Comunicação e Sociedade, vol. 50, 2026, e026016
https://doi.org/10.17231/comsoc.50(2026).7084

Materiality of Hate Governance on Instagram: Immigration-Related Conflicts in Portugal

Dalvacir Xavier de Oliveira Andrade
Centro de Estudos de Comunicação e Sociedade, Instituto de Ciências Sociais, Universidade do Minho, Braga, Portugal/ Programa de Pós-Graduação em Comunicação e Cultura Contemporâneas, Faculdade de Comunicação, Universidade Federal da Bahia, Salvador, Brazil

Abstract

This article investigates the materiality of Instagram's hate governance, reflecting on the role of this sociotechnical architecture in conflicts surrounding immigration in Portugal. The study is situated within a context of transformations in migration governance, the rise of the far right, and the centrality of digital platforms in struggles over visibility and participation. Its primary objective is to examine how the definition and regulation of anti-immigrant hate are organised through the platform's infrastructure. To this end, the study adopts the neomaterialist methodology of asocial communication (Lemos, 2020) and conducts a material-discursive analysis (Barad, 2007; Kitchin & Dodge, 2011). The empirical focus is on Meta's community standards, particularly the sections "Hateful conduct", "Violence and incitement", and "Bullying and harassment", as well as the reporting interfaces available to Instagram users. The analysis identifies the formal recognition of "refugees, migrants, immigrants, and asylum seekers" as a protected group against severe attacks. However, it also uncovers hierarchies of harm that prioritise immediate physical violence over political, economic, and social exclusion. Additionally, the study highlights areas of ambiguity arising from criteria related to intent, context, satire, and the platform's internal classification of reports. The article argues that Instagram's hate governance operates as a sociotechnical mediator, offering protection while shifting responsibility for safety onto users. This approach renders anti-immigrant hate only partially visible in policy and minimally discernible through the interface. The conclusion discusses the implications of these findings for the visibility, safety, and contestation experienced by immigrants within digital environments.

Keywords
platform governance, hate speech, immigration in Portugal, Instagram, communication conflict

Materialidade da Governação do Ódio no Instagram: Conflitos em Torno da Imigração em Portugal

Resumo

Este artigo investiga a materialidade da governação do ódio no Instagram, refletindo sobre o papel desta arquitetura sociotécnica nos conflitos em torno da imigração em Portugal. O estudo insere-se num cenário de transformações na governação migratória, ascensão da extrema-direita e centralidade das plataformas digitais na disputa por visibilidade e participação. O objetivo central é interrogar como a definição e o controlo do ódio anti-imigrante são infraestruturalmente organizados pela plataforma. Para tal, adota-se a metodologia neomaterialista da comunicação associal (Lemos, 2020) e realiza-se uma análise material-discursiva (Barad, 2007; Kitchin & Dodge, 2011). O foco empírico é dado aos padrões da comunidade da Meta, especialmente nas

secções "Conduta de ódio", "Violência e incitação" e "Bullying e assédio", assim como nas interfaces de denúncia disponíveis aos utilizadores do Instagram. A análise demonstra, por um lado, o reconhecimento formal de "refugiados, migrantes, imigrantes e pessoas que procurem asilo" como grupo protegido contra ataques graves. Por outro lado, o estudo evidencia hierarquias de dano que privilegiam a violência física imediata em detrimento de formas de exclusão política, económica e social, bem como zonas de ambiguidade produzidas por critérios de intenção, contexto, sátira e classificação interna das denúncias. Argumenta-se que a governação do ódio no Instagram funciona como mediador sociotécnico que simultaneamente promete proteção e redistribui responsabilidades de segurança pelos utilizadores, tornando o ódio anti-imigrante parcialmente visível nas políticas e pouco legível na interface. O artigo conclui discutindo as implicações desses achados para as condições de visibilidade, segurança e contestação de pessoas imigrantes no espaço digital.

Palavras-chave
governação de plataformas, discurso de ódio, imigração em Portugal, Instagram, conflito comunicacional

1. Introduction

During the first decades of the 21st century, increasing human mobility has been accompanied by an intensification of controversies surrounding migration. Economic and political crises, structural inequalities, armed conflicts, and climate emergencies have contributed to the growing number of both forced and voluntary movements, while States and institutions contest competing narratives concerning borders, security, and rights.

In the Portuguese context, these dynamics are embedded in successive transformations in migration governance, marked by the dissolution of the Immigration and Borders Service and the establishment of the Agency for Integration, Migration and Asylum in 2023, which assumed responsibility for the administration of migration affairs in Portugal. Since the Agency for Integration, Migration and Asylum began operating, reports of delays, communication failures, and difficulties in accessing rights have multiplied, prompting interventions by bodies such as the Ombudsperson, the Portuguese Bar Association, and civil society organisations. At the same time, changes in government and successive revisions to migration legislation have taken place in a context marked by the electoral rise of the far right, whose agenda emphasises stricter immigration policies, anti-immigration rhetoric, and the association of immigration with issues such as crime and insecurity.

Against this backdrop, digital platforms have become central to how migration is imagined, planned, experienced, and narrated. Environments such as YouTube, Instagram, and TikTok simultaneously function as sources of information, sociability, and visibility for immigrants, where practical advice, everyday experiences, and narratives of new beginnings circulate. However, this visibility is permeated by communication-related conflicts shaped by hate and content moderation, which bear upon the right to exist, work, and speak as an immigrant in the digital public sphere.

Within the context of the "platformisation" of social life (van Dijck et al., 2018) and the Platformisation, Datafication, and Algorithmic Performativity (PDPA) regime (Lemos,

2021), these conflicts are inseparable from the sociotechnical infrastructures that organise their circulation. Moderation is not a secondary aspect but rather a constitutive element of platforms themselves (Gillespie, 2018). Thus, beyond examining what is said about immigration, attention must also be paid to the infrastructural materiality that shapes this circulation: the categories through which platforms define "hate", the reporting and filtering mechanisms they deploy, the flows of human and algorithmic moderation, the sanctions they apply, and the affordances that encourage, constrain, or redirect particular forms of use.

Against this theoretical backdrop, this article approaches conflict as a sociotechnical controversy in which human and non-human actors (users, algorithms, interfaces, moderation policies, and reporting mechanisms) co-produce forms of visibility and vulnerability (Latour, 1994, 2012). Accordingly, this article maps the materiality of hate governance on Instagram, with particular emphasis on the infrastructural mechanisms that shape content relating to immigration. The study adopts a neomaterialist methodology of asocial communication (Lemos, 2020) and conducts a material-discursive analysis (Barad, 2007; Kitchin & Dodge, 2011) to examine how Instagram configures the governance of anti-immigrant hate through an analysis of its community standards and the reporting mechanisms made available to users, while reflecting on the role of this governance in shaping immigration-related conflicts in Portugal.

In this article, Portugal serves as the empirical-political framing of the controversy rather than as the exclusive scale of Instagram's technical governance. The policies and interfaces analysed are corporate and transnational; however, they are enacted in locally situated conflicts at a time when immigration has become the subject of intense institutional, media, and party-political contestation in the country. It is precisely at the intersection of a global platform architecture and a nationally situated conflict that the ambiguities of anti-immigrant hate moderation become particularly pronounced in the Portuguese case.

The remainder of this article is organised into five sections. Section 2 presents the theoretical framework on conflict, migration, and platform governance. Section 3 outlines the methodological approach and the empirical scope of the study. Section 4 analyses Meta's policies and the reporting process on Instagram. Section 5 brings the findings together and discusses their regulatory implications. Section 6 presents the final considerations.

2. Communication and Conflicts in Digital Platform Governance

2.1. Platformisation, Platform Governance, and Algorithmic Content Moderation

The notion of "platformisation" (van Dijck et al., 2018) and the PDPA regime (Lemos, 2021) help explain how digital platforms have assumed a structuring role in the organisation

of social and economic sectors, functioning as infrastructures that aggregate data, intermediation, and interactions on a large scale. This process imposes its own logics of classification and monetisation, translated into specific forms of content governance, whereby decisions regarding what is (or is not) made visible are determined by corporate rules and algorithmic systems. In this sense, platforms such as Instagram, TikTok, YouTube, and X exercise infrastructural power that shapes the circulation of information and the conditions of digital citizenship (Plantin & Punathambekar, 2019).

Content moderation is no longer an ancillary service but has become part of the "product" that platforms offer, alongside the promise of a safe and orderly environment (Gillespie, 2018, 2020). This form of sociotechnical governance involves design choices and policies that define, for example, categories of "hate", guidance frameworks, the sanctions applied, and the languages and cultural contexts prioritised (Birhane, 2023; Birhane & Cummins, 2019).

Within this framework, algorithmic content moderation has become a central axis of platform governance, combining automated filters, human moderation teams, and user reporting within a hybrid and opaque governance regime (Douek, 2022; Gillespie, 2018, 2020; Roberts, 2019). Regulatory pressure, such as that exerted by the European Union's Digital Services Act (União Europeia, 2022), coexists with commercial imperatives to maximise attention and reduce costs, producing compromises over what is classified as "hate" and when intervention occurs. In the case examined here, this means that Instagram's definition, detection, and sanctioning of anti-immigrant hate result from situated infrastructural decisions embedded in the platform's normative and operational materialities.

2.2. Communication-Related Conflicts, Hate, and Migration in Digital Culture

Conflicts surrounding immigration extend beyond the boundaries of institutional politics, traditional media, and legal forums, increasingly manifesting themselves within digital culture. In this context, sociotechnical platforms function as crucial infrastructures in struggles over visibility, recognition, and power, reshaping markets, labour, sociability, and civic participation (Lemos, 2021; van Dijck et al., 2018).

The centrality of these infrastructures to transnational migration is manifested at multiple levels (Leurs & Ponzanesi, 2018): from searching for information on migration routes and administrative procedures to building affective communities and support networks, as well as producing content that transforms the migratory experience into a communicative commodity (Andrade, 2025). At the same time, these spaces are permeated by discourses that challenge the legitimacy of migrants' presence, associating it with issues such as crime, the abuse of social benefits, or cultural threat.

These discursive disputes frequently take the form of hostile statements directed at individuals or groups on the basis of their national origin, race, ethnicity, religion, or migration status and are understood here, in a broad sense, as forms of prejudice-driven

anti-immigrant hate that seek to incite, justify, or normalise exclusion, hostility, or violence (Smets et al., 2022; Waldron, 2012).

In these terms, communication-related conflict in the digital sphere surrounding immigration should be understood as a phenomenon that is simultaneously symbolic and material: it involves narratives, framings, and representations of who "deserves" to belong within a given territory, while also depending on sociotechnical infrastructures that shape what circulates, who it reaches, and with what intensity. This article is situated at the intersection of symbolic contestation and infrastructural governance.

2.3. Xenophobia and Anti-Immigrant Hate in the Portuguese Digital Space

In the European context, hate speech is defined by the Committee of Ministers of the Council of Europe as all forms of expression which "spread, incite, promote or justify racial hatred, xenophobia, antisemitism or other forms of hatred based on intolerance" (Keen & Georgescu, 2016, p. 11), including hostility against minorities, migrants and people of immigrant origin. It constitutes a form of harm that extends beyond individual offence insofar as it seeks to undermine the equal status and public dignity of targeted groups, thereby excluding them from full participation in society (Waldron, 2012). Across Europe, racism and xenophobia are among the main reported motivations for hate-motivated threats and attacks (Organization for Security and Co-operation in Europe – Office for Democratic Institutions and Human Rights, 2021), and the available data for Portugal indicate that immigrants are among the most frequent victims of these forms of violence (Comissão Europeia contra o Racismo e a Intolerância, 2025).

The report of *Racismo e Xenofobia em Portugal: A Normalização dos Discursos de Ódio no Espaço Público da Internet* (Racism and Xenophobia in Portugal: The Normalisation of Hate Speech in the Public Sphere of the Internet; 2022) demonstrates the normalisation of such expressions in comments on news articles and social media, showing how hate speech becomes part of the everyday media landscape. Narratives linking immigration to threat, crime, incivility, or the "abuse" of social benefits circulate across digital environments, frequently intertwined with racial and ethnic stereotypes.

Recent reports reinforce this diagnosis. The *#MigraMyths* project shows that hate directed at migrants has become recurrent in the Portuguese public debate, within a context of political polarisation and the increasing politicisation of migration, in which populist and extremist movements exploit themes such as "national identity", "insecurity", and "economic crisis" to produce misleading narratives targeting migrants and refugees (Costa, 2025). The sixth report of the European Commission against Racism and Intolerance (Comissão Europeia contra o Racismo e a Intolerância, 2025) identifies a marked increase in hate speech in Portugal, directed primarily at migrants, Roma communities, Black people, and LGBTQIA+ people, warning against its "trivialisation" under the guise of freedom of speech. At the same time, data from the Portuguese Association for Victim Support (Associação Portuguesa de Apoio à Vítima, 2025) indicate an increase in reports of discrimination and incitement to hatred and violence, including through the internet.

The Brazilian community has been a particular target of online hostility. News reports and analyses point to a "wave of hate" directed at Brazilians in Portugal, in which anti-immigration discourse and stigmatisation circulate through comments, memes, and social media posts, reinforcing perceptions of insecurity and non-belonging (Santos, 2023). These discourses intersect with the electoral rise of the far right and the activities of organised hate groups, which use digital platforms to disseminate anti-immigration propaganda and mobilise supporters.

Data from the report published by the Casa do Brasil de Lisboa (Costa, 2025) help to clarify the role of platforms in this context: 79.8% of the immigrants surveyed reported having experienced online hate speech in Portugal, particularly on Instagram (36.9%). For the purposes of this article, the concern is less with quantifying these episodes than with understanding how platform infrastructures shape the conditions of their emergence, visibility, and governance. As a hybrid environment for social interaction, self-promotion, and digital labour, Instagram is one of the spaces in which these dynamics unfold, whether through hostile content and comments or through algorithmically mediated reporting, content removal, and (de-)recommendation.

At this point, a theoretical and methodological framework is required that can treat community standards, categories of "hate", reporting interfaces, and moderation tools as actants within the controversy (Latour, 1994, 2012; Lemos, 2020) surrounding immigration. Such a framework is presented in the following section.

3. Theoretical and Methodological Principles and Procedures

3.1. Exploratory Focus: Instagram

The empirical focus of this study is Instagram, as it is a central platform in the digital everyday lives of immigrants in Portugal, serving as a space for information, sociability, public visibility, and, in many cases, a source of income. As noted above, recent surveys on hate speech targeting immigrants in Portugal identify Instagram as the primary online platform where such incidents occur (Costa, 2025), reinforcing the relevance of this empirical focus.

Instagram also constitutes a paradigmatic case of the platformisation of social life (van Dijck et al., 2018) and the PDPA regime (Lemos, 2021), combining short-form audiovisual content, algorithmic recommendation systems, and engagement metrics that shape user behaviour. Moreover, the platform has become the subject of increasing regulatory and public scrutiny, making it particularly pertinent to examine the materiality of its hate governance mechanisms.

This empirical focus takes the form of a material-discursive analysis (Barad, 2007; Kitchin & Dodge, 2011) of a delimited set of Instagram's infrastructural elements: Meta's community standards (with particular attention to the sections "Hateful conduct", "Violence and incitement", and "Bullying and harassment"), the reporting interfaces available within the application, and relevant pages of the Help Centre describing the reporting, review, and user notification processes.

Although the analytical focus is on Instagram, the case is representative of broader disputes surrounding hate governance across platforms, in which different companies articulate, in different ways, commitments to regulation, economic interests, and competing understandings of freedom of expression and safety.

3.2. Methodology

The study adopts the neomaterialist methodology of asocial communication (Lemos, 2020), based on the premise that digital communication is a process of distributed mediation between human and non-human actors (Latour, 1994, 2012). Accordingly, community standards, interfaces, systems, and mechanisms are treated as actants that actively participate in the controversy surrounding anti-immigrant hate.

Operationally, the study follows the four stages of neomaterialist cartography (Lemos, 2020): (a) *mode* — delimiting the controversy in terms of how Instagram defines, classifies, and moderates anti-immigrant hate content, and how this governance shapes conflicts surrounding immigration in Portugal; (b) *inventory* — systematically identifying and describing the relevant infrastructural actants: normative documents and the reporting interfaces available to users; (c) *transduction* — analysing the mediations produced by these actants, that is, how categories, standards, and reporting workflows transform the possibilities for the circulation, visibility, and governance of hate content; (d) *reaggregation* — bringing the findings together into analytical axes that enable a discussion of hate governance as a process of sociotechnical mediation with specific effects on the conditions of safety, visibility, and contestation experienced by immigrants.

3.3. Procedures: Material-Discursive Analysis of Documents and Interfaces

The analytical procedures are grounded in material-discursive analysis (Barad, 2007; Kitchin & Dodge, 2011), which treats texts, images, code, and interfaces as agentic materialities capable of configuring modes of action and meaning.

The *inventory* and *transduction* stages are organised around two main groups of materials:

- Normative and guidance documents — Meta's community standards, with particular attention to the sections "Hateful conduct", "Violence and incitement", and "Bullying and harassment", as well as the Instagram Help Centre pages on privacy, reporting, and safety. These versions were selected because they correspond to the categories directly mobilised by the reporting workflow for hate, violence, and harassment-related content. The material was collected on 4 December 2025 from the Portuguese-language pages of the Meta Transparency Centre and the Instagram Help Centre. The analysis focused on hate categories and protected groups, illustrative examples, sanctions, and the guarantees of confidentiality, review, and appeal.
- Instagram reporting interfaces and support request workflows — using a user account, the reporting pathways for posts, comments, and accounts were documented through screenshots and field notes concerning the affordances, categories, and pathways available for reporting hate or abusive behaviour, as well as the informational messages presented to users. The screenshots included in

this article correspond to the reporting workflow for a post observed on 9 December 2025, using a Samsung mobile device running the Android operating system, with the mobile application configured in Portuguese and used in Portugal. As the objective was to examine the reporting workflow rather than evaluate any specific content, a random post from a public profile was selected. For ethical reasons, the report was not submitted, and the procedure was discontinued before the final submission stage. As interfaces may vary according to device, operating system, application version, region, language, and content type, these elements are treated as conditions of the materiality under examination.

4. Community Standards, Interfaces, and Reporting Workflows on Instagram

The policies governing Instagram are set out in Meta's community standards, a framework that applies across the company's technologies (Facebook, Instagram, Messenger, and Threads). According to Meta, these standards seek to balance "a place for expression" with safety by defining what is and is not permitted on its platforms, and they apply to all forms of content, including content generated by artificial intelligence (Meta, n.d.-a, n.d.-b, n.d.-c).

Each section is organised around a policy rationale, followed by specific policy guidelines that distinguish prohibited content from content requiring additional contextual assessment. For the purposes of this analysis, the sections "Hateful conduct", "Violence and incitement", and "Bullying and harassment" were selected because they constitute the core of Instagram's governance of hate, violence, and harassment, in conjunction with the practical guidance on reporting and tracking support requests.

Before proceeding with the analysis, it is useful to clarify the relationship between the terms employed in this article and Meta's own categories. "Hate" is used as a broad analytical category encompassing practices of hostility, dehumanisation, intimidation, or exclusion directed at vulnerable groups. "Hate speech" refers to its expressive forms, including statements, images, comments, or symbols that spread, incite, or justify hostility and discrimination. "Xenophobia" denotes the rejection or denigration of people perceived as foreign. "Anti-immigrant hate" refers to forms of hostility directed at individuals on the basis of their migration status, national origin, or perceived foreignness. These terms do not correspond directly to Meta's corporate categories. Rather, "Hateful conduct", "Violence and incitement", and "Bullying and harassment" are treated here as infrastructural forms of classification through which the platform translates, hierarchises, and renders certain forms of harm amenable to moderation.

From a neomaterialist perspective, these sections and interfaces are treated as mediators that delineate the boundaries of what may be said and done in relation to individuals and groups on the platform. The following subsections examine how these policies and infrastructural mechanisms configure protected categories, levels of severity, and areas of ambiguity that define the conditions of possibility for the circulation and governance of anti-immigrant hate on Instagram.

4.1. "Hateful Conduct": Definition, Protected Categories, and the Place of Migration Status

The "Hateful conduct" (formerly "Hate Speech") section of Meta's community standards (Meta, n.d.-b) begins by stating that people "use their voice and connect more freely when they don't feel attacked on the basis of who they are" (para. 1). Accordingly, Meta states that it does not allow hateful conduct on its platforms.

In the revision history, an update dated 7 January 2025 presents the text previously associated with "Hate speech" under the current heading "Hateful conduct". Although the revision history does not explicitly indicate a change of title, a comparison of the different versions suggests a terminological shift from an emphasis on "speech" to an emphasis on "conduct", indicating that the policy applies not only to linguistic expressions but also to broader forms of hostile behaviour. From the perspective of material-discursive practice (Barad, 2007), this change points to the processual and adaptive character of hate governance: the very definition of the phenomenon is itself subject to ongoing infrastructural adjustments.

In the version examined, Meta (n.d.-b) defines hateful conduct as "direct attacks against people — rather than concepts or institutions — on the basis of what we call protected characteristic(s)" (para. 2). These include race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease. Age is also treated as a protected characteristic when referenced along with another protected characteristic. The policy further states that Meta "also [protects] refugees, migrants, immigrants, and asylum seekers from the most severe attacks (Tier 1 below), though [it allows] commentary on and criticism of immigration policies" (Meta, n.d.-b, para. 2). By explicitly including "refugees, migrants, immigrants, and asylum seekers" within the scope of protection against severe attacks, the policy incorporates migration status into its internal classificatory framework, placing it alongside other characteristics considered particularly sensitive.

Content removal is organised according to levels of severity. Tier 1 covers the most severe forms of attack, including dehumanising speech (in the form of comparisons to animals, pathogens, or "sub-human life forms"), allegations of serious immorality or criminality, and calls and hopes for specific individuals to suffer harm (such as contracting a disease, dying, or experiencing a natural disaster), where the targets are defined by protected characteristics or immigration status. Tier 2 includes calls or support for exclusion or segregation (for example, denying a group political, economic, or social rights), as well as insults and expressions of contempt, disgust, or revulsion directed at people on the basis of the same characteristics.

From the perspective of protecting migrant populations, the distinction between Tier 1 and Tier 2 is particularly revealing. Whereas Tier 1 is reserved for direct threats and explicit dehumanisation, Tier 2 encompasses practices that, from legal and sociological perspectives, more closely resemble structural violence and systematic discrimination: political, economic, and social exclusion, encouragement of segregation, and repeated insults. Treating these as less severe than individual threats of physical harm privileges

direct physical violence over forms of exclusion that, in more diffuse and enduring ways, undermine the capacity of vulnerable groups — including immigrants — to participate fully in society.

The policy also acknowledges that not all content containing potentially offensive language should be removed automatically. It recognises that users may share insults or hate speech produced by others in order to condemn or expose them, or may use potentially offensive terms in self-referential or "empowering" ways, stating that such speech is allowed when "the speaker's intention is clear" (Meta, n.d.-b, para. 4), while noting that content may be removed where the intention is unclear. It further explains that, in certain political and religious contexts (such as debates concerning transgender rights, immigration, or homosexuality), the policies were "designed to allow room for these types of speech" (para. 5), as well as for some uses of satire, provided that the violating elements are themselves being criticised or ridiculed.

These provisions introduce a second layer of ambivalence. By allowing potentially offensive content to remain when it is interpreted as condemning hateful conduct, as self-referential use, or as humour, and by providing additional latitude for political and religious debate (including on immigration), the policy shifts part of the moderation decision to situated acts of interpretation. In practice, hostile statements directed at immigrants may be framed as legitimate opinions on "border control" or "security", thereby avoiding removal. The latitude afforded to satire and other ambiguous forms of expression further expands these areas of indeterminacy, particularly in contexts of intense political polarisation surrounding immigration, such as that observed in Portugal.

The "Hateful conduct" policy therefore functions as an infrastructural mediator that defines which identity attributes merit enhanced protection, hierarchises forms of violence, and establishes the boundaries and exceptions that simultaneously shape the visibility, vulnerability, and possibilities for contestation available to immigrants.

4.2. "Violence and Incitement": Offline Harm and Levels of Protection

The "Violence and incitement" policy aims to "prevent potential offline violence that may be related to content" (Meta, n.d.-c, para. 1) on Meta's platforms. It states that the company removes language that incites or facilitates violence, as well as credible threats to public or personal safety, including violent speech targeting a person or group of people "on the basis of their protected characteristic(s) or immigration status" (para. 1).

The section is organised around a set of "universal protections" (which apply to everyone regarding high- and medium-severity threats of violence) and "additional protections" granted to individuals considered more vulnerable, including children, people at heightened risk, and individuals or groups targeted on the basis of protected characteristic(s). Since attacks against immigrants may be directed at them on the basis of their national origin, ethnicity, religion, or migration status itself, the policy allows, at least in principle, for them to benefit from this enhanced level of protection.

In addition to direct threats, the section covers "other [types of] violence", including requests for or offers of violent services (for example, hiring assassins), instructions

on the use of weapons or explosives with the intention of causing serious harm, and threats to bring weapons to sensitive locations (educational facilities, places of worship, or polling places). In all these cases, the policy again recognises contextual exceptions (for example, situations involving self-defence or military or sports training), reinforcing the centrality of situated interpretation.

A substantial portion of the policy is devoted to situations in which the application of the standards requires "additional information and/or context", such as coded or implicit threats, references to historical incidents of violence, invitations to third parties to carry out violent acts, or suggestions that someone possesses sensitive personal information (such as a residential address, place of employment, or daily routines). Here too, intervention depends on the platform's assessment of the credibility and imminence of the harm.

This policy reinforces the primacy of the criterion of offline harm: what primarily justifies content removal is its potential to result in physical violence beyond the screen. In relation to the controversy surrounding immigration in Portugal, this suggests a dual effect. On the one hand, the reference to immigration status confirms the recognition of immigrants as a vulnerable group. On the other, by prioritising credible threats and immediate physical risk, the policy tends to undervalue intimidation and symbolic violence that contribute to a hostile environment and to the delegitimisation of migrant populations. The contextual exceptions and case-by-case assessments preserve areas in which hostile statements may circulate as political opinion, expressions of indignation, or figurative rhetoric.

4.3. "Bullying and Harassment": Personal Attacks and Mass Harassment

The "Bullying and harassment" policy shifts the focus from the protection of groups defined by protected characteristics to the protection of private individuals and public figures (Meta, n.d.-a). Meta (n.d.-a) states that it does not tolerate behaviour that "prevents people from feeling safe and respected" (para. 1) and distinguishes between public figures and private individuals, providing stronger safeguards for the latter, particularly minors.

The section establishes different levels of protection, covering behaviours ranging from repeated malicious contact, threats to disclose personal information (doxxing), encouragement of self-injury, degradation, and unwanted sexualisation, to attacks on physical appearance and coordinated mass harassment. Private minors, private adults, and minor involuntary public figures receive additional layers of protection, particularly when they report the content themselves.

In the context of immigration, this policy is relevant because it addresses the forms of attack most frequently directed at immigrants with a public profile: persistent hostile contact, repeated derogatory comments, and attempts to humiliate or intimidate through the disclosure of personal information. Although the section does not explicitly refer to "immigrants", it does mention "individuals at heightened risk of offline harm", including human rights defenders, ethnic and religious minorities in conflict zones, and victims of violent events/tragedies. As such, its scope may extend to immigrant content creators and

activists who become the targets of coordinated harassment campaigns because of the content they produce.

At the same time, the policy reinforces reliance on self-reporting: in many cases, platform intervention depends on the victim — or an authorised representative — indicating that the content is unwanted. For immigrants who work as content creators, this means that protection is contingent upon their ability to recognise the attack, report it, and manage the moderation tools available to them, thereby accumulating additional layers of work and emotional strain.

"Bullying and harassment" functions as a mediator that regulates the boundary between legitimate criticism, public exposure, and symbolic violence directed at specific individuals. Together with "Hateful conduct" and "Violence and incitement", it constitutes a normative-infrastructural core that, while promising enhanced protection, simultaneously produces limits and ambivalences regarding what counts as an unacceptable attack, with direct implications for immigrants' experiences of safety and the sustainability of their digital work on Instagram and other Meta platforms.

4.4. Reporting Workflows and Support Requests on Instagram: Between Guidance and Opacity

From the user's perspective, the documented workflow for reporting a post begins by selecting the three-dot icon associated with the content in the feed, reels, or stories. The first screen (Figure 1) presents options such as "captions", "why you're seeing this post", "not interested", "interested", and "about this account", followed by a link to "content preferences". The "report" option, highlighted in red, marks the transition from a regime of *feed* personalisation to one of content governance: once selected, a subjective experience of hate or violence is translated into the infrastructural categories recognised by the platform.

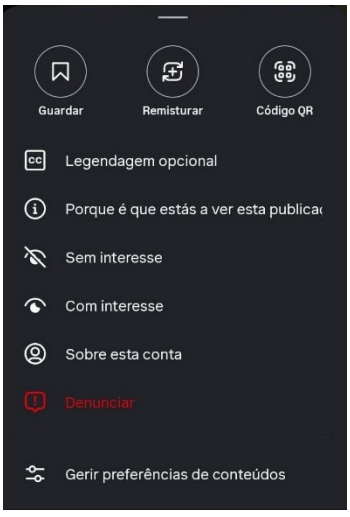

Figure 1. *First screen of the reporting workflow*

Note. Screenshot, 9 December 2025.

On the next screen (Figure 2), the user is asked, "why are you reporting this post?", accompanied by the statement that "your report is anonymous" and an instruction to contact the emergency services in situations of immediate danger. A set of reporting categories is then displayed, including "I just don't like it", "bullying or unwanted contact", "suicide, self-injury or eating disorders", "violence, hate or exploitation", "selling or promoting restricted items", "nudity or sexual activity", "scam, fraud or spam", "false information", "intellectual property", and "report as unlawful".

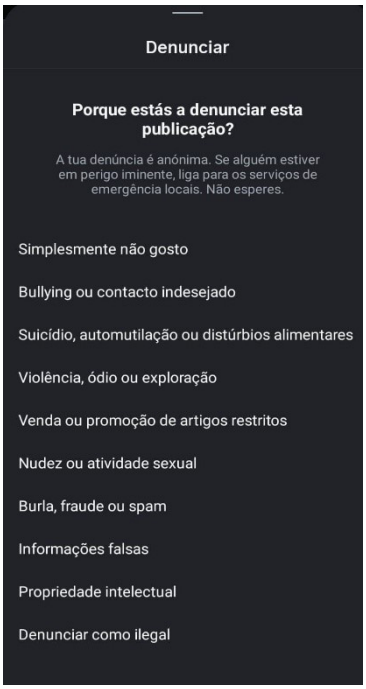

Figure 2. *Initial reporting categories*

Note. Screenshot, 9 December 2025.

When the user selects "violence, hate or exploitation", a further screen (Figure 3) requests greater specificity by asking, "how does this involve violence, hate or exploitation?", offering subcategories such as "credible threat to safety", "seems like terrorism or organised crime", "seems like exploitation", "hate speech or symbols", "calling for violence", "showing violence, death or severe injury", and "animal abuse".

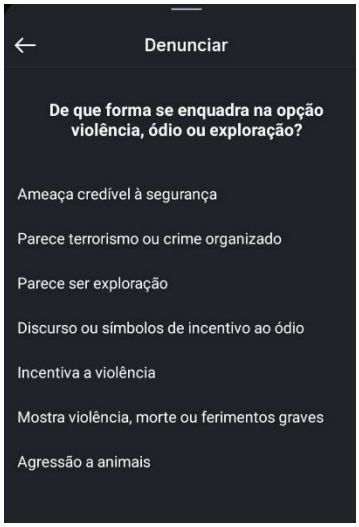

Figure 3. *Subcategories under "violence, hate or exploitation" in the reporting workflow*

Note. Screenshot, 9 December 2025.

On the following screen (Figure 4), Instagram presents a summary of the report: "you're about to submit a report. We only remove content that goes against our Community Standards", together with a direct link to that document, followed by the fields "why are you reporting this post?" and "how does this involve violence, hate or exploitation?", now populated with the selected options ("violence, hate or exploitation"/ "hate speech or symbols"), and the "submit" button. This final stage explicitly embeds the reporting workflow within the normative architecture analysed in the previous sections: the platform emphasises that not all offensive content will be removed, but only content that falls within its own definitions of a policy violation.

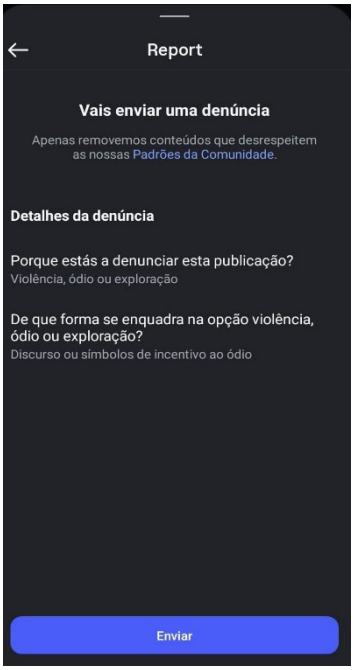

Figure 4. *Confirmation screen before submitting the report*

Note. Screenshot, 9 December 2025.

This organisation reinforces the interpretation advanced in the policy analysis. Although Meta recognises "refugees, migrants, immigrants, and asylum seekers" as protected targets in its community standards, the category of "migrant" does not appear explicitly in the reporting interface. Users are instead required to locate their experience of anti-immigrant hate within the generic category of "hateful speech or symbols". In terms of transduction (Lemos, 2020), situated experiences of xenophobia are translated into an abstract normative grammar without becoming visible at the interface level.

Accordingly, the reporting workflow highlights three central aspects of hate governance on Instagram:

- Redistribution of classificatory responsibility — the interface delegates to the user the task of translating lived experience into the platform's internal system of categories, thereby performing classificatory work that feeds both the moderation process and the datafication of conflict.
- Displacement between migrant experience and normative grammar — migration status is recognised within the textual framework of the policies but does not become directly visible within the reporting workflow, which treats anti-immigrant hate as a generic instance of "hate speech".
- Production of zones of ambiguity — the coexistence of options such as "I just don't like it" and "Violence, hate or exploitation", together with the need to distinguish a "credible" threat from figurative language, may shift experiences of anti-immigrant hate into the realm of disagreement, political opinion, or satire.

The Help Centre (Centro de Ajuda, n.d.) guidance reinforces and complements this representation. Within the "Privacy and reporting" section, guidance on reporting harassment, bullying, or abusive content explains that accounts, comments, and posts can be reported directly within the application using the integrated reporting tool. Users are also encouraged to block or restrict the reported account, and the Help Centre reiterates that reports are anonymous and that no information about the reporting user is shared with the reported individual.

The Help Centre (Centro de Ajuda, n.d.) also explains that users can check the status of reported content through the "Support requests" area, accessible via the settings menu. Users may review some previous reports, determine whether the platform has taken action, read additional explanations regarding the policies applied, and, in certain cases, request a further review of the decision. At the same time, it is made explicit that not all reports appear under "Support requests" and that the option to appeal is not always available, introducing a degree of opacity regarding how the platform prioritises and communicates particular cases.

Read together with the "Hateful conduct", "Violence and incitement", and "Bullying and harassment" policies, this workflow illustrates a dual dynamic. On the one hand, there is an investment in guidance: the interface simplifies classification, provides assurances of anonymity, directs users to safety resources, and suggests complementary actions, such as blocking or restricting accounts. Responsibility for activating these protective mechanisms is thus redistributed to the user, who must recognise the harmful situation, navigate the appropriate categories, monitor the status of the report, and, where applicable, request a review. On the other hand, areas of opacity remain. Users are not given access to the specific criteria that determine whether content reported for hate

or violence will be removed, retained, or merely subjected to reduced visibility. Nor do they know why some reports appear under "Support requests" while others do not, or under what conditions an appeal is available.

In contexts of polarisation surrounding immigration, such as that found in Portugal, this combination of guidance and opacity configures, in situated ways, the conditions of safety, visibility, and contestation experienced by immigrants on Instagram, mediating how experiences of anti-immigrant hate are — or are not — recognised, translated, or reclassified by the platform.

5. Hate Governance, Immigration and Platforms

5.1. Axes of Reaggregation of the Controversy

Drawing on the methodological trajectory of asocial communication (Lemos, 2020), *reaggregation* entails revisiting the path traced through the stages of *mode*, *inventory*, and *transduction* in order to propose an interpretative synthesis of the controversy. In the case examined here, three analytical axes summarise the relationship between hate governance, immigration, and digital platforms.

The first axis concerns the partial recognition of migrant status. At the normative level, the "Hateful conduct" section explicitly includes "refugees, migrants, immigrants, and asylum seekers" among the protected targets against the most severe attacks, aligning migrant status with other protected characteristics, such as race, ethnicity, nationality, and religion.

However, this recognition remains largely confined to the textual layer of the policies. Within the reporting interface, experiences of anti-immigrant hate are subsumed under generic categories such as "violence, hate or exploitation" or "hate speech or symbols", without any explicit reference to national origin or migrant status. Migrant status, which is central to the Portuguese sociopolitical controversy, is thus visible within the platform's normative grammar but absent from the operational surface of the application.

The second axis concerns hierarchies of harm. The tiered structure of the "Hateful conduct" policy associates Tier 1 with dehumanising attacks, allegations of serious criminality, and calls and hopes for individuals to suffer physical harm. In contrast, Tier 2 encompasses advocacy of political, economic, and social exclusion, segregation, and repeated insults.

From the perspective of the migrant experience, this hierarchy is problematic because practices of exclusion, stigmatisation, and systematic delegitimisation are formally classified as less severe than direct threats of physical harm, even though they may contribute to the production of fear, withdrawal, and the precarisation of public participation. Similarly, the "Violence and incitement" policy prioritises the criterion of a credible risk of offline harm, tending to undervalue forms of symbolic intimidation that do not immediately translate into physical aggression but nonetheless form part of the hostile environment experienced by immigrants.

The third axis concerns the asymmetrical redistribution of responsibility. Both the policies and the reporting workflow shift the task of recognising hate, translating it into the platform's internal categories, and activating the available protective mechanisms onto the user. At the same time, Meta reserves for itself the final decision as to whether content does or does not violate its community standards, thereby introducing opacity regarding the specific criteria underpinning moderation decisions, the prioritisation of cases, and the availability of appeals.

When immigration becomes a focal point of intense political and affective polarisation, this combination of guidance and ambiguity creates openings through which hostile discourse targeting immigrants may be reclassified as political opinion, satire, or figurative rhetoric, thereby avoiding removal even when it contributes to the normalisation of anti-immigrant hate.

Viewed through the lens of the PDPA regime (Lemos, 2021), these findings acquire an additional analytical dimension. The policies and interfaces examined not only define what may be said but also structure the transformation of reports into data, feeding recommendation systems, automated detection mechanisms, and integrity metrics that modulate the visibility of content and accounts. Hate governance in relation to immigration thus constitutes a process of sociotechnical mediation in which normative categories, reporting workflows, and algorithmic infrastructures become intertwined in producing different degrees of exposure, vulnerability, and capacity for contestation for immigrants.

5.2. Regulatory Implications in the European Context

These findings also have implications for the regulatory interpretation of content moderation in the European context. The Digital Services Act (União Europeia, 2022) does not substantively define what should constitute anti-immigrant hate in each national context; rather, it places platforms under a regime of enhanced due diligence, transparency, and contestability. Among other requirements, it mandates accessible notice-and-action mechanisms, clear information regarding moderation decisions, internal complaint-handling systems, and, for very large online platforms, the assessment and mitigation of systemic risks associated with the dissemination of illegal content and its effects on fundamental rights.

In light of these requirements, the protection of immigrants does not depend solely on the formal existence of a protected category. It also depends on the legibility of that category within the interface, the extent to which the reporting process can capture contextual dimensions such as national origin and migrant status, the provision of explanations for moderation decisions, and the availability of effective avenues of redress when the platform leaves hostile content online. The regulatory question is therefore no longer limited to whether an individual piece of content violates an internal rule; it also encompasses whether the reporting and moderation architecture recognises recurrent sociotechnical harms, assesses context-specific risks, and responds proportionately to the vulnerabilities of the affected groups.

6. Final Considerations

Starting from the Portuguese context, characterised by changes in migration governance, institutional delays, and anti-immigration political forces, this article has investigated the materiality of hate governance on Instagram, focusing on Meta's community standards and the reporting workflow available to users. Adopting the neomaterialist methodology of asocial communication (Lemos, 2020), it conducted a material-discursive analysis (Barad, 2007; Kitchin & Dodge, 2011), treating policies, interfaces, and moderation categories as actants that actively participate in the controversy surrounding anti-immigrant hate. The stages of *mode*, *inventory*, *transduction*, and *reaggregation* enabled an analysis of the normative and operational infrastructures that frame this conflict.

Empirically, the findings indicate that, although Meta formally recognises refugees, migrants, immigrants, and asylum seekers as protected targets, this recognition is constrained by three main dynamics: (a) the invisibility of migrant status within the reporting interface, which requires anti-immigrant hate to be subsumed under generic categories; (b) the hierarchy of harm, which treats forms of political, economic, and social exclusion that are central to anti-immigrant hate as less severe; and (c) the combination of delegated responsibility placed on users and decisional opacity, which allows hostile content to remain online under the guise of political opinion or satire. Taken together, these elements configure a regime of protection that is both necessary and limited, in which immigrants are encouraged to place their trust in the platform while remaining exposed to environments characterised by uncertainty and asymmetries of power.

For the Portuguese context, these findings extend beyond the infrastructural dimension. In a context in which empirical evidence points to the centrality of platforms — and Instagram in particular — in the everyday experience of hate directed at immigrants, how the platform defines, hierarchises, and operationalises "hate" forms part, albeit opaquely, of the broader ecology of migration governance. This observation also raises questions for the European regulatory framework, since the protection of vulnerable groups depends not only on the removal of illegal content but also on the transparency, contestability, and legibility of moderation systems.

This article represents an initial contribution to the study of the controversy surrounding anti-immigrant hate in Portugal within the context of digital platforms, focusing on Instagram and on a specific corpus of materials (policies and the reporting workflow). Future research could broaden this focus to include other platforms, examining these infrastructural tensions through approaches such as systematic content analysis, online observation/digital ethnography, and direct engagement with immigrants, thereby deepening our understanding of the relationships, assemblages, and affects between human and non-human actants in the constitution of the controversy (Fox & Alldred, 2022; Latour, 1994, 2012; Lemos, 2020).

By demonstrating that hate governance is not merely a matter of "content moderation", but rather a process of sociotechnical mediation (Latour, 1994, 2012; Lemos, 2020) involving norms, interfaces, algorithms, and other actants, this article contributes to debates on communication and immigration-related conflicts. It proposes a research

agenda centred on the materiality of platforms and its implications for the dissemination of anti-immigrant hate practices, as well as for the protection and participation of immigrants in the digital public sphere.

Machine Translation Post-Editing: Anabela Delgado

Acknowledgements

This work is funded under the multiannual funding of the Communication and Society Research Centre 2025–2029 (reference UID/00736/2025), by the Foundation for Science and Technology.

References

Andrade, D. (2025). *YouTubers e migração: A dinâmica comunicacional sociotécnica do trabalho digital de imigrantes brasileiros em Portugal* [Doctoral dissertation, Universidade do Minho]. RepositóriUM. https://hdl.handle.net/1822/98534

Associação Portuguesa de Apoio à Vítima. (2025). *Estatísticas APAV: Relatório anual 2024*. APAV.

Barad, K. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Duke University Press.

Birhane, A. (2023). Algorithmic colonization of Africa. In S. Cave & K. Dihal (Eds.), *Imagining AI: How the world sees intelligent machines* (pp. 247–260). Oxford University Press. https://doi.org/10.1093/oso/9780192865366.003.0016

Birhane, A., & Cummins, F. (2019). *Algorithmic injustices: Towards a relational ethics*. arXiv. https://doi.org/10.48550/arXiv.1912.07376

Centro de Ajuda. (n.d.). *Como denunciar algo no Instagram*. Retrieved 4 December 2025, from https://help.instagram.com/2922067214679225/

Comissão Europeia contra o Racismo e a Intolerância. (2025). *Relatório da ECRI sobre Portugal: Sexto ciclo de monitorização*. Conselho da Europa. https://rm.coe.int/sixth-report-on-portugal-translation-in-portuguese-/1680b6668f

Costa, A. P. (2025). *Discurso de ódio e imigração em Portugal*: *Relatório #MigraMyths – Desmistificando a imigração*. Casa do Brasil de Lisboa; #migramyths.

Douek, E. (2022). Content moderation as systems thinking. *Harvard Law Review*, *136*(2), 58–144.

Fox, N., & Alldred, P. (2022). New materialism. In P. Atkinson, S. Delamont, A. Cernat, J. W. Sakshaug, & R. A. Williams (Eds.), *SAGE research methods foundations* (pp. 1–16). SAGE. https://doi.org/10.4135/9781526421036768465

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, *7*(2), 1–5. https://doi.org/10.1177/2053951720943234

Keen, E., & Georgescu, M. (2016). *Manual para o combate do discurso de ódio online através da educação para os direitos humanos: Edição revista com a inclusão do guia dos direitos humanos para os utilizadores da internet*. Conselho da Europa; Fundação Calouste Gulbenkian.

Kitchin, R., & Dodge, M. (2011). *Code/space: Software and everyday life*. The MIT Press.

Latour, B. (1994). On technical mediation - Philosophy, sociology, genealogy. *Common Knowledge*, *3*(2), 29–64.

Latour, B. (2012). *Reagregando o social: Uma introdução à teoria do ator-rede*. EDUFBA; EDUSC.

Lemos, A. (2020). Epistemologia da comunicação, neomaterialismo e cultura digital. *Galáxia*, (43), 54-66. https://doi.org/10.1590/1982-25532020143970

Lemos, A. (2021). *A tecnologia é um vírus: Pandemia e cultura digital*. Editora Sulina.

Leurs, K., & Ponzanesi, S. (2018). Connected migrants: Encapsulation and cosmopolitanization. *Popular Communication*, *16*(1), 4–20.

Meta. (n. d.-a). *Padrões da comunidade: Bullying e assédio*. Meta Transparency Center. Retrieved 4 December 2025, from https://transparency.meta.com/policies/community-standards/bullying-harassment/

Meta. (n.d.-b). *Padrões da comunidade: Conduta de ódio*. Meta Transparency Center. Retrieved 4 December 2025, from https://transparency.meta.com/policies/community-standards/hate-speech/

Meta. (n.d.-c). *Padrões da comunidade: Violência e incitação*. Meta Transparency Center. Retrieved 4 December 2025, from https://transparency.meta.com/policies/community-standards/violence-incitement/

Organization for Security and Co-operation in Europe – Office for Democratic Institutions and Human Rights. (2021). *2020 hate crime data: Key findings*. https://hatecrime.osce.org/sites/default/files/2021-11/2020%20Infographic_EN.pdf

Plantin, J.-C., & Punathambekar, A. (2019). Digital media infrastructures: Pipes, platforms, and politics. *Media, Culture & Society*, *41*(2), 163–175. https://doi.org/10.1177/0163443718818376

Racismo e xenofobia em Portugal: A normalização dos discursos de ódio no espaço público da internet. (2022). Centro de Rede de Investigação em Antropologia. https://racismoexenofobia.cria.org.pt/wp-content/uploads/2022/07/Relato%CC%81rio_Racismo_Xenofobia.pdf

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Santos, J. (2023, December 17). Onda de ódio contra brasileiros em Portugal. *Visão*. https://visao.pt/atualidade/sociedade/2023-12-17-a-onda-de-odio-contra-os-brasileiros-em-portugal/

Smets, K., Toffano, G., & Almenara Niebla, S. (2022). Refugee and mediated lives. In M. Powers (Ed.), *Oxford research encyclopedia of communication* (pp. 1–22). Oxford University Press. https://doi.org/10.1093/acrefore/9780190228613.013.1263

van Dijck, J., Poell, T., & de Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press.

União Europeia. (2022). *Regulamento (UE) 2022/2065 de 19 de outubro de 2022 relativo a um mercado único de serviços digitais (Regulamento dos Serviços Digitais)*. https://eur-lex.europa.eu/legal-content/PT/TXT/HTML/?uri=CELEX:32022R2065

Waldron, J. (2012). *The harm in hate speech*. Harvard University Press.

Biographical Note

Dalvacir Xavier de Oliveira Andrade holds a PhD in Communication Sciences from the University of Minho. Her doctoral thesis examined the sociotechnical communicational dynamics of the digital labour of Brazilian YouTubers in Portugal. Her research focuses on the infrastructural mediations of digital platforms, sociotechnical controversies, migration and digital labour within platform ecosystems. She collaborates with the Communication and Society Research Centre, University of Minho, and with Lab404 - Research Laboratory in Digital Media, Networks and Space, at the Federal University of Bahia, where she contributes to projects on digital culture, platformisation, datafication and algorithmic performativity. Her research interests also include neomaterialist methodologies, digital ethnographies, and critical studies of labour precarity and inequalities in digital environments.

ORCID: https://orcid.org/0000-0001-8886-4038

Email: dalvacirx@gmail.com

Address: Lab404 – Laboratório de Pesquisa em Mídia Digital, Redes e Espaço, Programa de Pós-Graduação em Comunicação e Cultura Contemporâneas, Faculdade de Comunicação, Universidade Federal da Bahia, Rua Barão de Jeremoabo, s/n, Campus Universitário de Ondina, 40170-115 Salvador – BA, Brasil

Submitted: 30/12/2025 | Accepted: 19/06/2026