Meta-identidades e extremismo político nas redes digitais: classificação algorítmica, radicalização e disputas democráticas

Allan Herison Ferreira, Ana Carolina Trevisan , Camila Sayuri Shirakura

https://doi.org/10.1590/SciELOPreprints.17784

Submetido em: 2026-09-01
Postado em: 2026-09-03 (versão 1)
(AAAA-MM-DD)

A moderação deste preprint recebeu o(s) endosso(s) de:

- Belén Casas-Mas (ORCID: https://orcid.org/0000-0001-8329-0856)
- Fabrizio Macagno (ORCID: https://orcid.org/0000-0003-0712-421X)

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

ALLAN HERISON FERREIRA
ORCID: https://orcid.org/0000-0003-3606-2089
allanherison@gmail.com
Universidade NOVA de Lisboa, Instituto de Comunicação (ICNOVA); Laboratório de Pesquisa Social (LAPS/USP). Lisboa, Portugal

ANA CAROLINA TREVISAN
ORCID: https://orcid.org/0000-0002-2818-5858
anacarolinatcf@gmail.com
Universidade NOVA de Lisboa, Instituto de Filosofia (IFILNOVA); Laboratório de Pesquisa Social (LAPS/USP). Lisboa, Portugal

CAMILA SAYURI SHIRAKURA
ORCID: https://orcid.org/0009-0002-8291-2945
camila.shirakura@gmail.com
Universidade Estadual de Maringá (UEM), Núcleo de Pesquisas em Participação Política (Nuppol); TPDS-UEM. Maringá, Brasil

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

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

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 analyses 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

INTRODUÇÃO

Em novembro de 2024, Francisco Wanderley Luiz, conhecido como Tiu França, detonou artefatos explosivos diante do Supremo Tribunal Federal, em Brasília, e morreu no local. Nos dias que antecederam a ação, vinha publicando de forma constante em redes sociais mensagens que anunciavam o ato e indicavam possível premeditação (MANFRIN, 2024; FERREIRA; TREVISAN 2025). No dia seguinte, o ministro Alexandre de Moraes relacionou o atentado a um contexto mais amplo que, em sua avaliação, remontava ao chamado "gabinete do ódio" (GRANDI, 2024). O episódio condensa o problema abordado neste artigo. A explicação disponível oscila entre dois polos igualmente insuficientes: de um lado, a atribuição do ato a convicções individuais; de outro, a imputação difusa a "algoritmos" que o teriam levado a se engajar cada vez mais em conteúdo radicalizado. Nenhum dos dois, isoladamente, descreve os mecanismos que podem mediar a passagem entre o discurso e o ato.

Essa insuficiência tem lastro na teoria da radicalização. O modelo das duas pirâmides (MCCAULEY; MOSKALENKO, 2017) distingue a radicalização de opinião da radicalização de ação e sustenta, com apoio empírico consolidado, que as duas não coincidem, pois a maioria dos que sustentam ideias radicais nunca age, e parte dos que agem não partilha convicções particularmente radicais. É precisamente nessa distância entre opinião e ação — que o modelo demonstra não ser automática — que se abre espaço analítico para investigar mecanismos mediadores adicionais. Entre eles, este artigo torna analisável a camada classificatória produzida pelas infraestruturas digitais. Importante ressaltar a advertência de que o próprio conceito de radicalização carrega pressupostos securitários e tende a individualizar processos estruturais (KUNDNANI, 2015), o que recomenda tratá-lo como categoria analítica, e não como diagnóstico.

Entre esses polos há uma camada que raramente é nomeada: a infraestrutura que classifica sujeitos, conteúdos e coletivos, e que decide, a partir dessa classificação, o que cada um verá, a quem será recomendado, o que será moderado e quem será tratado como suspeito. É essa camada que o conceito de meta-identidade procura tornar analisável.

O argumento aqui desenvolvido é que a circulação online de repertórios extremistas não pode ser analisada apenas pela exposição a conteúdos extremos, mas também pelos ecossistemas classificatórios que estabilizam afinidades, públicos, sinais de engajamento e categorias de risco. A pergunta que organiza o texto é: de que modo a literatura documenta classificações que podem

favorecer, limitar ou reorganizar a circulação de repertórios extremistas, autoritários e antidemocráticos?

Três esclarecimentos delimitam o alcance da proposta. Primeiro, não se afirma que sistemas de recomendação causam radicalização. O deslocamento proposto é outro: da pergunta sobre o efeito do algoritmo à pergunta sobre o constructo classificatório atribuído e sobre o tratamento diferencial que ele autoriza. Segundo, a responsabilidade discutida aqui decorre menos da intenção do que da posição ocupada, pois ao exercer controle unilateral sobre parâmetros com efeitos socialmente relevantes isso implica em responder também por suas consequências, ainda que não antecipadas ou não desejadas. Terceiro, o artigo não dispõe de acesso ao interior de sistema algorítmico algum, e não o reivindica; opera sobre o que é observável de fora da caixa preta ou da caixa cinza.

A aplicação do mesmo constructo à contestabilidade das decisões automatizadas na administração pública, incluindo o problema da correspondência entre o critério operante e o registro documental, é desenvolvida separadamente em Shirakura et al. (2026). O presente artigo concentra-se nas plataformas digitais e mobiliza a meta-identidade como grade de leitura da literatura sobre extremismo político, recomendação, moderação e operações de influência.

O texto está organizado em cinco seções. A primeira delimita o constructo e sua arquitetura. A segunda propõe uma divisão do trabalho classificatório entre autoridades que parametrizam e a arena em que essa parametrização se torna ou não contestável. A terceira descreve o procedimento de síntese de evidências, a quarta apresenta o que a literatura recuperada documenta e a quinta discute as implicações democráticas.

1. META-IDENTIDADE: O CONSTRUCTO E QUEM O PARAMETRIZA

1.1. Delimitação

Meta-identidade é um constructo classificatório sociotécnico — inferido, agregado, opaco e operacional — produzido e controlado por plataformas digitais e sistemas de inteligência artificial sobre entidades que circulam em suas infraestruturas, e não pelos alvos delas, seus usuários, autores e produtores. Resulta do cruzamento contínuo de dados pessoais, contextuais, comportamentais e relacionais e opera como referência decisória que modula visibilidade, acesso, reputação, associação e circulação, com efeitos materiais dentro e fora do ambiente digital (FERREIRA; TREVISAN, 2025; FERREIRA ET AL., 2026).

Dois termos constituem o vocabulário técnico empregado neste estudo. “Portador” designa a entidade sobre a qual recai a classificação — um sujeito, uma conta, um grupo, uma organização, um território ou um conteúdo/artefato —, e não apenas o usuário individual de uma plataforma. “Predicado” designa o rótulo classificatório atribuído ao portador, como “propenso a engajamento com discurso de ódio”, “conta de risco” ou “conteúdo extremista”.

Quatro características são constitutivas do conceito. A primeira é a inferência: a classificação é deduzida de padrões e parâmetros analíticos, e não simplesmente declarada pelo portador. A segunda é a agregação: a classificação resulta da combinação de múltiplas fontes, incluindo interações, comportamentos, transações, parametrizações internas e categorias institucionais externas. A terceira é a opacidade: os critérios e mecanismos classificatórios não são plenamente acessíveis ao portador nem integralmente submetidos à contestação ou à auditoria. A quarta é a operacionalidade: a classificação produz efeitos concretos sobre circulação, acesso, visibilidade, reputação, associação ou responsabilização.

Para localizar como esses processos, ainda que não diretamente relacionados a esta terminologia, mas sim aos seus processos, práticas e efeitos, aparecem na literatura, esta síntese emprega cinco critérios complementares de observação: atribuição, quando existe um predicado aplicado a um portador determinado; inferência, quando o predicado deriva do processamento de sinais; estabilização, quando persiste ou se reaplica para além do episódio que o gerou; operacionalidade, quando serve de referência para tratamento diferencial; e exterioridade, quando o portador não dispõe de controle substantivo sobre a produção e a revisão do predicado. Esses cinco critérios não substituem as quatro características constitutivas do conceito. Eles organizam aquilo que pode ser rastreado em estudos realizados, em sua maioria, do lado de fora das infraestruturas. A codificação, portanto, não testa a presença integral do constructo em cada estudo, mas localiza traços empiricamente observáveis de diferentes segmentos do circuito classificatório descrito pelo framework.

Exterioridade e opacidade designam assimetrias distintas e complementares. A exterioridade diz respeito à distribuição da autoridade: o portador não controla substantivamente a produção ou a revisão da classificação. A opacidade diz respeito à distribuição do conhecimento: o portador não possui acesso suficiente ao predicado, aos dados, aos critérios, aos pesos e aos objetivos que orientam sua produção. Uma classificação pode tornar-se parcialmente transparente e continuar exterior ao

portador; do mesmo modo, a possibilidade de recorrer de uma decisão não torna necessariamente seus critérios transparentes.

A delimitação distingue o constructo do que lhe é vizinho. A digital persona de Clarke (1994) designa um modelo do indivíduo construído a partir de dados e mobilizado como seu representante em processos informacionais. A identidade algorítmica (CHENEY-LIPPOLD, 2011, 2017) estabelece que categorias como gênero, classe ou cidadania são reconstituídas probabilisticamente a partir do comportamento, mas permanece centrada no sujeito de dados e não alcança com a mesma facilidade a classificação dos artefatos, coletivos ou outros objetos com os quais ele se relaciona. O *data double* (HAGGERTY; ERICSON, 2000) descreve a remontagem de rastros dispersos em duplicatas informacionais utilizáveis por aparatos de vigilância, sem explicitar como essas representações passam a produzir diferenças de tratamento ou efeitos decisórios. O *digital twin*, por sua vez, designa uma contraparte digital de um objeto, processo ou sistema, idealmente mantida por fluxos contínuos de dados para fins de monitoramento, simulação ou previsão (GRIEVES; VICKERS, 2017; SCHROEDER et al., 2021). Sua lógica pressupõe uma aspiração de correspondência entre o modelo digital e aquilo que ele representa. Nas plataformas digitais, contudo, a representação não apenas pode ser parcial: sua própria finalidade pode dispensar a completude, selecionando e inferindo apenas os atributos relevantes para os objetivos do sistema. Em contextos de mediação algorítmica, essas representações tendem, além disso, a ser probabilísticas, opacas e funcionalmente orientadas, sem que sua incompletude, ou eventuais distorções, impeçam que produzam efeitos concretos sobre os sujeitos, conteúdos ou objetos classificados. A meta-identidade parte precisamente desse deslocamento: interessa menos a intenção de fidelidade da representação ao referente do que o modo como uma classificação, mesmo incompleta e inferida de rastros heterogêneos, torna-se operacional. Sua parcialidade tampouco deve ser tratada apenas como limitação técnica, pois pode resultar de escolhas institucionais sobre quais dados coletar, quais atributos inferir, quais categorias estabilizar e quais objetivos otimizar. A distinção central, portanto, está na assimetria classificatória entre quem define e controla a representação e quem é por ela tornado legível, sobretudo quando essa representação passa a orientar decisões sobre visibilidade, recomendação, moderação, acesso ou tratamento diferencial. Enquanto o *digital twin* enfatiza a correspondência entre modelo e referente, a meta-identidade enfatiza a incompletude, a opacidade e a eficácia operacional de classificações que podem produzir consequências sobre terceiros sem que estes conheçam, controlem ou consigam contestar os critérios pelos quais foram tornados legíveis. A situação classificatória, ou *classification situation* (FOURCADE; HEALY, 2013), mostra como instrumentos de classificação e avaliação

estruturam chances de vida, tomando escores atribuídos a indivíduos como elementos capazes de condicionar seu acesso a bens, recursos e oportunidades. Muito antes do surgimento das plataformas digitais, Bowker e Star (1999) demonstraram que sistemas classificatórios — como taxonomias médicas, classificações raciais, censos ou nomenclaturas técnicas — nunca são neutros: incorporam escolhas normativas, sedimentam-se em infraestrutura e tendem a tornar-se invisíveis enquanto funcionam. Esses conceitos, entretanto, não foram formulados para acompanhar conjuntamente a produção da classificação, sua integração algorítmica, seus efeitos operacionais e a eventual retroação dos classificados sobre os sinais que alimentam novas classificações.

A fronteira do constructo não é determinada pela natureza técnica do objeto envolvido, mas pela função que ele ocupa no circuito classificatório. Nem todo dado, conteúdo, interface ou operação submetidos a processamento algorítmico constituem, por isso, uma meta-identidade. O constructo se configura quando sinais relativos a um portador são processados de modo a produzir ou atualizar um predicado inferido que passa a servir de referência para tratamento diferencial. Elementos que não participam dessa relação podem funcionar como insumos, portadores, interfaces ou outputs do sistema sem se confundirem com a meta-identidade; e um mesmo elemento pode assumir funções distintas conforme sua inserção no circuito. A delimitação é, portanto, funcional e relacional: o que caracteriza a meta-identidade não é a mera presença de mediação algorítmica, mas a transformação de sinais sobre um portador em uma classificação inferida e operacionalmente consequente.

A meta-identidade, por sua vez, propõe uma unidade de análise voltada precisamente a esse circuito sociotécnico. Ela não se restringe ao indivíduo: sujeitos, conteúdos, coletivos, organizações, obras, lugares ou conceitos podem tornar-se objetos de classificação operacional. Permite distinguir os sinais e critérios produzidos por atores humanos e instituições de sua integração computacional, acompanhar a classificação ao longo do tempo e observar os diferentes outputs pelos quais ela se torna consequente — recomendação, ordenamento, visibilidade, moderação, segmentação, monetização ou restrição. Assim, torna possível seguir um mesmo constructo desde as autoridades, regras e infraestruturas que tornam determinada classificação possível até os efeitos que ela produz, inclusive quando os próprios classificados aprendem a manipular, contornar ou contestar os sinais dos quais essa classificação depende.

1.2. Dinâmicas, estados, modalidades temporais e outputs

A arquitetura da meta-identidade, exposta introdutoriamente em Ferreira e Trevisan (2025) e aplicada empiricamente em Ferreira et al. (2026), distingue quatro planos. Retoma-se aqui a formulação por remissão, sem reexposição integral.

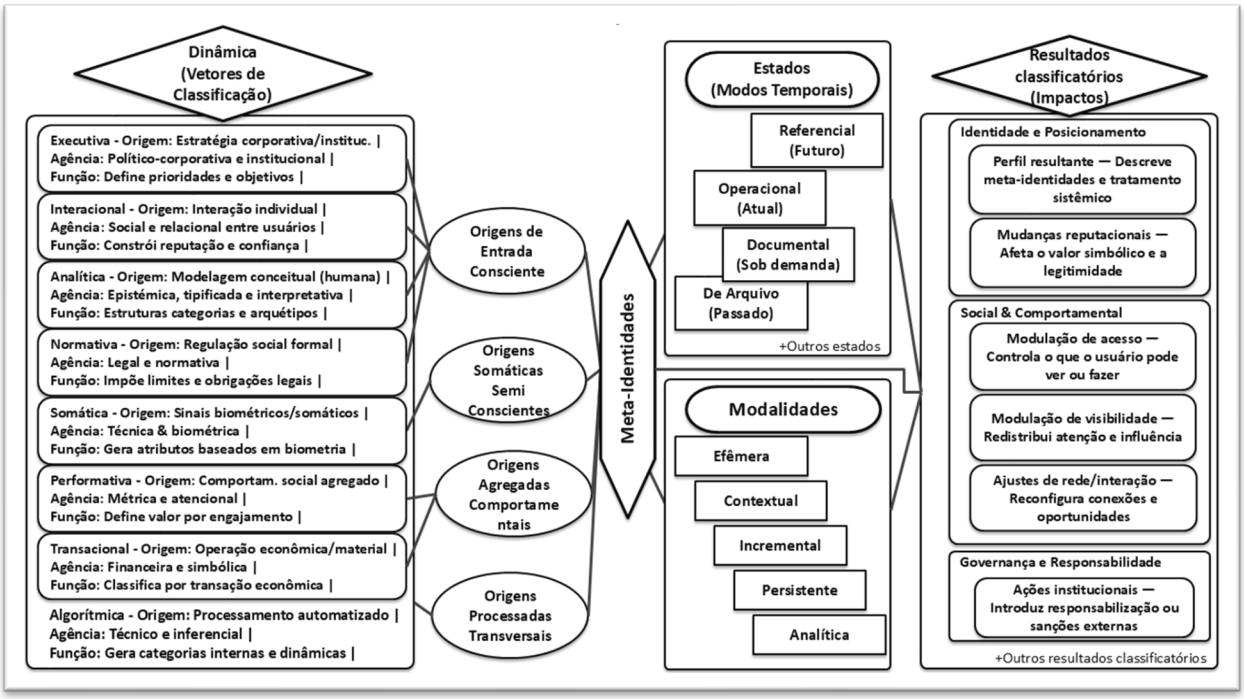
As dinâmicas da meta-identidade indicam de onde vêm os vetores de classificação. Sete operam como vetores de entrada — executiva, interacional, analítica, normativa, somática, performativa e transacional — e por fim, a oitava dinâmica, a algorítmica, integra e compõe as demais. A distinção impede que se atribua ao processamento automatizado uma autonomia que ele não tem e localiza a agência nos responsáveis por sua parametrização.

Os estados indicam em que condição a meta-identidade atua: operacional, o presente ativo que modula acesso e visibilidade; de arquivo, o passado armazenado, disponível para comparação, reativação ou treinamento; referencial, o futuro projetado, isto é, o molde de perfil desejado para o qual se induz o portador; e documental, o registro formalizado sob demanda para auditoria, prestação de contas ou uso jurídico. O estado referencial é o mais consequente, porque desloca a classificação do diagnóstico para a indução: não se trata de futuro provável, mas de futuro desejado por quem controla a infraestrutura. A lógica aproxima-se do *hypernudge* de Yeung (2017), no qual a análise contínua de dados permite reorganizar o contexto informacional e canalizar atenção e decisão em direções preferidas por quem projeta a arquitetura de escolha. Distingue-se, porém, por tomar como objeto analítico não apenas a condução da escolha, mas o próprio estado-alvo classificatório em relação ao qual o sujeito, conteúdo ou coletivo é continuamente situado e tratado

As modalidades indicam em que temporalidade e escopo a meta-identidade está sendo constituída, segundo sua duração, densidade informacional, estabilidade, finalidade operacional e grau de institucionalização. São elas: a modalidade efêmera, constituída durante uma sessão ou interação específica; a contextual, vinculada a determinada situação, tema ou conjuntura; a incremental, progressivamente ajustada pelo acúmulo ponderado de sinais; a persistente, estabilizada como referência duradoura para reputação, segmentação, acesso ou risco; e a modalidade analítica, construída intencionalmente por pesquisadores, auditores ou reguladores como instrumento de escrutínio das próprias meta-identidades. As quatro primeiras podem coexistir em camadas e seguir uma trajetória típica, mas não necessária, do efêmero ao persistente. A modalidade analítica possui estatuto distinto: a categorização permanece com quem investiga, em vez de ser incorporada pela plataforma como parâmetro classificatório. É nesse sentido que ela descreve o procedimento deste artigo, sendo necessário frisar a diferença conceitual previamente explicada entre a modalidade e a

dinâmica analítica: a dinâmica analítica é a categorização incorporada pela plataforma como parâmetro; a modalidade analítica é a categorização retida pelo pesquisador, auditor ou regulador como instrumento de escrutínio.

Figura 1 — Arquitetura das dinâmicas, estados e resultados classificatórios da meta-identidade



Nota: a figura sintetiza os vetores de classificação (dinâmicas), os modos temporais de operação (estados), as formas de constituição (modalidades) e os efeitos produzidos (resultados classificatórios) que compõem a arquitetura da meta-identidade analisada no artigo. Fonte: Adaptado de Ferreira et al., 2026.

Os outputs organizam-se em três eixos: identidade e posicionamento, que alteram a reputação e a legitimidade; social e comportamental, que modulam o acesso, a visibilidade, as redes e as oportunidades; e governança e responsabilização, que produzem sanções, bloqueios, desmonetização, banimento e encaminhamentos institucionais.

2. A DIVISÃO DO TRABALHO CLASSIFICATÓRIO

Atribuir a uma entidade abstrata chamada "algoritmo" os danos decorrentes do tratamento diferencial que a meta-identidade autoriza — sanções, bloqueios, desmonetização, banimento, exposição ampliada a repertórios extremistas — é insuficiente: não há, por trás do sistema, um agente autônomo dotado de vontade própria, mas sistemas heterogêneos e diversos, trazidos à existência

pelas práticas de quem os projeta e de quem os usa (SEAVER, 2017). O que se manifesta como comportamento emergente do sistema — inclusive quando produz efeitos que nenhum indivíduo escolheu ou autorizou especificamente — decorre de decisões humanas tomadas em camadas anteriores: qual objetivo otimizar, com quais dados, sob quais restrições — decisões que se tornam, na prática operacional do sistema, opacas mesmo para quem o projeta, dada a escala em que esses algoritmos precisam operar (BURRELL, 2016). Isso não deriva de segredo institucional, mas das próprias características técnicas do aprendizado de máquina — e é justamente por isso que responsabilizar o sistema isoladamente é insuficiente: a responsabilização precisa alcançar a rede de decisões e atores que o constitui, não apenas o artefato final (ANANNY & CRAWFORD, 2018). Para entender esses desdobramentos, o conceito de meta-identidades descreve diferentes tipos de dinâmicas internas aos mecanismos que operam no interior das plataformas digitais.

A dinâmica executiva traduz prioridades estratégicas em parâmetros. Ela se manifesta em três níveis encadeados: decisões de direção que fixam metas; sua tradução operacional em indicadores e diretrizes de produto; e a recalibração algorítmica que delas decorre. Weber (2004), ao tratar do diletante, reconhece que este pode ser mais produtivo que o especialista, mas questiona sua capacidade de avaliar as implicações do que produz — quem define objetivos raramente domina o trabalho técnico que os implementa, e quem implementa raramente decide sobre as finalidades. Mas o ponto sociológico não é a intenção de quem decide, nem sua capacidade de antever cada resultado específico: é a posição que ocupa. A responsabilidade não depende de prever o resultado técnico exato — depende de ocupar o lugar de onde se define, a montante, o que o sistema deve otimizar. É essa posição, não a previsão, que a opacidade técnica descrita anteriormente não dissolve.

A dinâmica analítica, por sua vez, reúne as classificações socialmente produzidas que são incorporadas aos sistemas: tipologias, arquétipos, categorias censitárias, tipificações clínicas, jurídicas e criminológicas, além da modelagem interna de equipes de dados e de *marketing*. Antes que um sistema rotule alguém como suscetível ao extremismo político, tal noção, categoria ou tipificação já circulou em artigos, relatórios, práticas profissionais e em outras interações formais ou informais — e pode depois ser absorvida como parâmetro de segmentação.

Daí decorre a tese específica desta seção: a dinâmica executiva define para onde o sistema deve tender; a dinâmica analítica define com que categorias ele o fará. É nesta segunda operação que parte relevante do conteúdo normativo pode entrar no sistema — sob a forma de escolha técnica, o que tende a torná-la menos visível à discussão pública.

A consequência é reflexiva e vale registrar. As categorias com que a comunidade acadêmica descreve o extremismo digital são um insumo potencial para a mesma infraestrutura que ela analisa. Isso não estabelece equivalência de poder entre quem propõe uma tipologia e quem define um parâmetro; estabelece que a autoridade epistêmica também produz efeitos classificatórios e que a responsabilidade acompanha a autoridade.

A dinâmica normativa não parametriza no mesmo sentido da executiva e da analítica. Ela pode servir ao interesse público, por meio de representação legislativa e mobilização social, ou aos interesses de quem conduz a dinâmica executiva, por meio de pressão organizada e de desenho regulatório favorável. É, portanto, o campo em que se decide se a parametrização das outras dinâmicas será auditável, explicável e recorrível.

Tratada como arena, a dinâmica normativa pode ser pensada como o lugar onde a disputa democrática efetivamente ocorre. É também onde se torna visível uma lacuna que a seção 4 documenta: o debate regulatório recuperado concentra-se em quem opera a infraestrutura classificatória, enquanto sua instrumentalização por atores externos aparece muito menos desenvolvida.

3. PROCEDIMENTO DE SÍNTESE DE EVIDÊNCIAS

3.1. Desenho e princípio de recuperação

O procedimento metodológico adotado é uma síntese estruturada de evidências com codificação em duas passagens, e não uma revisão sistemática no sentido protocolar do termo: não houve registro prévio de protocolo nem dupla triagem integral, e isso é declarado como limitação.

Alguns termos desta seção vêm da tradição de síntese de evidências e merecem definição breve. Semente é uma obra previamente conhecida, escolhida por sua centralidade no debate, que serve de ponto de partida para o rastreamento de citações; ela não entra no corpus por busca, mas por decisão declarada. O rastreamento tem duas direções: ascendente, que segue as referências citadas pela semente, e descendente, que segue os trabalhos que a citam. Sementes de calibragem são obras que foram incorporadas durante a verificação das sementes iniciais, e não no desenho original; a distinção importa porque o rendimento que produzem não pode ser atribuído ao desenho inicial. Grafo de citações é a rede formada pelas citações de autores (quem cita quem) mantida por um índice bibliográfico — cada índice tem cobertura própria, e nenhum é completo, do que decorre o viés de veículo, a distorção produzida quando um índice cobre desigualmente diferentes tipos de publicação.

Ganho marginal exclusivo é o número de trabalhos inéditos que uma fonte ou semente acrescenta depois de descontado tudo o que as demais já haviam trazido, e existe para impedir dupla contagem. Pela mesma razão distingue-se ocorrência de documento distinto: um trabalho que cita três sementes é recuperado três vezes, de modo que a soma das recuperações não equivale ao número de trabalhos. Distingue-se ainda registro, a entrada bibliográfica — *preprint* e versão publicada são dois registros — , de estudo, a unidade intelectual que os reúne e que é a unidade de análise. Teste de sensibilidade é a medição de quanto um resultado muda quando se altera uma decisão do procedimento mantendo as demais constantes. E corpus congelado designa o momento a partir do qual nada mais é incorporado: obras encontradas depois vão para registro separado e só entram mediante regra explícita.

O princípio que definiu a busca foi o seguinte: não se procurou o conceito de meta-identidade, mas sim os processos e fenômenos que ele prevê, de acordo com a terminologia que consta em sua literatura original. Nenhuma expressão de busca contém "meta-identidade" ou suas variantes. Sem essa regra, a síntese seria circular — o conceito encontraria apenas os textos que já adotam o termo e conceito.

A recuperação partiu de quatro blocos combinados a um bloco adicional de objeto político (extremismo, radicalização, direita radical, autoritarismo, discurso de ódio): inferência e perfilamento; distribuição e recomendação; restrição e moderação; e ação coordenada sobre o sistema. O recorte temporal é 2016–2026, tomando a eleição estadunidense de 2016 como o momento em que a classificação realizada por plataformas se torna um problema político público. A fonte primária foi a base de dados aberta OpenAlex (PRIEM et al., 2022), complementada pelo rastreamento de citações.

3.2. Dois achados de calibragem

Duas medições realizadas antes de qualquer leitura substantiva produziram achados que interessam por si.

O primeiro é a disjunção de vocabulários. Mantidas constantes a consulta sobre ação coordenada e a restrição de plataforma, o vocabulário de extremismo recuperou 54 registros e o vocabulário de operações de influência recuperou 178. São dois vocabulários que recuperam de forma fortemente desigual trabalhos sobre momentos relacionados do mesmo circuito. A ação organizada sobre a infraestrutura classificatória é recuperada com muito maior frequência sob a rubrica de operações de influência do que sob a de extremismo — o que ajuda a explicar por que o campo do extremismo digital raramente a tematiza.

O segundo é a subcobertura seletiva do grafo de citações. O rastreamento descendente parte de um conjunto de obras-semente e segue os trabalhos que as citam. As seis sementes iniciais foram escolhidas por cobrirem deliberadamente os dois lados da controvérsia sobre radicalização algorítmica, e não apenas o lado que sustenta a tese: Ribeiro et al. (2020), que documentam trilhas de migração para canais extremos no YouTube; Ledwich e Zaitsev (2020), que encontram o resultado oposto, com o algoritmo favorecendo conteúdo de veículos consolidados; Hosseinmardi et al. (2021), que não encontram evidência de que a recomendação dirija a atenção ao conteúdo radical; Munger e Phillips (2022), que deslocam a explicação da oferta algorítmica para a demanda do público; Haroon et al. (2023), que encontram recomendação congenial crescente ao longo da trilha; e Piccardi et al. (2025), cujo experimento de reordenação de feed mede efeito causal sobre a polarização afetiva. Um conjunto de sementes composto apenas pelos estudos que confirmam a tese produziria um corpus enviesado por construção.

A conferência dessas sementes — processo destinado apenas a verificar se o casamento por título estava correto — revelou contagens de citação implausíveis. Ribeiro et al. (2020) figurava com 48 trabalhos citantes indexados e Ledwich e Zaitsev (2020) com 14, ordens de grandeza abaixo do que outros índices registram para obras dessa centralidade. A verificação descartou a hipótese de registro duplicado e identificou subcobertura do grafo do OpenAlex para dois tipos de veículo: anais de conferência de computação e periódicos de acesso aberto independentes.

A mesma consulta de verificação teve efeito colateral produtivo. Ao devolver, junto de cada semente, suas vizinhas mais próximas no índice, expôs cinco obras que satisfaziam o critério de inclusão, recorriam na vizinhança de mais de uma semente e não constavam do conjunto inicial — sinal de centralidade que o desenho original havia deixado passar. Foram promovidas a sementes, com a rodada de entrada registrada no fluxo de seleção: Whittaker et al. (2021), sobre sistemas de recomendação e amplificação de conteúdo extremista; Yesilada e Lewandowsky (2022), revisão sistemática sobre recomendação e conteúdo problemático no YouTube; Shaw (2023), síntese sobre mídias sociais, extremismo e radicalização; Bouchaud e Ramaciotti (2024), exame metodológico das próprias auditorias de sistemas de recomendação; e Gauthier et al. (2026), sobre os efeitos políticos do algoritmo de feed do X.

O rastreamento foi então refeito sobre o Semantic Scholar, cujo grafo cobre melhor esses veículos, mantendo o OpenAlex como fonte de metadados para preservar a homogeneidade do corpus. Como essas mudanças simultâneas — a fonte de citantes e o conjunto de sementes —, os

efeitos foram decompostos. Mantidas constantes as seis sementes iniciais, o OpenAlex devolveu 14 trabalhos citantes elegíveis e inéditos, e o Semantic Scholar devolveu 171. As cinco sementes da calibragem contribuíram com 25 registros exclusivos, isto é, já descontados os que as seis originais haviam alcançado. O universo inédito combinado foi de 196 registros, dos quais 186 foram incorporados após recuperação de metadados e filtros de tipo e idioma. O ganho é atribuível sobretudo à cobertura da fonte, e não à ampliação das sementes.

A implicação metodológica ultrapassa este artigo, e convém explicitá-la. Índices bibliográficos não observam a literatura diretamente: reconstroem o grafo de citações a partir do que é indexado, e essa cobertura pode variar entre tipos de veículo. No conjunto examinado, a subcobertura observada incidiu especialmente sobre anais de conferência de computação e periódicos de acesso aberto independentes, veículos importantes para o núcleo empírico da pesquisa sobre auditoria algorítmica. O padrão é, portanto, compatível com uma distorção de cobertura alinhada à divisão epistêmica do campo. Uma síntese que dependa de um único grafo pode sub-recuperar parte da produção empírica e produzir uma imagem enviesada da literatura. É essa sensibilidade sistemática da recuperação ao tipo de veículo e ao grafo utilizado que se designa aqui por viés de veículo, recomendando que qualquer síntese sobre este objeto declare o grafo utilizado e, quando possível, o submeta a teste de sensibilidade contra um segundo índice.

3.3. Composição do corpus e regra de fronteira

O corpus foi congelado em 813 registros bibliográficos, consolidados em 749 estudos. A redução de 64 registros decorre de 62 grupos de versões: 60 grupos formados por dois registros e dois grupos formados por três registros. *Preprints* e versões publicadas do mesmo trabalho foram triados separadamente, mas contabilizados como uma única unidade de evidência. Foram descartados, com registro, 12 itens de paratexto, seis artefatos suplementares, duas duplicatas de identificador e treze registros cuja elegibilidade decorria de erro de metadado do indexador, que havia atribuído o resumo de um estudo de auditoria a artigos sem relação.

A regra que embasou a decisão de inclusão foi a semântica: incluíram-se os estudos cujos achados dizem algo sobre como um sistema de plataforma infere, ordena, distribui, restringe ou monetiza — ou sobre como atores agem para alterar esses sinais; excluíram-se os estudos cujos achados dizem respeito ao desempenho de um instrumento construído pelos próprios pesquisadores.

Durante a pesquisa, foram preservados 145 arquivos de texto integral obtidos por vias legítimas de acesso. Após a consolidação das versões e a correção dos metadados, 136 arquivos correspondiam a registros do corpus atual e representavam 127 estudos distintos; nove arquivos remanescentes de etapas anteriores foram mantidos exclusivamente para auditoria. A localização assistida por regras léxicas auditáveis foi aplicada aos 127 estudos com texto recuperado e identificou ao menos um candidato em 118 deles. Para cada variável, foram preservados o candidato, a frase literal e a seção em que ocorria. Frases que continham marcadores de citação bibliográfica foram sinalizadas porque poderiam relatar achados de terceiros, e não do estudo codificado. O corpus, os scripts, os protocolos e os materiais de auditoria correspondentes, incluindo as regras e os procedimentos de conferência empregados na localização assistida, estão depositados no Zenodo por Ferreira, Trevisan e Shirakura (2026).

3.4. Escala de base de atribuição

Uma dificuldade recorrente na literatura sobre algoritmos é inferir a arquitetura interna a partir de outputs. Para tratá-la, distingue-se a base sobre a qual a operação é atribuída, em cinco níveis: documentada pela própria plataforma, em relatório de transparência, documentação técnica, código aberto ou submissão regulatória; observada em operação, mediante auditoria com contas sintéticas ou experimento; inferida de padrão de output, em dados de rastro ou análise de rede; inferida de relato de usuários e criadores; e postulada como premissa teórica.

Este eixo não coincide com a força do desenho. Uma auditoria com contas sintéticas é um desenho robusto e acesso nulo; a documentação de um sistema de agrupamento publicada pela própria plataforma tem acesso máximo e não é, ela mesma, um estudo empírico.

4. O QUE A LITERATURA DOCUMENTA

Os resultados desta seção decorrem de uma localização assistida por regras aplicada aos 127 estudos com texto integral, dos quais 118 apresentaram ao menos um candidato. Os estudos recuperados observam diferentes segmentos do circuito classificatório; a meta-identidade é mobilizada aqui como grade de integração dessas evidências parciais, e não como variável causal diretamente testada por cada estudo. Os trabalhos discutidos nominalmente a seguir são casos ilustrativos selecionados para tornar visíveis diferentes mecanismos, desenhos e resultados encontrados no corpus, e não constituem uma segunda amostra.

4.1. Composição e o que ela já informa

A distribuição por plataforma é desigual de forma reveladora. X/Twitter aparece em 63 estudos; YouTube, em 36; Facebook, em 35; TikTok e Instagram, em 16 cada; Reddit, em 13; WhatsApp, em 10; Telegram, em 5; Gab, em 4; BitChute, em 3; Twitch, em 2. A ordem não mede peso político, mas sim o acesso a dados. X/Twitter ofereceu historicamente acesso relativamente amplo a dados públicos e APIs, o que favoreceu sua presença na pesquisa acadêmica; TikTok e Telegram, politicamente centrais em vários contextos, são residuais. A composição do corpus é, ela mesma, um efeito do regime de acesso que o artigo analisa.

Quanto ao portador da classificação, usuário ou conta aparece em 90 estudos; coletivo ou rede, em 81; obra ou artefato, em 80; tópico e organização, em 75 cada; e território, em 38. A presença do território é relevante porque o conceito reivindica alcance sobre entidades não humanas e sobre produções, e o campo o documenta mais do que se supunha: há estudos que registram tratamento diferencial por região e por país.

4.2. Inferência, distribuição e a controvérsia da radicalização

A literatura de auditoria de sistemas de recomendação constitui o núcleo de evidência mais robusto do corpus, e é também a mais dividida.

Huszár et al. (2022) conduziram, ainda no Twitter (renomeado X em 2023), um experimento randomizado em larga escala com um grupo de controle de quase dois milhões de contas ativas diárias submetidas a uma linha do tempo cronológica reversa, cobrindo mensagens de legisladores eleitos em sete países. Encontraram amplificação consistente da direita majoritária em seis dos sete casos — e, contra a expectativa corrente, não encontraram vantagem para a extrema-direita frente à direita majoritária. É o único caso do corpus em que a plataforma audita a si própria e publica o resultado, algo que deixou de ocorrer após a aquisição da plataforma por Elon Musk em 2022.

Haroon et al. (2023) auditaram o sistema de recomendação do YouTube com cem mil contas sintéticas e encontraram recomendação congenial crescente à medida que se avança na trilha, sobretudo para perfis à direita, sem que isso equivalha a uma progressão necessária da extremidade ideológica do conteúdo recomendado. Tomlein et al. (2021) e Srba et al. (2023) mostraram que bolhas de desinformação se formam rapidamente, mas que também podem ser revertidas quando o perfil passa a consumir conteúdo de desmentido — resultado que interessa diretamente à gradação entre modalidades efêmera e persistente. Lahsaini et al. (2024), auditando recomendações sobre aborto em

seis perfis, encontraram resultado contrário ao esperado: escore negativo em todos os perfis, isto é, predominância de conteúdo favorável ao aborto e de desmentido.

Ribeiro et al. (2023) oferecem a síntese que impede uma conclusão fácil. Diante do contraste entre auditorias que encontram viés e dados de uso que não encontram consumo correspondente, mostram por simulação que a natureza colaborativa dos sistemas de recomendação e o caráter de nicho do conteúdo extremo bastam para dissolver o paradoxo: o algoritmo pode favorecer sem dirigir, porque oferta algorítmica e consumo efetivo não são equivalentes. Hilbert et al. (2024), em meta-análise de 151 auditorias extraídas de 33 estudos, estimam que 8% a 10% das recomendações são prejudiciais, sem diferença estatisticamente significativa entre plataformas nem entre tipos de risco.

O conjunto sustenta o deslocamento proposto na introdução. Não é necessário afirmar que a recomendação radicaliza para sustentar que existem classificações atribuídas que servem de referência para tratamento diferencial — e esta segunda afirmação é observável em rastros. Recomendação, exposição, consumo, desinformação, polarização ou mudança de atitudes não são tratados aqui como medidas equivalentes de radicalização, mas como diferentes outputs ou segmentos do circuito classificatório documentados pelos respectivos desenhos.

4.3. Restrição e moderação

O segundo núcleo é o da restrição, e nele os achados mais fortes são de ausência de efeito.

Tonneau et al. (2026) realizaram auditoria global da moderação de discurso de ódio no Twitter/X a partir de amostra representativa construída sobre captura integral de 24 horas de mensagens públicas, com 540 mil mensagens anotadas por avaliadores treinados em oito idiomas. Cinco meses após a publicação, 80% das mensagens de ódio permaneciam disponíveis, inclusive incitação explícita à violência, e não apresentavam probabilidade de remoção superior à de mensagens não ofensivas: nem a severidade nem a visibilidade aumentaram a chance de remoção.

Shukla et al. (2025) auditaram o sistema automatizado de moderação do Twitch mediante contas de transmissão criadas como ambientes isolados, enviando mais de 107 mil comentários provenientes de quatro conjuntos de dados. Até 94% das mensagens abertamente ofensivas escaparam à moderação em alguns conjuntos; a adição contextual de termos-gatilho elevou a detecção a 100%, revelando dependência lexical; e até 89,5% dos exemplos benignos, que empregavam termos sensíveis em contexto pedagógico ou afirmativo, foram bloqueados. O estudo documenta atribuição, inferência

e operacionalidade, mas não permite estabelecer de maneira inequívoca a estabilização do predicado para além do evento de moderação.

Hartmann et al. (2025) deslocam o portador. Auditaram cinco interfaces comerciais de moderação, analisando cinco milhões de consultas, e mostraram que os serviços frequentemente decidem a partir de termos de identidade de grupo, submoderando discurso de ódio implícito — sobretudo contra pessoas LGBTQIA+ — e sobremoderando contradiscurso e termos reapropriados. O classificador, aqui, não pertence à plataforma: é uma infraestrutura vendida a plataformas, o que estende o alcance analítico do constructo a terceiros associados.

Gennaro et al. (2025) documentam a distribuição: em cinco conjuntos originais, coletados em múltiplas plataformas na Suíça e nos Estados Unidos, a maior parte do discurso de ódio provém de fração reduzida de usuários; e experimento de campo pré-registrado indica que estratégias de contradiscurso produzem reduções pequenas, sendo ineficazes justamente contra os produtores mais prolíficos.

4.4. O acionamento adversarial das dinâmicas de entrada

O acionamento adversarial é o que mais interessa à pergunta do artigo e o menos documentado. Trata-se da ação deliberada de atores políticos organizados sobre os vetores de entrada — sobretudo as dinâmicas interacional, performativa e transacional — com a finalidade de alterar a meta-identidade atribuída. A ação assume três modos, e a distinção é decisiva: autorreferente, quando visa a própria conta ou o próprio conteúdo, mediante engajamento fabricado e inflação de alcance; heterodirigida, quando visa terceiros, mediante denúncia coordenada, assédio e fabricação de sinal de risco sobre adversários; e ambiental, quando visa o espaço informacional, por inundação e manipulação de tendências.

Somente o modo heterodirigido sustenta a proposição específica de que atores organizados podem agir sobre sinais dirigidos a terceiros, e agregar os três produziria falsa confirmação, porque a literatura sobre engajamento fabricado é abundante e quase toda autorreferente. Nesse modo, a evidência permite identificar a produção deliberada de sinais dirigidos a terceiros; isso não implica, salvo quando diretamente observado, demonstrar a classificação interna posteriormente produzida pela plataforma. Na localização assistida, o modo autorreferente aparece em 11 estudos, o heterodirigido em sete e o ambiental em quatro; em 99 dos 118 estudos não há candidato de qualquer

modo. Os modos não são mutuamente exclusivos no nível do estudo, razão pela qual suas frequências somadas excedem os 19 estudos em que apareceu ao menos um candidato.

O caso mais completo é o de Recabarren et al. (2023). Os autores realizaram 19 entrevistas semiestruturadas com participantes de operações de influência que atuam sobre a imagem da Venezuela e validaram parte das respostas com dados coletados durante quase quatro meses das contas que esses participantes controlam. Documentam o *Patria*, plataforma governamental em que operadores registram atividades e recebem benefícios, e sistematizam as estratégias declaradas para promover conteúdo político e para evadir-se de penalidades da plataforma. De todo o corpus, este é o registro mais direto da ação sobre o classificador descrita pelo lado do ator.

A escassez dos demais casos exige cautela. Pode indicar que o fenômeno é raro; pode indicar que ele é descrito por vocabulários distintos daqueles predominantes na literatura de extremismo, conforme a disjunção medida na calibragem; e pode ainda refletir o alcance do instrumento de localização. Essas duas últimas possibilidades não são independentes: a dispersão terminológica entre campos pode reduzir a capacidade de regras léxicas de localizar manifestações equivalentes do fenômeno. Não é possível, nesta etapa, estimar quanto cada uma dessas hipóteses contribui para a baixa frequência observada.

4.5. Que condições a literatura alcança

A codificação preliminar examinou quais dos cinco critérios complementares de observação — atribuição, inferência, estabilização, operacionalidade e exterioridade — aparecem documentados em cada estudo.

Tabela 1 — Indício de cada condição observacional nos estudos com candidatos localizados (n = 118)

Condição	Estudos	%
Estabilização	66	56
Inferência	59	50
Operacionalidade	31	26
Atribuição	28	24
Exterioridade	6	5

Fonte: elaboração dos autores com base no conjunto de dados depositado por Ferreira, Trevisan e Shirakura (2026). Nota: As frequências não são mutuamente exclusivas e representam a presença de cada condição separadamente. A ausência de candidato foi interpretada como ausência de documentação

localizada no estudo, e não como evidência de ausência da condição. Nenhum estudo apresentou indícios simultâneos das cinco condições.

A análise exploratória não identificou estudo que apresentasse indícios simultâneos das cinco condições. A interpretação desse resultado ainda não pode ser conclusiva, pois ele reflete, em proporção desconhecida, tanto a literatura recuperada quanto o alcance dos padrões de localização. Mais robusta que os valores absolutos é, por ora, a ordenação observada: a literatura documenta com maior frequência a inferência e a estabilização das classificações; com menor frequência, sua atribuição a um portador determinado e sua operacionalidade; e raramente a exterioridade — condição que evidencia a assimetria de autoridade sobre a produção e a revisão do predicado e que, articulada à opacidade, limita sua contestação. O mesmo padrão aparece nos indicadores de contestabilidade codificados nesta síntese. O recurso contra decisões automatizadas aparece em 51 estudos, e a auditoria independente, em 34; o acesso do portador às categorias que lhe foram atribuídas aparece em apenas três. Nesta etapa exploratória, a convergência entre a baixa documentação da exterioridade, a raridade do acesso às categorias e os casos localizados de ausência explícita de razões ou de revisão é tratada como hipótese interpretativa; a coocorrência entre esses indicadores não foi examinada.

Nesse ponto, a meta-identidade revela sua utilidade como teoria de médio alcance, no sentido proposto por Merton (1968), ao oferecer uma taxonomia e um vocabulário de tradução sociotécnica entre campos que frequentemente observam partes distintas do mesmo processo. Para as ciências sociais e as humanidades, essa arquitetura permite sintetizar operações classificatórias complexas sem dispensar, quando necessário, o acesso à literatura técnica específica das engenharias e da ciência de dados. Para pesquisadores e desenvolvedores das áreas técnicas, o mesmo vocabulário torna legíveis o alcance e as consequências materiais, econômicas, sociais e políticas das categorias, dos parâmetros e dos sistemas que concebem e operam. Sua função, portanto, não é substituir a especialização de cada campo, mas estabelecer uma camada comum de análise que permita relacionar mecanismos técnicos, decisões humanas e consequências sociais.

5. A DISPUTA PELO REGIME CLASSIFICATÓRIO

Três consequências decorrem do quadro descrito. A primeira é que extremismos e autoritarismos contemporâneos não circulam apenas por discursos, lideranças e organizações, mas também por infraestruturas que classificam e modulam públicos, conteúdos e redes. A meta-identidade oferece uma linguagem para descrever como interações efêmeras com conteúdos

radicalizados podem integrar classificações contextuais, incrementais ou persistentes e, assim, compor trajetórias de engajamento e exposição seletiva.

A segunda é que a assimetria fundamental não está na quantidade de dados, mas na distribuição do conhecimento sobre os critérios. A plataforma dispõe de múltiplas dimensões do portador; o portador não dispõe plenamente dos critérios que o classificam. Essa assimetria encontra tradução jurídica no debate sobre o processo decisório automatizado: Yeung e Lodge (2019) situam nesse processo parte das preocupações normativas da regulação algorítmica, enquanto Bayamlıoğlu (2022) formula mais diretamente o direito de contestação como garantia procedimental que ultrapassa o mero fornecimento de explicações. Sem acesso às categorias atribuídas nem via para revisão, o classificado não dispõe das condições procedimentais mínimas para contestar significativamente a classificação. Os dados desta síntese são compatíveis com essa assimetria, embora não a demonstrem diretamente — devido limitação ou inacessibilidade aos modelos de funcionamento das plataformas digitais. Tendo esta particularidade em consideração é possível considerar que a exterioridade é a condição menos documentada, e o acesso do portador às categorias que lhe foram atribuídas é o indicador de contestabilidade menos frequente. A convergência entre esses resultados evidencia sobretudo uma lacuna da literatura: os estudos examinam com maior frequência a decisão e seus efeitos do que o conhecimento e o controle de que dispõe o classificado sobre o processo classificatório.

A terceira consequência é regulatória e decorre da discussão sobre a dinâmica normativa apresentada na seção 2. Instrumentos como o Digital Services Act (PARLAMENTO EUROPEU; CONSELHO DA UNIÃO EUROPEIA, 2022), com obrigações de transparência, avaliação de risco sistêmico e auditoria independente para plataformas de muito grande dimensão, dirigem-se principalmente a quem opera a infraestrutura classificatória. No corpus recuperado, não foi identificado um tratamento regulatório direto da ação coordenada de atores externos destinada a induzir classificações sobre adversários. O resultado deve ser entendido como lacuna da literatura recuperada, e não como demonstração de inexistência de qualquer dispositivo jurídico aplicável.

Não se sustenta, aqui, que a classificação algorítmica determine trajetórias. Os classificados produzem os sinais, adaptam comportamentos, disputam categorias e acionam instituições — e o caso venezuelano documenta atores que aprendem sistematicamente a operar o classificador. Em outro registro, que não integrou a síntese estruturada de evidências, Zenkl (2025) mostra como trabalhadores inseridos em um regime tecnoburocrático procuram "domesticar" (*tame*) futuros algorítmicos, negociando seus limites e reivindicando margens de discricionariedade, em uma relação que não se

reduz à adesão ou à resistência. Sustenta-se algo mais modesto e mais verificável: a disputa democrática contemporânea envolve não apenas a regulação de mensagens extremistas, mas a contestação dos regimes classificatórios que definem o que será visto, recomendado, moderado, monetizado ou tornado suspeito.

Essa disputa pode ser situada na teoria da esfera pública. Habermas (2023) sustenta que a estrutura de plataforma orientada por algoritmos fragmenta a esfera pública em pseudo-públicos fechados, corroendo a formação discursiva da opinião e as instituições de que a democracia depende. A meta-identidade pode ser mobilizada como uma chave sociotécnica para analisar essa fragmentação: é pela modulação da visibilidade e da recomendação a partir de classificações inacessíveis ao classificado que a infraestrutura pode reorganizar quem fala com quem. Convém, no entanto, ressaltar que o engajamento político em rede é também afetivo (PAPACHARISSI, 2015) e agonístico (MOUFFE, 2013), de modo que a resposta democrática não se reduz a restaurar o debate racional, mas passa por tornar contestáveis os regimes que distribuem visibilidade e suspeição.

CONSIDERAÇÕES FINAIS

O artigo mobilizou a meta-identidade como grade analítica para examinar a literatura sobre extremismo político em redes digitais. A síntese não procurou trabalhos que empregassem o conceito especialmente para não incorrer em autorreferenciação no corpus analisado, mas buscou estudos dos fenômenos que ele permite descrever e analisar. A codificação admitiu a resposta "não se enquadra", e os limites do procedimento foram explicitados antes da interpretação dos resultados.

Três resultados preliminares merecem registro. A literatura sobre extremismo digital e a literatura sobre operações de influência alcançam momentos relacionados do mesmo circuito classificatório por vocabulários distintos. Entre os cinco critérios complementares de observação, inferência e estabilização aparecem com maior frequência, enquanto a exterioridade é rara; nenhum dos 118 estudos reuniu indícios simultâneos dos cinco critérios. A ação organizada sobre os vetores de entrada dirigida a terceiros, que constitui a proposição mais específica do artigo, aparece em número reduzido de estudos, permanecendo como hipótese sociotécnica com lacuna empírica declarada, e não como achado consolidado.

Algumas limitações delimitam o alcance deste estudo. Não houve acesso ao interior de sistema algorítmico algum, e a síntese opera sobre o que é observável de fora. A análise integral ficou restrita aos 127 estudos para os quais ao menos um texto completo foi recuperado, e estudos de acesso

fechado estão sub-representados. Não se presume, portanto, que esse subconjunto seja representativo do corpus bibliográfico consolidado, de modo que não se pode excluir viés de acesso. A composição do corpus por plataforma reflete o regime de disponibilidade de dados, não o peso político das plataformas.

A agenda que daí decorre é dupla. No plano da pesquisa, é necessário submeter o modo heterodirigido de acionamento adversarial a desenho empírico próprio, capaz de distinguir a ação sobre a própria classificação da ação sobre a classificação alheia. No plano público, é necessário que o debate sobre regulação de plataformas incorpore uma dimensão ainda insuficientemente tratada: não apenas o que as plataformas fazem com as classificações que produzem, mas como atores organizados podem agir sobre seus sinais para tentar induzir classificações e tratamentos algorítmicos sobre terceiros.

REFERÊNCIAS

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.

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*. Versão 1.0. [Conjunto de dados]. 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. Disponível em: https://www.gazetadopovo.com.br/republica/moraes-explosao-fato-isolado-considera-segundo-pior-apos-8-1/. Acesso em: 31 ago. 2026.

GRIEVES, M.; VICKERS, J. Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. In: KAHLEN, F.-J.; FLUMERFELT, S.; ALVES, A. (org.). *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. Tradução: 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. Disponível em: https://www.gazetadopovo.com.br/republica/autor-de-explosoes-em-brasilia-anunciou-atos-nas-redes-sociais-se-despediu-e-deixou-recado-a-pf/. Acesso em: 31 ago. 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*. Ed. ampliada. 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.

PARLAMENTO EUROPEU; CONSELHO DA UNIÃO EUROPEIA. Regulamento (UE) 2022/2065 do Parlamento Europeu e do Conselho, de 19 de outubro de 2022, relativo a um mercado único para os serviços digitais e que altera a Diretiva 2000/31/CE (Regulamento dos Serviços Digitais). *Jornal Oficial da União Europeia*, L 277, p. 1–102, 2022. Disponível em: https://eur-lex.europa.eu/eli/reg/2022/2065/oj. Acesso em: 31 ago. 2026.

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. Disponível em: https://arxiv.org/abs/2205.01833. Acesso em: 31 ago. 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. Disponível em: https://www.usenix.org/conference/usenixsecurity23/presentation/recabarren. Acesso em: 31 ago. 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. *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*. 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. (org.). *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.

DECLARAÇÃO DE DISPONIBILIDADE DE DADOS DA PESQUISA: O corpus bibliográfico consolidado, o mapa entre registros e estudos, o fluxo de seleção, os registros descartados com seus motivos, o protocolo analítico, os scripts e os materiais de auditoria estão depositados no Zenodo, atualmente em acesso restrito mediante autorização dos autores (Ferreira, Trevisan e Shirakura, 2026): https://doi.org/10.5281/zenodo.22237155. A restrição incide sobre os materiais depositados durante esta etapa de circulação e avaliação do trabalho e não impede sua preservação nem sua rastreabilidade. Os textos integrais dos estudos analisados, por serem materiais de terceiros, não integram o depósito; para fins de auditoria, foram preservados seus identificadores, sua situação de recuperação e seus *hashes*, permitindo verificar a correspondência entre os arquivos utilizados no processamento e os registros documentados no pipeline.

FINANCIAMENTO: Allan Herison Ferreira é bolseiro de doutoramento da Fundação para a Ciência e a Tecnologia (FCT), Portugal, referência: https://doi.org/10.54499/2022.14807.BD. Ana Carolina Trevisan é bolsista de doutoramento da Fundação para a Ciência e a Tecnologia (FCT), Portugal, referência: https://doi.org/10.54499/UI/BD/153755/2022. Este trabalho é financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia, I.P., no âmbito do Projecto referência UID/05021/2023.

CONTRIBUIÇÃO DOS AUTORES: 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.

DECLARAÇÃO DE CONFLITO DE INTERESSE: Os autores declaram que não há conflito de interesses a mencionar.

DECLARAÇÃO DE USO DE INTELIGÊNCIA ARTIFICIAL: Recursos de inteligência artificial foram utilizados de modo assistivo em quatro etapas, sob supervisão e responsabilidade dos autores: construção e execução das consultas às bases bibliográficas abertas; localização, por regras léxicas auditáveis, de trechos candidatos no texto integral dos estudos recuperados; organização e normalização da estrutura editorial; e revisão linguística. O conteúdo científico é de inteira responsabilidade dos autores.

MINIBIOGRAFIAS DOS AUTORES

Allan Herison Ferreira é doutorando em Ciências da Comunicação na NOVA de Lisboa, com bolsa da FCT no ICNOVA. Mestre em Sociologia pela USP, pesquisa identidades sociais, trabalho em tecnologia da informação, inteligência artificial e mediação algorítmica.

Ana Carolina Trevisan é doutoranda em Ciências da Comunicação na NOVA de Lisboa, com bolsa da FCT pelo IFILNOVA. Mestre em Sociologia pela USP, pesquisa estratégias argumentativas em redes sociais, sociologia política e análise socioargumentativa.

Camila Sayuri Shirakura é mestranda no Programa de Pós-Graduação em Ciências Sociais da Universidade Estadual de Maringá (UEM). Investiga a relação entre captura tecnológica, poder do Estado e direitos dos cidadãos.

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença Creative Commons CC-BY.
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.