Meta-Identities and Political Extremism in Digital Networks: Algorithmic Classification, Radicalization and Democratic Disputes

Allan Herison Ferreira, Ana Carolina Trevisan, Camila Sayuri Shirakura

https://doi.org/10.1590/SciELOPreprints.18139

Submitted on: 2026-09-22
Posted on: 2026-09-22 (version 1)
(YYYY-MM-DD)

META-IDENTITIES AND POLITICAL EXTREMISM IN DIGITAL NETWORKS: ALGORITHMIC CLASSIFICATION, RADICALIZATION AND DEMOCRATIC DISPUTES

ALLAN HERISON FERREIRA
ORCID: https://orcid.org/0000-0003-3606-2089
allanherison@gmail.com
Universidade NOVA de Lisboa, Institute of Communication (ICNOVA); Social Research Laboratory (LAPS/USP). Lisbon, Portugal

ANA CAROLINA TREVISAN
ORCID: https://orcid.org/0000-0002-2818-5858
anacarolinatcf@gmail.com
Universidade NOVA de Lisboa, Institute of Philosophy (IFILNOVA); Social Research Laboratory (LAPS/USP). Lisbon, Portugal

CAMILA SAYURI SHIRAKURA
ORCID: https://orcid.org/0009-0002-8291-2945
camila.shirakura@gmail.com
Universidade Estadual de Maringá (UEM), Center for Research on Political Participation (Nuppol); TPDS-UEM. Maringá, Brazil

ABSTRACT: This article examines how the literature documents classificatory processes produced by digital platforms and artificial intelligence systems about subjects, content, and collectives involved in radicalized political disputes. Meta-identity is understood as a sociotechnical classificatory construct — inferred, aggregated, opaque and operational — produced about entities circulating within digital infrastructures, rather than by them, and mobilized as a reference for recommendation, moderation, visibility, segmentation and suspicion. The central question is how the literature documents classifications that may favor, limit or reorganize the circulation of extremist and anti-democratic repertoires. The procedure is a structured evidence synthesis: 813 records retrieved from open bibliographic sources between 2016 and 2026, consolidated into 749 studies, of which 127 had at least one full text retrieved; in 118 of them, rule-based, auditable lexical localization identified at least one candidate. The search did not look for the concept of meta-identity, but for the phenomena it analyzes and describes, using the vocabulary of the original literature. Preliminary results indicate that the literature readily documents the inference and stabilization of classifications, less frequently their attribution to determinate bearers, and almost never exteriority — the condition whereby the classified party lacks substantive control over the production and revision of the predicate. Adversarial activation of input dynamics by organized political actors appears in a small number of studies. We conclude that democratic dispute involves not only the regulation of extremist messages, but the

contestation of the classificatory regimes that define what will be seen, recommended, moderated or rendered suspect.

Keywords: meta-identity, algorithmic classification, political extremism, content moderation, contestability

META-IDENTIDADES E EXTREMISMO POLÍTICO NAS REDES DIGITAIS: CLASSIFICAÇÃO ALGORÍTMICA, RADICALIZAÇÃO E DISPUTAS DEMOCRÁTICAS

RESUMO: O artigo investiga como a literatura documenta processos classificatórios produzidos por plataformas digitais e sistemas de inteligência artificial sobre sujeitos, conteúdos e coletivos envolvidos em disputas políticas radicalizadas. Entende-se meta-identidade como um constructo classificatório sociotécnico — inferido, agregado, opaco e operacional — produzido sobre entidades que circulam em infraestruturas digitais, e não por elas, e mobilizado como referência para recomendação, moderação, visibilidade, segmentação e suspeição. A questão central é de que modo a literatura documenta classificações que podem favorecer, limitar ou reorganizar a circulação de repertórios extremistas e antidemocráticos. A metodologia consiste em uma síntese estruturada de evidências: 813 registros recuperados em bases abertas entre 2016 e 2026, consolidados em 749 estudos, dos quais 127 tiveram ao menos um texto integral recuperado; em 118 deles, a localização assistida por regras léxicas auditáveis identificou ao menos um candidato. A busca não se concentrou no conceito de meta-identidade, mas nos fenômenos que ele analisa e descreve. Os resultados preliminares indicam que a literatura documenta com maior frequência a inferência e a estabilização das classificações operadas nas infraestruturas analisadas, com menor frequência a atribuição a portadores determinados, e raramente a exterioridade — condição segundo a qual o classificado não controla integralmente a produção nem a revisão do predicado. A ação deliberada de atores políticos organizados sobre os vetores de entrada é abordada em um número reduzido de estudos. Conclui-se que a disputa democrática envolve não apenas a regulação de mensagens extremistas, mas a contestação dos regimes classificatórios que definem o que será visto, recomendado, moderado ou tornado suspeito.

Palavras-chave: meta-identidade, classificação algorítmica, extremismo político, moderação de conteúdo, contestabilidade

INTRODUCTION

In November 2024, the Brazilian Francisco Wanderley Luiz, known as *Tiu França*, detonated explosive devices in front of the Supreme Federal Court in Brasília and died at the scene. In the days leading up to the act, he had been posting steadily on social media, announcing what he intended to do and suggesting premeditation (MANFRIN, 2024; FERREIRA; TREVISAN, 2025). The following day, Justice Minister Alexandre de Moraes linked the attack to a broader context that, in his assessment, went back to the so-called "*gabinete do ódio*" or "office of hate" (GRANDI, 2024). The episode condenses the problem addressed in this article. The available explanation oscillates between two equally insufficient poles: on one side, the attribution of the act to individual convictions; on the other, the diffuse imputation to "algorithms" that supposedly drew him into ever deeper engagement with radicalized content. Neither one, taken alone, describes the mechanisms that may mediate the passage from discourse to act.

This insufficiency is grounded in radicalization theory. The two-pyramids model (MCCAULEY; MOSKALENKO, 2017) distinguishes radicalization of opinion from radicalization of action and holds, with consolidated empirical support, that the two do not coincide: most of those who hold radical ideas never act, and some of those who act do not share particularly radical convictions. It is precisely in this distance between opinion and action — which the model shows is not automatic — that analytical room opens up for investigating additional mediating mechanisms. Among them, this article renders analyzable the classificatory layer produced by digital infrastructures. It is important to stress the caution that the concept of radicalization itself carries securitarian presuppositions and tends to individualize structural processes (KUNDNANI, 2015), which recommends treating it as an analytical category rather than as a diagnosis.

Between these poles lies a layer that is rarely named: the infrastructure that classifies subjects, content and collectives, and that decides, on the basis of that classification, what each of them will see, to whom they will be recommended, what will be moderated and who will be treated as suspect. It is this layer that the concept of meta-identity seeks to render analyzable.

The argument developed here is that the online circulation of extremist repertoires cannot be analyzed solely through exposure to extreme content, but also through the classificatory ecosystems that stabilize affinities, audiences, engagement signals and risk categories. The question that organizes the text is: how does the literature document classifications that may favor, limit, or reorganize the circulation of extremist, authoritarian, and anti-democratic repertoires?

Three clarifications delimit the scope of the proposal. First, it is not claimed that recommender systems cause radicalization. The displacement proposed is a different one: from the question about the effect of the algorithm to the question about the classificatory construct that is attributed and about the differential treatment it authorizes. Second, the responsibility discussed here derives less from intention than from the position occupied, since exercising unilateral control over parameters with socially relevant effects entails answering for their consequences as well, even when these are neither anticipated nor desired. Third, the article has no access to the interior of any algorithmic system, and claims none; it operates on what is observable from outside the black box or the gray box.

The application of the same construct to the contestability of automated decisions in public administration, including the problem of correspondence between the operative criterion and the documentary record, is developed separately in Shirakura et al. (2026). The present article concentrates on digital platforms and mobilizes meta-identity as a reading grid for the literature on political extremism, recommendation, moderation and influence operations.

The text is organized into five sections. The first delimits the construct and its architecture. The second proposes a division of classificatory labor between the authorities that set parameters and the arena in which that parameterization does or does not become contestable. The third describes the evidence synthesis procedure, the fourth presents what the retrieved literature documents, and the fifth discusses the democratic implications.

1. META-IDENTITY: THE CONSTRUCT AND WHO PARAMETERIZES IT

1.1. Delimitation

Meta-identity is a sociotechnical classificatory construct — inferred, aggregated, opaque and operational — produced and controlled by digital platforms and artificial intelligence systems about entities that circulate within their infrastructures, and not by the targets of those infrastructures, their users, authors and producers. It results from the continuous crossing of personal, contextual, behavioral and relational data and operates as a decision-making reference that modulates visibility, access, reputation, association and circulation, with material effects both inside and outside the digital environment (FERREIRA; TREVISAN, 2025; FERREIRA et al., 2026).

Two terms make up the technical vocabulary employed in this study. "Bearer" designates the entity upon which the classification falls — a subject, an account, a group, an organization, a territory or a piece of content/artifact — and not merely the individual user of a platform. "Predicate"

designates the classificatory label attributed to the bearer, such as "prone to engagement with hate speech", "at-risk account" or "extremist content".

Four characteristics are constitutive of the concept. The first is inference: the classification is deduced from patterns and analytical parameters, rather than simply declared by the bearer. The second is aggregation: the classification results from combining multiple sources, including interactions, behaviors, transactions, internal parameterizations and external institutional categories. The third is opacity: the classificatory criteria and mechanisms are neither fully accessible to the bearer nor fully subject to contestation or audit. The fourth is operationality: the classification produces concrete effects on circulation, access, visibility, reputation, association or accountability.

In order to locate how these processes appear in the literature — not under this terminology, but through their processes, practices and effects — this synthesis employs five complementary observational criteria: attribution, when a predicate is applied to a determinate bearer; inference, when the predicate derives from the processing of signals; stabilization, when it persists or is reapplied beyond the episode that generated it; operationality, when it serves as a reference for differential treatment; and exteriority, when the bearer has no substantive control over the production and revision of the predicate. These five criteria do not replace the four constitutive characteristics of the concept. They organize what can be traced in studies conducted, for the most part, from outside the infrastructures. The coding therefore does not test the full presence of the construct in each study, but locates empirically observable traces of different segments of the classificatory circuit described by the framework.

Exteriority and opacity designate distinct and complementary asymmetries. Exteriority concerns the distribution of authority: the bearer does not substantively control the production or revision of the classification. Opacity concerns the distribution of knowledge: the bearer does not have sufficient access to the predicate, the data, the criteria, the weights and the objectives that guide its production. A classification may become partially transparent and still remain exterior to the bearer; likewise, the possibility of appealing a decision does not necessarily render its criteria transparent.

The delimitation distinguishes the construct from its neighbors. Clarke's (1994) digital persona designates a model of the individual built from data and mobilized as their representative in informational processes. Algorithmic identity (CHENEY-LIPPOLD, 2011, 2017) establishes that categories such as gender, class, or citizenship are probabilistically reconstituted from behavior, but it remains centered on the data subject and does not reach, with the same ease, the classification of

artifacts, collectives, or other objects with which that subject is related. The *data double* (HAGGERTY; ERICSON, 2000) describes the reassembly of dispersed traces into informational duplicates usable by surveillance apparatuses, without making explicit how those representations come to produce differences in treatment or decision-making effects. The *digital twin*, in turn, designates a digital counterpart of an object, process or system, ideally maintained by continuous data flows for purposes of monitoring, simulation or prediction (GRIEVES; VICKERS, 2017; SCHROEDER et al., 2021). Its logic presupposes an aspiration to correspondence between the model and what it represents. On digital platforms, however, the representation may not only be partial: its very purpose may dispense with completeness, selecting and inferring only the attributes relevant to the system's objectives. In contexts of algorithmic mediation, these representations also tend to be probabilistic, opaque, and functionally oriented, without their incompleteness, or eventual distortions, preventing them from producing concrete effects on the subjects, content, or objects classified. Meta-identity starts precisely from this displacement: what matters is less the representation's intended fidelity to its referent than the way in which a classification, even when incomplete and inferred from heterogeneous traces, becomes operational. Nor should its partiality be treated merely as a technical limitation, since it may result from institutional choices about which data to collect, which attributes to infer, which categories to stabilize and which objectives to optimize. The central distinction, therefore, lies in the classificatory asymmetry between those who define and control the representation and those who are rendered legible by it, especially when that representation comes to guide decisions about visibility, recommendation, moderation, access or differential treatment. Whereas the *digital twin* emphasizes correspondence between model and referent, meta-identity emphasizes the incompleteness, opacity, and operational efficacy of classifications that may produce consequences for third parties without those parties knowing, controlling, or being able to contest the criteria by which they were made legible. The *classification situation* (FOURCADE; HEALY, 2013) shows how instruments of classification and evaluation structure life chances, taking scores attributed to individuals as elements capable of conditioning their access to goods, resources and opportunities. Long before the emergence of digital platforms, Bowker and Star (1999) demonstrated that classificatory systems — such as medical taxonomies, racial classifications, censuses or technical nomenclatures — are never neutral: they incorporate normative choices, sediment into infrastructure and tend to become invisible while they work. These concepts, however, were not formulated to track jointly the production of the classification, its algorithmic integration, its operational effects and the eventual feedback of the classified upon the signals that feed new classifications.

The boundary of the construct is not determined by the technical nature of the object involved, but by the function it occupies in the classificatory circuit. Not every datum, piece of content, interface, or operation submitted to algorithmic processing therefore constitutes a meta-identity. The construct takes shape when signals relating to a bearer are processed so as to produce or update an inferred predicate that comes to serve as a reference for differential treatment. Elements that do not take part in this relation may function as inputs, bearers, interfaces or outputs of the system without being confused with the meta-identity; and one and the same element may take on distinct functions depending on where it sits in the circuit. The delimitation is therefore functional and relational: what characterizes meta-identity is not the mere presence of algorithmic mediation, but the transformation of signals about a bearer into an inferred and operationally consequential classification.

Meta-identity, in turn, proposes a unit of analysis directed precisely at this sociotechnical circuit. It is not restricted to the individual: subjects, content, collectives, organizations, works, places, or concepts may all become objects of operational classification. It allows the signals and criteria produced by human actors and institutions to be distinguished from their computational integration, the classification to be tracked over time, and the different outputs through which it becomes consequential to be observed — recommendation, ranking, visibility, moderation, segmentation, monetization or restriction. It thus makes it possible to follow one and the same construct from the authorities, rules and infrastructures that make a given classification possible through to the effects it produces, including when the classified themselves learn to manipulate, circumvent or contest the signals on which that classification depends.

1.2. Dynamics, states, temporal modalities and outputs

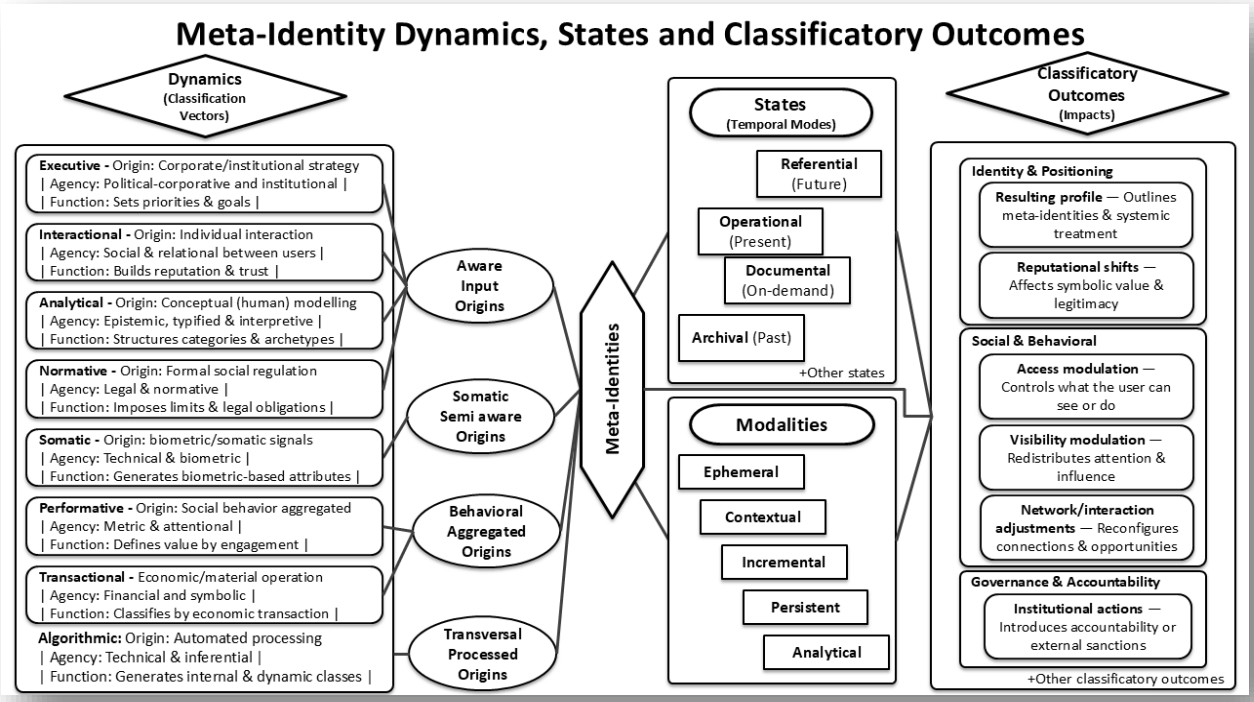
The architecture of meta-identity, set out introductorily in Ferreira and Trevisan (2025) and applied empirically in Ferreira et al. (2026), distinguishes four planes. It is taken up here by reference, without full re-exposition.

The dynamics of meta-identity indicate where the classification vectors come from. Seven operate as input vectors — executive, interactional, analytical, normative, somatic, performative and transactional — and finally the eighth dynamic, the algorithmic one, integrates and composes all the others. The distinction prevents automated processing from being credited with an autonomy it does not have, and locates agency in those responsible for its parameterization.

The states indicate the condition in which the meta-identity operates: operational, the active present that modulates access and visibility; archival, the stored past, available for comparison, reactivation, or training; referential, the projected future, that is, the desired profile template toward which the bearer is induced; and documentary, the record formalized on demand for audit, accountability, or legal use. The referential state is the most consequential, because it shifts classification from diagnosis to inducement: what is at stake is not a probable future, but a future desired by those who control the infrastructure. The logic comes close to Yeung's (2017) *hypernudge*, in which the continuous analysis of data makes it possible to reorganize the informational context and channel attention and decision in directions preferred by those who design the choice architecture. It differs, however, in taking as its analytical object not only the steering of choice, but the classificatory target state itself, in relation to which the subject, content or collective is continuously situated and treated

The modalities indicate the temporality and scope in which the meta-identity is being constituted, according to its duration, informational density, stability, operational purpose and degree of institutionalization. They are: the ephemeral modality, constituted during a specific session or interaction; the contextual, tied to a given situation, theme or conjuncture; the incremental, progressively adjusted by the weighted accumulation of signals; the persistent, stabilized as a lasting reference for reputation, segmentation, access or risk; and the analytical modality, intentionally built by researchers, auditors or regulators as an instrument for scrutinizing meta-identities themselves. The first four may coexist in layers and follow a typical, though not necessary, trajectory from the ephemeral to the persistent. The analytical modality has a distinct status: the categorization remains with whoever is conducting the inquiry, instead of being incorporated by the platform as a classificatory parameter. It is in this sense that it describes the procedure of this article, and it is necessary to stress the conceptual difference explained earlier between the analytical modality and the analytical dynamic: the analytical dynamic is the categorization incorporated by the platform as a parameter; the analytical modality is the categorization retained by the researcher, auditor, or regulator as an instrument of scrutiny.

Figure 1 — Architecture of the dynamics, states and classificatory outputs of meta-identity



Note: The figure summarizes the classification vectors (dynamics), the temporal modes of operation (states), the forms of constitution (modalities), and the effects produced (classificatory outputs) that make up the architecture of meta-identity analyzed in the article. Source: Adapted from Ferreira et al., 2026.

The outputs are organized along three axes: identity and positioning, which alter reputation and legitimacy; social and behavioral, which modulate access, visibility, networks, and opportunities; and governance and accountability, which produce sanctions, blocks, demonetization, banning, and institutional referrals.

2. THE DIVISION OF CLASSIFICATORY LABOR

Attributing to an abstract entity called "the algorithm" the harms arising from the differential treatment that meta-identity authorizes — sanctions, blocks, demonetization, banning, amplified exposure to extremist repertoires — is insufficient: behind the system there is no autonomous agent endowed with a will of its own, but heterogeneous and diverse systems, brought into existence by the practices of those who design them and those who use them (SEAVER, 2017). What manifests itself as emergent system behavior — including when it produces effects that no individual specifically chose or authorized — follows from human decisions taken in earlier layers: which objective to optimize, with which data, under which constraints — decisions that become, in the system's

operational practice, opaque even to those who design it, given the scale at which these algorithms must operate (BURRELL, 2016). This does not derive from institutional secrecy, but from the technical characteristics of machine learning itself — and it is precisely for this reason that holding the system alone accountable is insufficient: accountability needs to reach the network of decisions and actors that constitutes it, not just the final artifact (ANANNY; CRAWFORD, 2018). To understand these developments, the concept of meta-identity describes different types of dynamics internal to the mechanisms operating within digital platforms.

The executive dynamic translates strategic priorities into parameters. It manifests itself at three linked levels: directional decisions that set goals; their operational translation into indicators and product guidelines; and the algorithmic recalibration that follows from them. Weber (2004), in discussing the dilettante, acknowledges that the dilettante may be more productive than the specialist, but questions their capacity to assess the implications of what they produce — those who define objectives rarely master the technical work that implements them, and those who implement rarely decide on the ends. But the sociological point is neither the intention of those who decide nor their capacity to foresee every specific outcome: it is the position they occupy. Responsibility does not depend on predicting the exact technical result — it depends on occupying the place from which it is defined, upstream, what the system is to optimize. It is this position, not the prediction, that the technical opacity described above does not dissolve.

The analytical dynamic, in turn, gathers the socially produced classifications that are incorporated into systems: typologies, archetypes, census categories, clinical, legal and criminological typifications, as well as the internal modeling carried out by data and *marketing* teams. Before a system labels someone as susceptible to political extremism, that notion, category or typification has already circulated in articles, reports, professional practices and other formal or informal interactions — and it may later be absorbed as a segmentation parameter.

From this follows the specific thesis of this section: the executive dynamic defines where the system should tend; the analytical dynamic defines with which categories it will do so. It is in this second operation that a relevant part of the normative content may enter the system — in the form of a technical choice, which tends to make it less visible to public discussion.

The consequence is reflexive and worth recording. The categories with which the academic community describes digital extremism are a potential input for the very infrastructure it analyzes. This does not establish an equivalence of power between those who propose a typology and those

who set a parameter; it establishes that epistemic authority also produces classificatory effects and that responsibility follows authority.

The normative dynamic does not set parameters in the same sense as the executive and analytical ones. It may serve the public interest, through legislative representation and social mobilization, or the interests of those who conduct the executive dynamic, through organized pressure and favorable regulatory design. It is therefore the field in which it is decided whether the parameterization of the other dynamics will be auditable, explainable and appealable.

Treated as an arena, the normative dynamic can be thought of as the place where democratic dispute actually takes place. It is also where a gap that section 4 documents becomes visible: the regulatory debate retrieved here concentrates on those who operate the classificatory infrastructure, while its instrumentalization by external actors appears much less developed.

3. EVIDENCE SYNTHESIS PROCEDURE

3.1. Design and retrieval principle

The methodological procedure adopted is a structured evidence synthesis with two-pass coding, and not a systematic review in the protocol sense of the term: there was no prior protocol registration and no full double screening, and this is declared as a limitation.

Some terms in this section come from the evidence synthesis tradition and warrant brief definition. A seed is a previously known work, chosen for its centrality in the debate, that serves as a starting point for citation tracking; it does not enter the corpus through search, but by declared decision. Tracking has two directions: backward, which follows the references cited by the seed, and forward, which follows the works that cite it. Calibration seeds are works incorporated during the verification of the initial seeds rather than in the original design; the distinction matters because the yield they produce cannot be attributed to the initial design. A citation graph is the network formed by authors' citations (who cites whom) maintained by a bibliographic index — each index has its own coverage, and none is complete, from which follows venue bias, the distortion produced when an index covers different publication types unevenly. Exclusive marginal yield is the number of previously unseen works that a source or seed adds once everything the others had already brought in is discounted, and it exists to prevent double counting. For the same reason, occurrences are distinguished from distinct documents: a work citing three seeds is retrieved three times, so that the sum of retrievals does not equal the number of works. A further distinction is drawn between a record,

the bibliographic entry — a *preprint* and a published version are two records —, and a study, the intellectual unit that brings them together and that is the unit of analysis. A sensitivity test is the measurement of how much a result changes when one procedural decision is altered while the others are held constant. And a frozen corpus designates the point beyond which nothing further is incorporated: works found later go into a separate register and enter only under an explicit rule.

The principle that defined the search was the following: the search did not look for the concept of meta-identity, but for the processes and phenomena it predicts, according to the terminology found in their original literature. No search expression contains "meta-identity" or its variants. Without this rule, the synthesis would be circular — the concept would find only the texts that already adopt the term and the concept.

Retrieval started from four blocks combined with an additional political-object block (extremism, radicalization, radical right, authoritarianism, hate speech): inference and profiling; distribution and recommendation; restriction and moderation; and coordinated action upon the system. The time frame is 2016–2026, taking the 2016 United States election as the moment when platform classification becomes a public political problem. The primary source was the open database OpenAlex (PRIEM et al., 2022), complemented by citation tracking.

3.2. Two calibration findings

Two measurements carried out before any substantive reading produced findings that are of interest in their own right.

The first is the disjunction of vocabularies. Holding constant the query on coordinated action and the platform restriction, the extremism vocabulary retrieved 54 records and the influence operations vocabulary retrieved 178. These are two vocabularies that retrieve, in a strongly unequal fashion, works on related moments of the same circuit. Organized action upon the classificatory infrastructure is retrieved far more frequently under the heading of influence operations than under that of extremism — which helps explain why the field of digital extremism rarely thematizes it.

The second is the selective under-coverage of the citation graph. Forward tracking starts from a set of seed works and follows the works that cite them. The six initial seeds were chosen because they deliberately cover both sides of the controversy over algorithmic radicalization, and not only the side that supports the thesis: Ribeiro et al. (2020), who document migration pathways toward extreme channels on YouTube; Ledwich and Zaitsev (2020), who find the opposite result, with the algorithm

favoring content from established outlets; Hosseinmardi et al. (2021), who find no evidence that recommendation directs attention to radical content; Munger and Phillips (2022), who shift the explanation from algorithmic supply to audience demand; Haroon et al. (2023), who find increasingly congenial recommendation along the pathway; and Piccardi et al. (2025), whose feed reranking experiment measures a causal effect on affective polarization. A seed set composed only of studies confirming the thesis would produce a corpus biased by construction.

Checking these seeds — a process intended only to verify whether title matching was correct — revealed implausible citation counts. Ribeiro et al. (2020) appeared with 48 indexed citing works and Ledwich and Zaitsev (2020) with 14, orders of magnitude below what other indexes record for works of such centrality. Verification ruled out the duplicate record hypothesis and identified under-coverage of the OpenAlex graph for two types of venue: computing conference proceedings and independent open access journals.

The same verification query had a productive side effect. By returning, alongside each seed, its nearest neighbors in the index, it exposed five works that satisfied the inclusion criterion, recurred in the neighborhood of more than one seed, and were absent from the initial set — a sign of centrality that the original design had missed. They were promoted to seeds, with their round of entry recorded in the selection flow: Whittaker et al. (2021), on recommender systems and the amplification of extremist content; Yesilada and Lewandowsky (2022), a systematic review on recommendation and problematic content on YouTube; Shaw (2023), a synthesis on social media, extremism and radicalization; Bouchaud and Ramaciotti (2024), a methodological examination of recommender system audits themselves; and Gauthier et al. (2026), on the political effects of X's feed algorithm.

Tracking was then redone on Semantic Scholar, whose graph covers these venues better, while OpenAlex was kept as the metadata source in order to preserve the homogeneity of the corpus. Since two changes were made simultaneously — the source of citing works and the seed set — the effects were decomposed. Holding the six initial seeds constant, OpenAlex returned 14 eligible and previously unseen citing works, and Semantic Scholar returned 171. The five calibration seeds contributed 25 exclusive records, that is, after discounting those the six original seeds had already reached. The combined pool of previously unseen items was 196 records, of which 186 were incorporated after metadata retrieval and type and language filters. The gain is attributable above all to source coverage, and not to the expansion of the seed set.

The methodological implication goes beyond this article and is worth making explicit. Bibliographic indexes do not observe the literature directly: they reconstruct the citation graph from what is indexed, and that coverage may vary across venue types. In the set examined here, the under-coverage observed fell especially on computing conference proceedings and independent open access journals, venues that are important to the empirical core of research on algorithmic auditing. The pattern is therefore consistent with a coverage distortion aligned with the epistemic division of the field. A synthesis that depends on a single graph may under-retrieve part of the empirical output and produce a biased picture of the literature. It is this systematic sensitivity of retrieval to venue type and to the graph used that is designated here as venue bias, which recommends that any synthesis on this object declare the graph used and, where possible, submit it to a sensitivity test against a second index.

3.3. Corpus composition and boundary rule

The corpus was frozen at 813 bibliographic records, consolidated into 749 studies. The reduction of 64 records follows from 62 version groups: 60 groups formed by two records and two groups formed by three records. *Preprints* and published versions of the same work were screened separately but counted as a single unit of evidence. Discarded, with a record of the decision, were 12 paratext items, six supplementary artifacts, two identifier duplicates, and thirteen records whose eligibility followed from an indexer metadata error that had attached the abstract of an audit study to unrelated articles.

The rule underpinning the inclusion decision was semantic: studies were included whose findings say something about how a platform system infers, ranks, distributes, restricts or monetizes — or about how actors act to alter those signals; studies were excluded whose findings concern the performance of an instrument built by the researchers themselves.

During the research, 145 full-text files obtained through legitimate access routes were preserved. After version consolidation and metadata correction, 136 files corresponded to records in the current corpus and represented 127 distinct studies; nine files remaining from earlier stages were kept exclusively for auditing. Rule-based, auditable lexical localization was applied to the 127 studies with retrieved text and identified at least one candidate in 118 of them. For each variable, the candidate, the literal sentence, and the section in which it occurred were preserved. Sentences containing bibliographic citation markers were flagged because they might report findings by third parties rather than by the coded study. The corpus, the scripts, the protocols and the corresponding

audit materials, including the rules and verification procedures employed in the assisted localization, are deposited in Zenodo by Ferreira, Trevisan and Shirakura (2026).

3.4. Scale of attribution basis

A recurrent difficulty in the literature on algorithms is inferring internal architecture from outputs. To address it, a distinction is drawn among five levels of the basis on which the operation is attributed: documented by the platform itself, in a transparency report, technical documentation, open source code or regulatory filing; observed in operation, through an audit with synthetic accounts or an experiment; inferred from output patterns, in trace data or network analysis; inferred from user and creator reports; and postulated as a theoretical premise.

This axis does not coincide with design strength. An audit with synthetic accounts is a robust design with zero access; documentation of a clustering system published by the platform itself has maximum access and is not, in itself, an empirical study.

4. WHAT THE LITERATURE DOCUMENTS

The results in this section follow from rule-assisted localization applied to the 127 studies with full text, of which 118 presented at least one candidate. The retrieved studies observe different segments of the classificatory circuit; meta-identity is mobilized here as a grid for integrating this partial evidence, and not as a causal variable directly tested by each study. The works discussed by name below are illustrative cases selected to make visible different mechanisms, designs, and results found in the corpus, and do not constitute a second sample.

4.1. Composition and what it already tells us

The distribution by platform is unequal in a revealing way. X/Twitter appears in 63 studies; YouTube in 36; Facebook in 35; TikTok and Instagram in 16 each; Reddit in 13; WhatsApp in 10; Telegram in 5; Gab in 4; BitChute in 3; Twitch in 2. The ordering does not measure political weight, but data access. X/Twitter historically offered relatively broad access to public data and APIs, which favored its presence in academic research; TikTok and Telegram, politically central in several contexts, are residual. The composition of the corpus is itself an effect of the access regime the article analyzes.

As for the bearer of the classification, user or account appears in 90 studies; collective or network in 81; work or artifact in 80; topic and organization in 75 each; and territory in 38. The presence of territory is relevant because the concept claims reach over non-human entities and over

productions, and the field documents this more than had been assumed: there are studies recording differential treatment by region and by country.

4.2. Inference, distribution and the radicalization controversy

The literature auditing recommender systems constitutes the most robust core of evidence in the corpus, and it is also the most divided.

Huszár et al. (2022) conducted, while the platform was still Twitter (renamed X in 2023), a large-scale randomized experiment with a control group of nearly two million daily active accounts subjected to a reverse-chronological timeline, covering messages from elected legislators in seven countries. They found consistent amplification of the mainstream right in six of the seven cases — and, against current expectations, found no advantage for the far right over the mainstream right. It is the only case in the corpus in which the platform audits itself and publishes the result, something that ceased to occur after the platform's acquisition by Elon Musk in 2022.

Haroon et al. (2023) audited YouTube's recommender system with one hundred thousand synthetic accounts and found increasingly congenial recommendations as one advances along the pathway, especially for right-leaning profiles, without this amounting to a necessary progression in the ideological extremity of the recommended content. Tomlein et al. (2021) and Srba et al. (2023) showed that misinformation bubbles form quickly, but that they can also be reversed when the profile starts consuming debunking content — a result that bears directly on the gradation between the ephemeral and persistent modalities. Lahsaini et al. (2024), auditing recommendations on abortion across six profiles, found a result contrary to expectation: a negative score in all profiles, that is, a predominance of content favorable to abortion and of debunking.

Ribeiro et al. (2023) offer the synthesis that forecloses any easy conclusion. Faced with the contrast between audits that find bias and usage data that find no corresponding consumption, they show by simulation that the collaborative nature of recommender systems and the niche character of extreme content suffice to dissolve the paradox: the algorithm may favor without directing, because algorithmic supply and actual consumption are not equivalent. Hilbert et al. (2024), in a meta-analysis of 151 audits drawn from 33 studies, estimate that 8% to 10% of recommendations are harmful, with no statistically significant difference across platforms or across risk types.

Taken together, this body of work supports the displacement proposed in the introduction. It is not necessary to claim that recommendation radicalizes in order to hold that there are attributed

classifications serving as a reference for differential treatment — and this second claim is observable in traces. Recommendation, exposure, consumption, misinformation, polarization, or attitude change are not treated here as equivalent measures of radicalization, but as different outputs or segments of the classificatory circuit documented by their respective designs.

4.3. Restriction and moderation

The second core is restriction, and here the strongest findings are findings of no effect.

Tonneau et al. (2026) carried out a global audit of hate speech moderation on Twitter/X, based on a representative sample built from a full 24-hour capture of public messages, with 540,000 messages annotated by trained raters across eight languages. Five months after posting, 80% of hate messages remained available, including explicit incitement to violence, and were no more likely to be removed than non-offensive messages: neither severity nor visibility increased the chance of removal.

Shukla et al. (2025) audited Twitch's automated moderation system using broadcasting accounts created as isolated environments, sending more than 107,000 comments drawn from four datasets. Up to 94% of overtly offensive messages escaped moderation in some datasets; the contextual addition of trigger terms raised detection to 100%, revealing lexical dependence; and up to 89.5% of benign examples, which used sensitive terms in pedagogical or affirmative contexts, were blocked. The study documents attribution, inference, and operationality, but does not allow the stabilization of the predicate beyond the moderation event to be established unequivocally.

Hartmann et al. (2025) shift the bearer. They audited five commercial moderation interfaces, analyzing five million queries, and showed that these services frequently decide on the basis of group identity terms, under-moderating implicit hate speech — especially against LGBTQIA+ people — and over-moderating counterspeech and reappropriated terms. Here the classifier does not belong to the platform: it is an infrastructure sold to platforms, which extends the analytical reach of the construct to associated third parties.

Gennaro et al. (2025) document distribution: across five original datasets, collected on multiple platforms in Switzerland and the United States, most hate speech comes from a small fraction of users; and a pre-registered field experiment indicates that counterspeech strategies produce small reductions, being ineffective precisely against the most prolific producers.

4.4. The adversarial activation of the input dynamics

Adversarial activation is what matters most to the article's question and is the least documented. It refers to deliberate action by organized political actors upon the input vectors — above all the interactional, performative and transactional dynamics — with the aim of altering the meta-identity attributed. The action takes three modes, and the distinction is decisive: self-referential, when it targets one's own account or content, through fabricated engagement and reach inflation; other-directed, when it targets third parties, through coordinated reporting, harassment and the fabrication of risk signals about adversaries; and environmental, when it targets the informational space, through flooding and trend manipulation.

Only the other-directed mode sustains the specific proposition that organized actors may act upon signals directed at third parties, and aggregating the three would produce false confirmation, because the literature on fabricated engagement is abundant and almost entirely self-referential. In this mode, the evidence makes it possible to identify the deliberate production of signals directed at third parties; this does not entail, except where directly observed, demonstrating the internal classification subsequently produced by the platform. In the assisted localization, the self-referential mode appears in 11 studies, the other-directed in seven and the environmental in four; in 99 of the 118 studies there is no candidate of any mode. The modes are not mutually exclusive at the study level, which is why their summed frequencies exceed the 19 studies in which at least one candidate appeared.

The most complete case is that of Recabarren et al. (2023). The authors conducted 19 semi-structured interviews with participants in influence operations acting on Venezuela's image and validated part of the responses with data collected over nearly four months from the accounts these participants control. They document *Patria*, a government platform on which operators log activities and receive benefits, and they systematize the declared strategies for promoting political content and for evading platform penalties. Of the whole corpus, this is the most direct record of action upon the classifier described from the actor's side.

The scarcity of the remaining cases calls for caution. It may indicate that the phenomenon is rare; it may indicate that it is described by vocabularies distinct from those prevailing in the extremism literature, in line with the disjunction measured in calibration; and it may still reflect the reach of the localization instrument. These last two possibilities are not independent: terminological dispersion across fields may reduce the capacity of lexical rules to locate equivalent manifestations of the phenomenon. It is not possible, at this stage, to estimate how much each of these hypotheses contributes to the low observed frequency.

4.5. Which conditions the literature reaches

The preliminary coding examined which of the five complementary observational criteria — attribution, inference, stabilization, operationality and exteriority — appear documented in each study.

Table 1 — Indication of each observational condition in the studies with located candidates (n = 118)

Condition	Studies	%
Stabilization	66	56
Inference	59	50
Operationality	31	26
Attribution	28	24
Exteriority	6	5

Source: prepared by the authors on the basis of the dataset deposited by Ferreira, Trevisan and Shirakura (2026). Note: The frequencies are not mutually exclusive and represent the presence of each condition separately. The absence of a candidate was interpreted as the absence of documentation located in the study, and not as evidence of the absence of the condition. No study presented simultaneous indications of all five conditions.

The exploratory analysis identified no study presenting simultaneous indications of all five conditions. The interpretation of this result cannot yet be conclusive, since it reflects, in unknown proportion, both the literature retrieved and the reach of the localization patterns. More robust than the absolute values, for now, is the observed ordering: the literature documents most frequently the inference and stabilization of classifications; less frequently their attribution to a determinate bearer and their operationality; and rarely exteriority — the condition that makes visible the asymmetry of authority over the production and revision of the predicate and that, articulated with opacity, limits its contestation. The same pattern appears in the contestability indicators coded in this synthesis. Appeal against automated decisions appears in 51 studies, and independent auditing in 34; the bearer's access to the categories attributed to them appears in only three. At this exploratory stage, the convergence among the low documentation of exteriority, the rarity of access to the categories, and the located cases of explicit absence of reasons or of review is treated as an interpretive hypothesis; the co-occurrence among these indicators was not examined.

At this point, meta-identity reveals its usefulness as a middle-range theory, in the sense proposed by Merton (1968), by offering a taxonomy and a vocabulary of sociotechnical translation between fields that frequently observe distinct parts of the same process. For the social sciences and

the humanities, this architecture makes it possible to synthesize complex classificatory operations without dispensing, where necessary, with access to the specific technical literature of engineering and data science. For researchers and developers in technical areas, the same vocabulary renders legible the reach and the material, economic, social and political consequences of the categories, parameters and systems they conceive and operate. Its function, therefore, is not to replace the specialization of each field, but to establish a common layer of analysis that allows technical mechanisms, human decisions and social consequences to be related to one another.

5. THE DISPUTE OVER THE CLASSIFICATORY REGIME

Three consequences follow from the picture described. The first is that contemporary extremisms and authoritarianisms circulate not only through discourses, leaderships and organizations, but also through infrastructures that classify and modulate audiences, content and networks. Meta-identity offers a language for describing how ephemeral interactions with radicalized content may feed into contextual, incremental, or persistent classifications and thus compose trajectories of engagement and selective exposure.

The second is that the fundamental asymmetry lies not in the quantity of data, but in the distribution of knowledge about the criteria. The platform holds multiple dimensions of the bearer; the bearer does not fully hold the criteria that classify them. This asymmetry finds legal translation in the debate on automated decision-making: Yeung and Lodge (2019) situate part of the normative concerns of algorithmic regulation within that process, while Bayamlıoğlu (2022) formulates more directly the right to contest as a procedural guarantee that goes beyond the mere provision of explanations. Without access to the attributed categories and without a route to review, the classified party lacks the minimum procedural conditions for meaningfully contesting the classification. The data from this synthesis are consistent with that asymmetry, although they do not demonstrate it directly — owing to the limitation or inaccessibility of digital platforms' operating models. Bearing this particularity in mind, it is possible to note that exteriority is the least documented condition, and the bearer's access to the categories attributed to them is the least frequent contestability indicator. The convergence between these results reveals above all a gap in the literature: studies examine the decision and its effects more often than the knowledge and the control available to the classified party over the classificatory process.

The third consequence is regulatory and follows from the discussion of the normative dynamic presented in section 2. Instruments such as the Digital Services Act (EUROPEAN PARLIAMENT; COUNCIL OF THE EUROPEAN UNION, 2022), with obligations of transparency, systemic risk assessment and independent auditing for very large platforms, are addressed primarily to those who operate the classificatory infrastructure. In the retrieved corpus, no direct regulatory treatment was identified of coordinated action by external actors aimed at inducing classifications about adversaries. The result should be understood as a gap in the retrieved literature, and not as a demonstration that no applicable legal provision exists.

It is not maintained here that algorithmic classification determines trajectories. The classified produce the signals, adapt behaviors, dispute categories and mobilize institutions — and the Venezuelan case documents actors who systematically learn to operate the classifier. In another register, which was not part of the structured evidence synthesis, Zenkl (2025) shows how workers embedded in a technobureaucratic regime seek to *tame* algorithmic futures, negotiating their limits and claiming margins of discretion, in a relation that cannot be reduced to either adherence or resistance. What is maintained is something more modest and more verifiable: contemporary democratic dispute involves not only the regulation of extremist messages, but the contestation of the classificatory regimes that define what will be seen, recommended, moderated, monetized, or rendered suspect.

This dispute can be situated within public sphere theory. Habermas (2023) argues that the algorithm-driven platform structure fragments the public sphere into closed pseudo-publics, eroding the discursive formation of opinion and the institutions on which democracy depends. Meta-identity can be mobilized as a sociotechnical key for analyzing this fragmentation: it is through the modulation of visibility and recommendation on the basis of classifications inaccessible to the classified party that the infrastructure can reorganize who speaks with whom. It is worth stressing, however, that networked political engagement is also affective (PAPACHARISSI, 2015) and agonistic (MOUFFE, 2013), so that the democratic response is not reducible to restoring rational debate, but runs through rendering contestable the regimes that distribute visibility and suspicion.

CONCLUDING REMARKS

The article mobilized meta-identity as an analytical grid for examining the literature on political extremism in digital networks. The synthesis did not look for works employing the concept, specifically so as not to incur self-reference within the corpus analyzed, but sought studies of the

phenomena the concept makes it possible to describe and analyze. The coding admitted the response "does not apply", and the limits of the procedure were made explicit before the results were interpreted.

Three preliminary results are worth recording. The literature on digital extremism and the literature on influence operations reach related moments of the same classificatory circuit through distinct vocabularies. Among the five complementary observational criteria, inference and stabilization appear most frequently, while exteriority is rare; none of the 118 studies brought together simultaneous indications of all five criteria. Organized action upon the input vectors directed at third parties, which constitutes the article's most specific proposition, appears in a small number of studies, remaining a sociotechnical hypothesis with a declared empirical gap rather than a consolidated finding.

Some limitations delimit the reach of this study. There was no access to the interior of any algorithmic system, and the synthesis operates on what is observable from outside. Full analysis was restricted to the 127 studies for which at least one complete text was retrieved, and closed-access studies are under-represented. It is therefore not assumed that this subset is representative of the consolidated bibliographic corpus, so access bias cannot be ruled out. The composition of the corpus by platform reflects the data availability regime, not the political weight of the platforms.

The agenda that follows is twofold. On the research side, the other-directed mode of adversarial activation needs to be submitted to an empirical design of its own, capable of distinguishing action upon one's own classification from action upon someone else's. On the public side, the debate on platform regulation needs to incorporate a dimension still insufficiently addressed: not only what platforms do with the classifications they produce, but how organized actors may act upon their signals in order to try to induce algorithmic classifications and treatments of third parties.

REFERENCES

ANANNY, M.; CRAWFORD, K. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, v. 20, n. 3, p. 973–989, 2018. DOI: https://doi.org/10.1177/1461444816676645.

BAYAMLIOĞLU, E. The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called "right to explanation". *Regulation & Governance*, v. 16, n. 4, p. 1058–1078, 2022. DOI: https://doi.org/10.1111/rego.12391.

BOUCHAUD, P.; RAMACIOTTI, P. Auditing the audits: Evaluating methodologies for social media recommender system audits. *Applied Network Science*, v. 9, n. 1, p. 59, 2024. DOI: https://doi.org/10.1007/s41109-024-00668-6.

BOWKER, G. C.; STAR, S. L. *Sorting things out: Classification and its consequences*. Cambridge: MIT Press, 1999.

BURRELL, J. How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society*, v. 3, n. 1, 2016. DOI: https://doi.org/10.1177/2053951715622512.

CHENEY-LIPPOLD, J. A new algorithmic identity: Soft biopolitics and the modulation of control. *Theory, Culture & Society*, v. 28, n. 6, p. 164–181, 2011. DOI: https://doi.org/10.1177/0263276411424420.

CHENEY-LIPPOLD, J. *We are data: Algorithms and the making of our digital selves*. New York: New York University Press, 2017.

CLARKE, R. The digital persona and its application to data surveillance. *The Information Society*, v. 10, n. 2, p. 77–92, 1994. DOI: https://doi.org/10.1080/01972243.1994.9960160.

EUROPEAN PARLIAMENT; COUNCIL OF THE EUROPEAN UNION. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). *Official Journal of the European Union*, L 277, p. 1–102, 2022. Available at: https://eur-lex.europa.eu/eli/reg/2022/2065/oj Accessed on: 31 Aug. 2026.

FERREIRA, A. H.; TREVISAN, A. C.; BAPTISTA, C. M.; RAMOS-ANTÓN, R.; COMIN, Á. A.; CARVALHO, H. F.; VENDRELL, S.; SÁ, V. O. Meta-identity and algorithmic mediation on digital platforms: A comparative analysis of AI–human content categorization. *Societies*, v. 16, n. 4, art. 132, 2026. DOI: https://doi.org/10.3390/soc16040132.

FERREIRA, A. H.; TREVISAN, A. C. Meta-identidade e plataformas digitais: Disputas por transparência e controle na criação algorítmica das identidades pelos sistemas de IA. *SciELO Preprints*, 2025. DOI: https://doi.org/10.1590/SciELOPreprints.13161.

FERREIRA, A. H.; TREVISAN, A. C.; SHIRAKURA, C. S. *Meta-identidades e extremismo político nas redes digitais: corpus, pipeline e materiais de auditoria* [Meta-identities and political extremism in digital

networks: corpus, pipeline and audit materials]. Version 1.0. [Data set]. Zenodo, 2026. DOI: https://doi.org/10.5281/zenodo.22237155.

FOURCADE, M.; HEALY, K. Classification situations: Life-chances in the neoliberal era. *Accounting, Organizations and Society*, v. 38, n. 8, p. 559–572, 2013. DOI: https://doi.org/10.1016/j.aos.2013.11.002.

GAUTHIER, G. et al. The political effects of X's feed algorithm. *Nature*, v. 652, n. 8109, p. 416–423, 2026. DOI: https://doi.org/10.1038/s41586-026-10098-2.

GENNARO, G. et al. The distribution of hate speech and its implications for content moderation. *Political Science Research and Methods*, p. 1–9, 2025. DOI: https://doi.org/10.1017/psrm.2025.10063.

GRANDI, G. Moraes diz que explosão não foi fato isolado e considera segundo pior ato após o 8/1. *Gazeta do Povo*, 14 Nov. 2024. Available at: https://www.gazetadopovo.com.br/republica/moraes-explosao-fato-isolado-considera-segundo-pior-apos-8-1/ Accessed on: 31 Aug. 2026.

GRIEVES, M.; VICKERS, J. Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. In: KAHLEN, F.-J.; FLUMERFELT, S.; ALVES, A. (eds.). *Transdisciplinary perspectives on complex systems: New findings and approaches*. Cham: Springer International Publishing, 2017. p. 85–113. DOI: https://doi.org/10.1007/978-3-319-38756-7_4.

HABERMAS, J. A new structural transformation of the public sphere and deliberative politics. Translated by C. Cronin. Cambridge: Polity Press, 2023.

HAGGERTY, K. D.; ERICSON, R. V. The surveillant assemblage. *The British Journal of Sociology*, v. 51, n. 4, p. 605–622, 2000. DOI: https://doi.org/10.1080/00071310020015280.

HAROON, M. et al. Auditing YouTube's recommendation system for ideologically congenial, extreme, and problematic recommendations. *Proceedings of the National Academy of Sciences*, v. 120, n. 50, e2213020120, 2023. DOI: https://doi.org/10.1073/pnas.2213020120.

HARTMANN, D. et al. Lost in moderation: How commercial content moderation APIs over- and under-moderate group-targeted hate speech and linguistic variations. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. New York: ACM, 2025. Article 175. DOI: https://doi.org/10.1145/3706598.3713998.

HILBERT, M. et al. 8–10% of algorithmic recommendations are "bad", but… an exploratory risk-utility meta-analysis and its regulatory implications. *International Journal of Information Management*, v. 75, 102743, 2024. DOI: https://doi.org/10.1016/j.ijinfomgt.2023.102743.

HOSSEINMARDI, H. et al. Examining the consumption of radical content on YouTube. *Proceedings of the National Academy of Sciences*, v. 118, n. 32, e2101967118, 2021. DOI: https://doi.org/10.1073/pnas.2101967118.

HUSZÁR, F. et al. Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences*, v. 119, n. 1, e2025334119, 2022. DOI: https://doi.org/10.1073/pnas.2025334119.

KUNDNANI, A. A decade lost: Rethinking radicalisation and extremism. London: Claystone, 2015.

LAHSAINI, M.; LECHIAKH, M.; MAURER, A. *Auditing health-related recommendations in social media: A case study of abortion on YouTube*. arXiv, 2024. DOI: https://doi.org/10.48550/arXiv.2404.07896.

LEDWICH, M.; ZAITSEV, A. Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *First Monday*, v. 25, n. 3, 2020. DOI: https://doi.org/10.5210/fm.v25i3.10419.

MANFRIN, J. Autor de explosões em Brasília anunciou atos nas redes sociais, se despediu e deixou recado à PF. *Gazeta do Povo*, 14 Nov. 2024. Available at: https://www.gazetadopovo.com.br/republica/autor-de-explosoes-em-brasilia-anunciou-atos-nas-redes-sociais-se-despediu-e-deixou-recado-a-pf/ Accessed on: 31 Aug. 2026.

MCCAULEY, C.; MOSKALENKO, S. Understanding political radicalization: The two-pyramids model. *American Psychologist*, v. 72, n. 3, p. 205–216, 2017. DOI: https://doi.org/10.1037/amp0000062.

MERTON, R. K. *Social theory and social structure*. Enlarged ed. New York: Free Press, 1968.

MOUFFE, C. Agonistics: Thinking the world politically. London: Verso, 2013.

MUNGER, K.; PHILLIPS, J. Right-wing YouTube: A supply and demand perspective. *The International Journal of Press/Politics*, v. 27, n. 1, p. 186–219, 2022. DOI: https://doi.org/10.1177/1940161220964767.

PAPACHARISSI, Z. *Affective publics: Sentiment, technology, and politics*. New York: Oxford University Press, 2015.

PICCARDI, T. et al. Reranking partisan animosity in algorithmic social media feeds alters affective polarization. *Science*, v. 390, n. 6776, eadu5584, 2025. DOI: https://doi.org/10.1126/science.adu5584.

PRIEM, J.; PIWOWAR, H.; ORR, R. *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. arXiv, 2022. Available at: https://arxiv.org/abs/2205.01833 Accessed on: 31 Aug. 2026.

RECABARREN, R. et al. Strategies and vulnerabilities of participants in Venezuelan influence operations. In: *32nd USENIX Security Symposium (USENIX Security 23)*. Berkeley: USENIX Association, 2023. p. 6683–6700. Available at: https://www.usenix.org/conference/usenixsecurity23/presentation/recabarren Accessed on: 31 Aug. 2026.

RIBEIRO, M. H. et al. Auditing radicalization pathways on YouTube. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York: ACM, 2020. p. 131–141. DOI: https://doi.org/10.1145/3351095.3372879.

RIBEIRO, M. H.; VESELOVSKY, V.; WEST, R. The amplification paradox in recommender systems. *Proceedings of the International AAAI Conference on Web and Social Media*, v. 17, n. 1, p. 1138–1142, 2023. DOI: https://doi.org/10.1609/icwsm.v17i1.22223.

SCHROEDER, G. N. et al. Digital Twin connectivity topologies. *IFAC-PapersOnLine*, v. 54, n. 1, p. 737–742, 2021. DOI: https://doi.org/10.1016/j.ifacol.2021.08.086.

SEAVER, N. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, v. 4, n. 2, 2017. DOI: https://doi.org/10.1177/2053951717738104.

SHAW, A. Social media, extremism, and radicalization. *Science Advances*, v. 9, n. 35, eadk2031, 2023. DOI: https://doi.org/10.1126/sciadv.adk2031.

SHIRAKURA, C. S.; FERREIRA, A. H.; TREVISAN, A. C.; CASAS MAS, B. Perfis algorítmicos e democracia: meta-identidades, poder classificatório e contestabilidade na esfera pública digital [Algorithmic profiles and democracy: meta-identities, classificatory power and contestability in the digital public sphere]. *SciELO Preprints*, 2026. DOI: https://doi.org/10.1590/SciELOPreprints.17781.

SHUKLA, P. et al. Silencing empowerment, allowing bigotry: Auditing the moderation of hate speech on Twitch. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Stroudsburg: ACL, 2025. p. 22771–22797. DOI: https://doi.org/10.18653/v1/2025.acl-long.1110.

SRBA, I. et al. Auditing YouTube's recommendation algorithm for misinformation filter bubbles. *ACM Transactions on Recommender Systems*, v. 1, n. 1, Article 6, p. 1–33, 2023. DOI: https://doi.org/10.1145/3568392.

TOMLEIN, M. et al. An audit of misinformation filter bubbles on YouTube: Bubble bursting and recent behavior changes. In: *Proceedings of the 15th ACM Conference on Recommender Systems*. New York: ACM, 2021. p. 1–11. DOI: https://doi.org/10.1145/3460231.3474241.

TONNEAU, M. et al. *The enforcement and feasibility of hate speech moderation on Twitter*. arXiv, 2026. DOI: https://doi.org/10.48550/arXiv.2604.12289.

WEBER, M. *Ciência e política: Duas vocações* [Science as a vocation; Politics as a vocation]. São Paulo: Cultrix, 2004.

WHITTAKER, J. et al. Recommender systems and the amplification of extremist content. *Internet Policy Review*, v. 10, n. 2, 2021. DOI: https://doi.org/10.14763/2021.2.1565.

YESILADA, M.; LEWANDOWSKY, S. Systematic review: YouTube recommendations and problematic content. *Internet Policy Review*, v. 11, n. 1, 2022. DOI: https://doi.org/10.14763/2022.1.1652.

YEUNG, K.' Hypernudge': Big Data as a mode of regulation by design. *Information, Communication & Society*, v. 20, n. 1, p. 118–136, 2017. DOI: https://doi.org/10.1080/1369118X.2016.1186713.

YEUNG, K.; LODGE, M. (eds.). *Algorithmic regulation*. Oxford: Oxford University Press, 2019. DOI: https://doi.org/10.1093/oso/9780198838494.001.0001.

ZENKL, T. The taming of sociodigital anticipations: AI in the digital welfare state. *Frontiers in Sociology*, v. 10, art. 1556675, 2025. DOI: https://doi.org/10.3389/fsoc.2025.1556675.

RESEARCH DATA AVAILABILITY STATEMENT: The consolidated bibliographic corpus, the record-to-study map, the selection flow, the discarded records with their reasons, the analytical protocol, the scripts and the audit materials are deposited in Zenodo, currently under restricted access subject to the authors' authorization (Ferreira, Trevisan and Shirakura, 2026): https://doi.org/10.5281/zenodo.22237155. The restriction applies to the materials deposited during this stage of circulation and evaluation of the work and does not prevent their preservation or their traceability. The full texts of the studies analyzed, being third-party materials, are not part of the

deposit; for audit purposes, their identifiers, their retrieval status and their *hashes* were preserved, making it possible to verify the correspondence between the files used in processing and the records documented in the pipeline.

FUNDING:

Allan Herison Ferreira is a doctoral fellow of the Fundação para a Ciência e a Tecnologia (FCT), Portugal, reference: https://doi.org/10.54499/2022.14807.BD. Ana Carolina Trevisan is a doctoral fellow of the Fundação para a Ciência e a Tecnologia (FCT), Portugal, reference: https://doi.org/10.54499/UI/BD/153755/2022. This work is funded by National Funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under Project reference UID/05021/2023.

AUTHOR CONTRIBUTIONS:

Allan Herison Ferreira: Conceptualization, Methodology, Investigation, Formal analysis, Writing – original draft. Ana Carolina Trevisan: Conceptualization, Methodology, Investigation, Writing – review & editing. Camila Sayuri Shirakura: Investigation, Writing – review & editing.

CONFLICT OF INTEREST STATEMENT:

The authors declare that there is no conflict of interest to report.

DECLARATION OF USE OF ARTIFICIAL INTELLIGENCE (AI):

Artificial intelligence resources were used in an assistive capacity at four stages, under the supervision and responsibility of the authors: the construction and execution of queries to open bibliographic databases; the rule-based, auditable lexical localization of candidate passages in the full text of the retrieved studies; the organization and normalization of the editorial structure; and language editing. The scientific content is the sole responsibility of the authors.

AUTHOR BIOGRAPHIES

Allan Herison Ferreira is a doctoral candidate in Communication Sciences at NOVA University Lisbon, with an FCT fellowship at ICNOVA. He holds a master's degree in Sociology from USP and researches social identities, work in information technology, artificial intelligence and algorithmic mediation.

Ana Carolina Trevisan is a doctoral candidate in Communication Sciences at NOVA University Lisbon, with an FCT fellowship at IFILNOVA. She holds a master's degree in Sociology from USP and researches argumentative strategies on social media, political sociology and socio-argumentative analysis.

Camila Sayuri Shirakura is a master's student in the Graduate Program in Social Sciences at the Universidade Estadual de Maringá (UEM). She investigates the relationship between technological capture, state power and citizens' rights.

This preprint was submitted under the following conditions:

- The authors declare that the necessary Terms of Free and Informed Consent of participants or patients in the research were obtained and are described in the manuscript, when applicable.
- The authors declare that the preparation of the manuscript followed the ethical norms of scientific communication.
- The authors declare that they are aware that they are solely responsible for the content of the preprint and that the deposit in SciELO Preprints does not mean any commitment on the part of SciELO, except its preservation and dissemination.
- The authors declare that the data, applications, and other content underlying the manuscript are referenced.
- The deposited manuscript is in PDF format.
- The authors declare that the research that originated the manuscript followed good ethical practices and that the necessary approvals from research ethics committees, when applicable, are described in the manuscript.
- The authors declare that once a manuscript is posted on the SciELO Preprints server, it can only be taken down on request to the SciELO Preprints server Editorial Secretariat, who will post a retraction notice in its place.
- The authors agree that the approved manuscript will be made available under a Creative Commons CC-BY license.
- The submitting author declares that the contributions of all authors and conflict of interest statement are included explicitly and in specific sections of the manuscript.
- The authors declare that the manuscript was not deposited and/or previously made available on another preprint server or published by a journal.
- If the manuscript is being reviewed or being prepared for publishing but not yet published by a journal, the authors declare that they have received authorization from the journal to make this deposit.
- The submitting author declares that all authors of the manuscript agree with the submission to SciELO Preprints.