, a Doctor of Social Sciences. On December 12, 2022, Gamarnik posted on her Facebook page about a trip involving judges, government officials, and members of the Clarín Group (*a media conglomerate*), in Argentina, along with the exchange of messages between them that had been made public that same week.

Facebook removed Gamarnik's post and subsequently suspended her account. "When I published that post, they suspended my account claiming that I was spreading hate speech. So I filed a claim and explained that my post was actually criticizing the hateful messages in the screenshot I shared," she explains. The suspension lasted for a few days.

Her post consisted of several screenshots from the chat, which included phrases like "let's clear out all the Mapuche"[42] and a text she wrote denouncing the implicit racism of such a discourse. "Clearly Facebook didn't read the text in question. After that, my posts started having very limited reach, which was totally different from what I was used to," explained the research associate of Argentina's National Council for Scientific and Technical Research (CONICET).

After noticing a change in the reach of her publications, with significantly fewer interactions than she had received before, Gamarnik gradually stopped using Facebook and began to be more active on other social networks, though she never fully abandoned the platform. "Up to that point," says Cora, "my Facebook account had a wide reach, and there were things I knew for certain would go viral or get shared immediately if I were to post them. What I started to see was that those same kinds of posts got no traction." What's more, many of her followers who regularly read her posts began telling her they could no longer see them on Facebook, and had even thought she'd stopped posting entirely on the platform.

Currently, if one examines the evolution of her metrics, it becomes clear how the interactions on Cora Gamarnik's posts started falling off from December 12, 2022 onwards. Whereas earlier posts on similar topics had received over 2,000 "likes," many dozens of comments, and more than 300 shares (such as in early December of that same year), subsequent posts dropped to barely 13 likes, 2 shares, and no comments.

The case of Cora Gamarnik joins that of other public figures, such as journalist Sebastián Lacunza, who in September 2023 posted on his X account that he had "*realized*" he had been *shadow banned*, when his "posts' 'impressions' dropped to a fifth of their usual number and interactions abruptly tailed

42 Translator's note: the indigenous inhabitants of south-central Chile and south-western Argentina

off." He also stated that he no longer appeared in X search results for users who didn't follow him, which all had a significant impact on the outreach and visibility of his account.

"Emma y yo", when comprehensive sex education is suppressed on Instagram

"I've come close to throwing in the towel and saying 'to hell with it all'... let the content stay where it is; oh *how lovely;* it's all over, I'm done."[43] The frustration expressed in the words of Alesia Lund, creator of the sex education project "@emmayyoperu"[44] on Instagram, underlines the invisible reality faced by digital educators in Latin America. Without warning or explanation, she saw her Instagram account -which had grown to nearly 68,000 followers- systematically fade away right before her eyes, losing more than 10,000 followers in two years, while her posts dropped from 15,000 or 20,000 views to barely three to five hundred. What Alesia describes fits squarely within the phenomenon examined in this investigation: a form of content moderation that operates like some type of algorithmic ghost, leaving no trace, issuing no notifications, simply making content disappear from the public radar.

"*Emma y yo*" was launched in 2019 as an urgent response to the lack of accessible content on sex education in the context of Latin America. Through illustrations, infographics, and carefully crafted educational materials, the project quickly became a regional reference for discussing sex and sexuality with children, adolescents, and young adults. Its approach combines informative content with an engaging visual style that demystifies taboo subjects, from anatomy to consent, using direct and easily accessible language. With two published books and a community made up mostly of young women between the ages of 18 and 24, Alesia's initiative succeeded in filling a critical educational gap in a region where institutional sex education remains derisory. The organic growth of Alesia's project on social media reflected not only the quality of its content but also the immense social need for trustworthy information on a historically suppressed topic.

What's most revealing about this case is the complete lack of communication from the platform. Unlike a traditional suspension, where a user is issued a notification for violating guidelines, Alesia never received any warning regarding her sex education content. This gradual invisibilization has occurred alongside changes to Meta's content policies, coinciding with the post-pandemic period during which various sex education accounts began reporting similar issues, Alesia told us. "I started seeing U.S. accounts I follow getting *banned*. Observing from afar I said to myself, *they'll be coming for us next*." Not long after, Alesia began losing followers and her posts became less visible. And this is a pattern that can be confirmed

43 Interview of Alesia Lund carried out by Observacom and Digital Action (2025), March 14, 2025
44 Emma y yo Peru, Instagram, April 14, 2025: https://www.instagram.com/emmayyoperu/

when compared to other on-line creators: content covering topics deemed controversial (sex education, feminism, reproductive rights) is subject to significantly lower distribution than other types of content.

The impact of this invisible, shadowy, covert form of moderation goes beyond numbers, given the significant consequences for the emotional and professional wellbeing of those affected. "My output has dropped a lot," Alesia confesses, describing cycles of frustration that led her to pause her work for weeks at a time: "Why on earth should I even try, if no one can see what I'm doing?" she ponders. This gradual erosion of reach not only dampens creative motivation but also directly threatens the sustainability of independent educational projects, pushing creators to constantly migrate across platforms (from Facebook to Instagram, then to TikTok and YouTube) in search of visibility. It's a situation, notes Alesia, that is truly exhausting. This is why she is now trying to move on from Instagram and focus on platforms like YouTube to see if, finally, for the first time in her life, she can monetize her educational project.

This case illustrates how *shadow banning* functions as a control mechanism, one that lacks transparency, while disproportionately affecting voices addressing sensitive topics. The algorithm appears to penalize terms like "sexuality," "vulva," or "sexual education" — even when presented in educational contexts. As a result, Alesia stopped using certain hashtags. What's most concerning is that, unlike traditional censorship, this system offers absolutely no capacity for users to appeal: with no notification of which rules were supposedly violated, content creators cannot simply adjust their content to comply with specific guidelines. This form of invisible control poses a fundamental challenge to freedom of expression in digital environments, particularly for those creators in Latin America addressing sexual health from educational and feminist perspectives.

Alesia's testimony also points to a broader phenomenon affecting the accounts of educators across the region, one closely associated with *shadow banning*: the commodification of perception on platforms like Instagram is creating increasing pressure to monetize reach. "They're pushing us to the point where we have to pay," she says, explaining how even new accounts with no controversial content face severe limitations unless they invest in promotion. "Before, growth was organic and fast, and now it's incredibly slow, views just don't happen," Alesia tells us. This reality raises important questions about democratic access to information, particularly when educational content on sexuality, widely regarded as a matter of public interest, is restricted by commercial algorithms that fail to distinguish between sensitive and educational content. This creates a double standard that especially penalizes non-profit initiatives.

Ultimately, the trends Alesia describes suggest a form of algorithmic discrimination based on subject matter. The way platforms prioritize individuals speaking over photos and illustrations reveals a structural bias: certain formats are favored, while creators addressing sensitive issues through educational texts or graphics are disproportionately affected. This technical bias worsens

invisibilization by showing, on the one hand, that entertainment videos without educational content on issues such as sexuality circulate freely, and on the other, that infographics with a scientific pedagogical approach are downgraded.

When algorithms judge bodies: the case of Love&Lust and digital invisibility

This kind of algorithmic discrimination and *shadow banning* not only affects educational content on sexuality. Independent entrepreneurs who build communities around so-called "sensitive" products face the same invisible wall of silencing. The case of Paula Labra and her inclusive lingerie business also illustrates this phenomenon.

"I filmed a friend trying to look up my account, and it just didn't exist,"[45] she explained when we interviewed her. With these words, Paula, owner of the e-commerce lingerie brand (@love-lust.cl), which has 240,000 followers, describes the moment she documented what thousands of creators suspect but rarely manage to prove: her account was being subject to *shadow banning*, rendering it virtually invisible on the very platform that sustains much of her business.

The case of Love&Lust clearly exposes the secretive moderation mechanisms that disproportionately affect content creators. Paula recounts how Instagram enforces discriminatory criteria based on body type. "On TikTok, skinnier women weren't taken down, but curvier women were. If they were a bit fuller or had bigger [breasts], forget it, they took down my posts." This algorithmic bias amounts to a form of censorship that goes beyond explicit policies and reinforces prejudice against non-normative bodies, particularly affecting content that celebrates female body diversity or features anatomy-related products like post-mastectomy nipple prosthetics or menstrual underwear.

The effects of *shadow banning* are both devastating and measurable. "I felt it in sales. It was as if we didn't exist, even though we're a brand that invests $3,000 dollars a month in advertising," Paula told us. Her testimony shows how Meta charges for advertising services, while simultaneously limiting the reach of the same accounts paying for visibility. This issue, as we'll see later, also extends to other services offered by the platform. Consequently, this situation has created a new form of algorithmic illusion in which creators pay USD 45 a month for Meta Verified "out of fear alone," and without any guarantee of protection from invisibilization. "I'm spending over $500 dollars a year just out of fear, that's ridiculous," she laments. The psycho-emotional toll is just as serious: "I live in fear. Before I fall asleep, I always pray that I don't lose the account," she admits, referring to a constant state of anxiety that affects her wellbeing and creative capacity.

Paula's response to this systematic censorship reveals the inventive resistance of Latin American creators confronting

45 Interview of Paula Labra carried out by Observacom and Digital Action (2025) March 19, 2025

secretive algorithms. She developed community-based strategies, forming a "group of influencer friends" to constantly repost her content, aiming to break the *shadow banning* through mass engagement. At the same time, she meticulously documents each instance of censorship, saving screenshots that reveal the differential treatment of small accounts versus large brands like Calvin Klein or Savage X Fenty, which can share much more explicit content without facing any negative consequences. Paula's case, while not centered on public-interest topics, clearly illustrates how algorithmic moderation is shaping a digital landscape where certain bodies and subjects are systematically made invisible. It also forces small entrepreneurs to divert significant resources, and not toward improving their products, but toward blindly fighting a covert system that subjects their existence to the constant fear of being excluded.

06.

Shadow banning in platform rules

One of the aims of this investigation was to explore what information digital platforms provide to users regarding content moderation measures that affect the reach and visibility of their posts or accounts —commonly referred to as *shadow banning*— as well as the options available to appeal or challenge these decisions.

This section examines what the terms and conditions of X (formerly Twitter) and Meta (parent company of Facebook and Instagram, which share unified rules) actually say. It also reviews statements by executives and spokespersons to determine how these platforms define, explain, and regulate "reach reduction" or "visibility reduction," and whether they offer users any guarantees or give them any rights to defend themselves, if such mechanisms do exist.

Terms and conditions of X (formerly Twitter)

In the case of X's rules, the platform includes a specific section outlining the measures it may apply when it deems that content violates its policies. It distinguishes between actions taken at different levels: measures applied to either a post or an account, or those applied to direct messages.

Among the measures X may apply to a post is the explicit possibility of "limiting post visibility." Its guidelines explain that content-level actions are taken "when a specific post violates the X Rules, including posts that share or reproduce other posts by posting screenshots, quote-posting, or sharing post URLs that violate our Rules."

Visibility limitations on posts are described as follows: "Where appropriate, we will restrict the reach of posts that violate our policies and create a negative experience for other users by making the post less discoverable on X." Possible measures include:

1. Excluding the post from search results, trends, and recommended notifications.
2. Removing the post from the For you and Following timelines.
3. Restricting the post's discoverability to the author's profile.
4. Downranking the post in replies.
5. Restricting Likes, replies, Reposts, Quote posts, bookmarks, share, pin to profile, or Edit post.

As of April 2023, X began publicly labeling posts identified as violating its policies, informing both the authors and readers that the visibility of the post is being restricted. Authors have the ability to request a review of these labels if they believe the visibility limitation is being applied in error. However, the platform does not clearly specify how to submit such a review request, unlike the more detailed procedures available for appealing accounts that are suspended or locked.[46]

46 https://help.x.com/en/forms/account-access/appeals/redirect

Furthermore, X provides for a "public interest exceptions"[47] for certain content that, although violating the rules, is considered to be of sufficient public relevance to remain accessible. This exception applies primarily to posts from high-profile accounts representing current or potential members of governmental or legislative bodies. In such cases, the post is placed behind a notice and its visibility is limited, but it remains accessible on the platform.

These are the criteria for such exceptions::

1. The post violates one or more X Rules;
2. The post was shared by a high profile account; and
3. The account represents a current or potential member of a local, state, national, or supra-national governmental or legislative body::

   a. Current holders of an elected or appointed leadership position in a governmental or legislative body, or
   b. Candidates or nominees for political office, or
   c. Registered political parties.

Terms and conditions of Meta (Facebook and Instagram)

Meta, the parent company of Facebook, Instagram, Threads, and Messenger, includes explicit references in its official documents to the practice of reducing the visibility of certain content, even when that content does not violate its Community Standards.[48] This measure falls under its content "curation" policy, which has been structured around a three-pronged approach since 2016: remove, reduce, and inform.

The "remove" option is the most easily recognized, as it involves classic forms of content moderation, such as removing posts or deleting user accounts. By contrast, the "reduce" approach, outlined in the *Reducing the Distribution of Problematic Content*[49] section of the company's Transparency Center, aims to limit the circulation of what it labels "problematic content," which, although it does not directly violate the rules, may "create negative experiences" or be considered "low quality." In such cases, Meta states that it reduces distribution in the feed and recommendations, without deleting the content or notifying the user that they have been sanctioned.

The company uses a broad and ambiguous classification of "problematic content" that may be subject to reduced distribution, as follows::

- Low-quality content such as clickbait and engagement bait.
- Links to websites overloaded with ads, slow to load, or not functioning properly.
- Low-quality comments that are copied and pasted repeatedly.

47 https://help.x.com/en/rules-and-policies/public-interest
48 As of November 12, 2024, Meta unified its community standards into just one that applies to its four social networking and messaging services, i.e., Facebook, Instagram, Messenger and Threads. The community standards can be read in full here: https://transparency.meta.com/en-us/policies/community-standards/
49 https://transparency.meta.com/en-us/enforcement/taking-action/lowering-distribution-of-problematic-content/

- Content with limited originality that is mostly repurposed from other sources.
- Low-quality videos that misuse video formats or livestream videos.
- Misinformation and disinformation.
- Content from creators who repeatedly violate Meta's policies.

As can be seen, the list ranges from clickbait and repetitive comments to unoriginal content and posts flagged as misinformation. It's important to note that the latter can result in algorithmic sanctions without any human review or accessible appeals process for affected users.

Meta justifies this practice as a way to prioritize user experience. However, the company does not provide notifications or clear appeals mechanisms in these cases, making it impossible for users to understand whether they are being sanctioned or why. This lack of transparency is especially concerning when automated decisions affect the circulation of legitimate content or content related to matters of public interest.

Moreover, the recommendation guidelines Meta applies to Facebook and Instagram reinforce this type of practice. These guidelines state that suggested content —such as what appears in "Explore," "Suggestions," or "Reels"— is governed by internal criteria designed to avoid amplifying material that the platform considers inappropriate or irrelevant for certain audiences. However, the exact parameters guiding these decisions are not made public, nor are transparent mechanisms provided for users to understand why their content has stopped circulating normally.

Unlike other forms of content moderation, such as post removal or account suspension, in cases of "reduced distribution of problematic content," Meta's policies do not provide the option to appeal or request a review. This further deepens the opaque nature of the process, as users receive no notification and have no tools to challenge or reverse decisions that directly affect the visibility of their posts.

Meta's Community Standards also include specific restrictions on content related to cannabis and its derivatives. In the section on Restricted Goods and Services, the company states it may restrict posts that "coordinate or promote (i.e., speak positively about, encourage use, or provide instructions for use or production of) marijuana and products containing THC or related psychoactive components."[50] While this policy does not necessarily result in account deletion for those who share such content, its inclusion in the Standards allows the company to apply visibility reduction or reach-limiting measures, at least for users under 18 years old. This regulatory framework could be linked to the case of the Instagram account "@muypaola," whose creator reported a sharp decline in the interactions and visibility of her posts. However, in her case, the reach limitation did not appear to apply only to minors, but to all users in general.

50 https://transparency.meta.com/en-us/policies/community-standards/regulated-goods/

07.

CONCLUSIONS

This investigation has enabled us to confirm that shadow banning comprises a set of increasingly frequent practices carried out across major digital platforms, and that its impact on freedom of expression, the circulation of public-interest information, and democratic participation in digital environments is considerable.

Through the analysis of specific cases in Latin America, interviews with experts, and a review of the terms and conditions of companies like Meta (Facebook and Instagram) and X (formerly Twitter), we can clearly state that these covert forms of moderation operate as mechanisms of algorithmic silencing that affect content creators, activists, journalists, and historically marginalized communities. Moreover, they conceal content through processes that are themselves hidden from users.

One of the main findings is that reduced visibility can have consequences comparable to, or even more lasting than, the outright removal of content. Although content is not deleted, its circulation is severely restricted, limiting both its reach and the ability of the user to participate in public debates. This type of sanction, often automated, can produce disproportionate effects that users only become aware of after noticing a sudden drop in interactions, views, or the discoverability of their accounts.

It was also found that platform transparency regarding these practices is minimal or entirely absent. Meta acknowledges applying measures to "reduce the distribution of problematic content" —a broad and ambiguous category— but does not inform users when such measures are applied or provide clear channels for appeal. X, for its part, notes that it may limit the visibility of certain posts and promises a possible review, but does not detail the procedures or guarantee that they are accessible or effective. In practice, this leaves affected users in a state of helplessness, with insufficient information to contest the penalty and no tools to reverse it.

Our investigation also reveals that decisions about which content to downgrade in visibility are not neutral, and that moderation algorithms not only reproduce existing social biases but also amplify structural inequalities, which they do by limiting access to dissenting or non-mainstream voices.

Lastly, the study confirms that these practices violate not only individual rights, but also undermine the public and democratic nature of digital spaces. If platforms continue to operate without transparency, accountability, or adequate mechanisms of redress for users, the risk is not just the covert censorship of certain voices, but the overall deterioration of democratic debate on the Internet.

In light of this scenario, there is an urgent need to move toward regulatory and legal frameworks that guarantee transparency, due process, and the right to a proper defense when major digital platforms apply measures to reduce reach and visibility.

OBSERVACOM

www. observacom.org