



Universidade Federal da Paraíba
Centro de Ciências Humanas, Letras e Artes
Departamento de Filosofia
Curso de Pós-Graduação Lato Sensu em
Ética em Inteligência Artificial, EaD

Lyvia Rocha de Jesus Araujo

A operacionalização da injustiça epistêmica na inteligência artificial: uma revisão de escopo

João Pessoa
2026

Lyvia Rocha de Jesus Araujo

A operacionalização da injustiça epistêmica na inteligência artificial: uma revisão de escopo

Trabalho de conclusão de curso apresentado ao Curso de Pós-Graduação Lato Sensu em Ética em Inteligência Artificial, EaD, do Departamento de Filosofia, do Centro de Ciências Humanas, Letras e Artes, da Universidade Federal da Paraíba, como requisito para obtenção do título de Especialista em Ética em Inteligência Artificial.

Orientador: Dr. Danilo Fraga Dantas

João Pessoa
2026

Catálogo na publicação
Seção de Catalogação e Classificação

A658o Araujo, Lyvia Rocha de Jesus.

A operacionalização da injustiça epistêmica na inteligência artificial: uma revisão de escopo / Lyvia Rocha de Jesus Araujo. - João Pessoa, 2026.

45 f. : il.

Orientador : Danilo Fraga Dantas.

TCC (Especialização) - Universidade Federal da Paraíba/Centro de Ciências Humanas, Letras e Artes, 2026.

1. Inteligência artificial. 2. Modelos de linguagem de grande escala. 3. Injustiça epistêmica. 4. Metodologias. I. Dantas, Danilo Fraga. II. Título.

UFPB/CCHLA

CDU 004.8

Lyvia Rocha de Jesus Araujo

A operacionalização da injustiça epistêmica na inteligência artificial: uma revisão de escopo

Trabalho de conclusão de curso apresentado ao Curso de Pós-Graduação *Lato Sensu* em Ética em Inteligência Artificial, EaD, do Departamento de Filosofia, do Centro de Ciências Humanas, Letras e Artes, da Universidade Federal da Paraíba, como requisito para obtenção do título de Especialista em Ética em Inteligência Artificial.

Aprovado em: ____/____/____.

Banca Examinadora

Prof. Dr. Danilo Fraga Dantas
Orientador
Universidade Federal da Paraíba - Departamento de Filosofia

Prof. Dr. Bruno Ribeiro Nascimento
Membro Interno
Universidade Federal da Paraíba - Departamento de Comunicação

Profª. Dra. Dayanne da Silva Prudencio
Membro Externo
Universidade Federal do Estado do Rio de Janeiro - Departamento de Biblioteconomia

Ao meu companheiro, que aguardou com paciência a conclusão desta pesquisa e ofereceu com carinho seus ouvidos às minhas incessantes deliberações sobre as implicações da IA como vetor de amplificação de desigualdades sociais.

AGRADECIMENTOS

Ao corpo docente, que nos guiou e compartilhou seu conhecimento durante toda a especialização. À Universidade Federal da Paraíba (UFPB), por oferecer a estrutura que tornou nossa formação possível. E aos meus colegas de turma, pelas conversas, parcerias e debates que tornaram essa jornada enriquecedora.

“[...] A humanidade está diante do espelho
mais potente que já criou, fascinada com o
reflexo, e a maioria das pessoas nem sequer
está se perguntando quem segura o espelho”
(*Gemini 3.5 Flash*, 2026)

ARAUJO, Lyvia Rocha de Jesus. **A operacionalização da injustiça epistêmica na inteligência artificial**: uma revisão de escopo. 2026. 45 f. Trabalho de Conclusão de Curso (Especialização em Ética em Inteligência Artificial) – Departamento de Filosofia, Centro de Ciências Humanas, Letras e Artes, Universidade Federal da Paraíba, João Pessoa, 2026.

RESUMO

Esse artigo é uma revisão de escopo conduzida de acordo com as diretrizes do *Joanna Briggs Institute* e do *Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews*. Tem como objetivo mapear e sistematizar as estratégias analíticas e metodológicas utilizadas na literatura científica para identificar e avaliar a ocorrência de injustiça epistêmica em sistemas baseados em Inteligência Artificial. A coleta de dados foi realizada em sete fontes de informação multidisciplinares. As referências recuperadas foram organizadas e deduplicadas no sistema *online* Rayyan em preparação para as etapas de seleção dos estudos e extração de dados. Foram incluídos 11 estudos de abordagens qualitativas, quantitativas e mistas, a partir dos quais foram elaboradas duas tabelas para representação das características demográficas de cada estudo e apresentação dos modelos metodológicos adotados e suas respectivas limitações. Metodologias mistas mostram-se ferramentas eficazes para identificar e mensurar injustiças epistêmicas em sistemas de IA.

Palavras-chave: inteligência artificial; modelos de linguagem de grande escala; injustiça epistêmica; metodologias.

ARAÚJO, Lyvia Rocha de Jesus. **A operacionalização da injustiça epistêmica na inteligência artificial**: uma revisão de escopo. 2026. 45 f. Trabalho de Conclusão de Curso (Especialização em Ética em Inteligência Artificial) – Departamento de Filosofia, Centro de Ciências Humanas, Letras e Artes, Universidade Federal da Paraíba, João Pessoa, 2026.

ABSTRACT

This article is a scoping review conducted in accordance with the guidelines of the Joanna Briggs Institute and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews. It aims to map and systematize the analytical and methodological strategies used in the scientific literature to identify and evaluate the occurrence of epistemic injustice in Artificial Intelligence-based systems. Data collection was performed across seven multidisciplinary information sources. The retrieved references were organized and deduplicated in the Rayyan online system in preparation for the study selection and data extraction stages. Eleven studies with qualitative, quantitative, and mixed-methods approaches were included, from which two tables were developed to represent the demographic characteristics of each study and to present the adopted methodological models and their respective limitations. Mixed methodologies prove to be effective tools for identifying and measuring epistemic injustices in AI systems.

Keywords: artificial intelligence; large-language models; epistemic injustice; methods.

LISTA DE FIGURAS E QUADROS

Figura 1 - Fluxograma de seleção dos estudos de acordo com as recomendações do PRISMA-ScR	22
--	----

LISTA DE TABELAS

Tabela 1 - Estratégias de busca detalhadas por base de dados	20
Tabela 2 - Perfil e caracterização demográfica dos estudos incluídos na revisão	23
Tabela 3 - Síntese descritiva dos estudos incluídos na revisão	24

LISTA DE ABREVIATURAS E SIGLAS

ACM	<i>Association for Computing Machinery</i>
BIBED	<i>Bengali Identity Bias Evaluation Dataset</i>
BSA	<i>Bangla Sentiment Analysis</i>
CAPES	<i>Coordenação de Aperfeiçoamento de Pessoal de Nível Superior</i>
ChatGPT	<i>Generative pre-trained transformer Chat</i>
CoLAKE	<i>Contextualized Language and Knowledge Embedding</i>
CrowS-Pairs	<i>Crowdsourced Stereotype Pairs</i>
DP	<i>Demographic Parity</i>
EBSCO	<i>Elton B. Stephens Company</i>
EGIDS	<i>Expanded Graded Intergenerational Disruption Scale</i>
HCI	<i>Human-Computer Interaction</i>
IA	<i>Inteligência Artificial</i>
IAG	<i>Inteligência Artificial Generativa</i>
ICAT	<i>Idealized Context Association Test</i>
IMRD	<i>Introdução, Metodologia, Resultados e Discussão</i>
JB	<i>Joanna Briggs Institute</i>
KG	<i>Knowledge Graph</i>
Kepler	<i>Knowledge Embedding and Pre-trained Language Representation</i>
LAMA	<i>Language Model Analysis</i>
LLM	<i>Modelos de Linguagem de Grande Escala</i>
MECE	<i>Mutually Exclusive and Collectively Exhaustive</i>
Mean P@1	<i>Mean Precision at 1</i>
OLAC	<i>Open Language Archives Community</i>
PCA	<i>Principal Component Analysis</i>
PCC	<i>População, Conceito e Contexto</i>
PCM	<i>Pairwise Comparison Metric</i>
PLN	<i>Processamento de Linguagem Natural</i>
PRISMA-ScR	<i>Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Review</i>
QA	<i>Questions and Answers</i>
Redalyc	<i>Red de Revistas Científicas de América Latina y el Caribe, España y Portugal</i>
RoBERTa	<i>Robustly Optimized BERT Pretraining Approach</i>
SciELO	<i>Scientific Electronic Library Online</i>
SEAT	<i>Sentence Embedding Association Test</i>
T-Rex	<i>Template-based Relation Extraction</i>
Triplas	<i>Sujeito, predicado, objeto.</i>
UFPB	<i>Universidade Federal da Paraíba</i>
UFRJ	<i>Universidade Federal do Rio de Janeiro</i>
WoS	<i>Web of Science</i>

SUMÁRIO

1 INTRODUÇÃO	13
2 FUNDAMENTAÇÃO TEÓRICA.....	14
3 METODOLOGIA.....	18
3.1 DEFINIÇÃO DOS CRITÉRIOS DE ELEGIBILIDADE.....	18
3.2 ESCOLHA DAS FONTES DE INFORMAÇÃO	19
3.3 BUSCA NAS BASES DE DADOS	19
3.4 SELEÇÃO DOS ESTUDOS	20
3.5 EXTRAÇÃO DOS DADOS.....	21
3.6 AVALIAÇÃO CRÍTICA E ABORDAGEM REFLEXIVA COMPLEMENTAR	21
3.7 SÍNTESE DOS RESULTADOS.....	21
4 RESULTADOS	22
5 DISCUSSÃO	32
5.1 ANÁLISES ALGORÍTMICAS.....	33
5.2 MAPEAMENTO DE ASSIMETRIAS ESTRUTURAIS	36
5.3 METODOLOGIAS PARTICIPATIVAS	38
6 CONCLUSÃO.....	41
7 LIMITAÇÕES	42
REFERÊNCIAS	43

1 INTRODUÇÃO

Vivemos, atualmente, o que muitos veículos de informação, teóricos e especialistas modernos nomearam de “A era da IA”. O rápido avanço e propagação de inteligências artificiais (IAs), principalmente dos modelos de linguagem, têm modificado a maneira com a qual interagimos com a tecnologia e a aplicamos no nosso dia a dia. Não é de se espantar que frases como “joga no ChatGPT” ou “pesquisa na IA” tenham sido repetidas com tanta frequência. Nós utilizamos as IAs para esclarecer dúvidas, planejar eventos ou verificar opções, coisas que antes resolvíamos com uma pesquisa no *Google* ou com a ajuda de um bom grupo de amigos.

Apesar de parecer inofensiva e até mesmo engraçada, essa mudança sutil nos nossos hábitos evidencia o quão facilmente ignoramos a natureza dos sistemas de informação digitais. Ao delegarmos escolhas e interpretações à IA com cada vez mais frequência, podemos passar a acreditar que essas ferramentas são capazes de cognição, tratando-as, cada vez mais, como seres “quase humanos”. Nesse sentido, é importante darmos um passo atrás e situarmos as novas tecnologias de acordo com o que elas são e não com o que gostaríamos que elas fossem ou esperamos que elas sejam.

Essa necessidade de encarar as ferramentas de forma realista ganha urgência quando se percebe que, por trás da aparente imparcialidade técnica, os sistemas automatizados herdam e multiplicam preconceitos históricos, agindo como engrenagens de silenciamento e exclusão (Eubanks, 2018). Investigar como a ciência consegue identificar e medir esses danos nas IAs ainda é um desafio recente. Por isso, esta pesquisa se construiu a partir de uma revisão de escopo, a fim de traçar um panorama geral das evidências existentes, organizando o conhecimento fragmentado entre as ciências computacionais, as ciências sociais e a filosofia para compreender como a literatura atual investiga e denuncia a violência algorítmica.

Diante disso, o objetivo deste artigo é mapear e sistematizar as estratégias analíticas e metodológicas utilizadas na literatura científica para identificar e avaliar a ocorrência de injustiça epistêmica em sistemas baseados em Inteligência Artificial. Alinhada a esse objetivo, foi elaborada a seguinte questão norteadora: Quais abordagens metodológicas e/ou estratégias analíticas têm sido utilizadas na literatura para identificar a injustiça epistêmica em sistemas de Inteligência Artificial no que tange aos impactos sobre grupos marginalizados?

Para responder a esta pergunta, este artigo está estruturado em cinco seções subsequentes. A fundamentação teórica apresenta os conceitos relacionados e fornece um panorama geral do objeto de estudo deste artigo. A seção de metodologia detalha o protocolo

da revisão de escopo, explicitando as estratégias de busca e os critérios de elegibilidade adotados. A seção de resultados apresenta a síntese descritiva da amostra final de artigos selecionados. A seção de discussão aprofunda a análise crítica dos dados extraídos, agrupando as abordagens metodológicas encontradas em categorias temáticas e confrontando-as com o referencial teórico. Por fim, a conclusão sintetiza os principais achados do estudo, aponta as lacunas remanescentes na literatura e sugere caminhos para futuras investigações.

2 FUNDAMENTAÇÃO TEÓRICA

Segundo Floridi (2024, p. 52), os primeiros passos na conceituação da Inteligência Artificial giravam em torno de uma premissa curiosa: fazer com que uma máquina se comportasse de tal forma que, se um ser humano agisse igual, este seria considerado "inteligente". Essa premissa está fundamentada no histórico Teste de Turing, o Jogo da Imitação, proposto por Alan Turing (1950). O seu objetivo inicial não era criar uma mente que compreendesse o mundo, mas sim um espelho técnico capaz de reproduzir respostas, reações e comportamentos humanos diante de qualquer indagação ou solicitação de informação.

A partir disso, um frenesi de possibilidades, receios e perspectivas infladas passou a orbitar as reais capacidades tecnológicas da IA. Gerou-se o mito urbano e mercadológico de que esses sistemas poderiam, facilmente, superar os seres humanos em grande parte de suas faculdades intelectuais (Floridi, 2024, p. 68). Contudo, como destaca o autor, é mais adequado entender a IA como uma ferramenta interativa e autônoma, capaz de aprender e assumir tarefas cada vez mais complexas que, de outra forma, exigiriam intervenção humana e muito tempo para serem realizadas.

Ao nos atermos a esta interpretação, conseguimos desconstruir o imaginário de que a IA, de alguma forma, vai nos salvar de nossas limitações humanas ou nos tornar cada vez mais dispensáveis.

Embora este seja um primeiro passo para reaprendermos como coexistir com essa tecnologia, ainda é preocupante o modo como estamos adaptando a nossa realidade à existência da IA. Floridi (2024, p. 70) define como “circunscrição” a criação ou remodelagem de estruturas que favorecem o desempenho de agentes inteligentes. O autor aponta como exemplos deste fenômeno o aumento de satélites em órbita, o aumento de serviços e dispositivos digitais disponíveis, entre outras coisas.

Assim, reaprendemos a atuar no mundo e nossas interações passam a favorecer esses agentes. É por meio do fluxo contínuo de interações na internet que as IAs coletam o *big data*

que precisam para seu treinamento, transformando as escolhas do nosso dia a dia no insumo essencial para sua própria evolução.

Contudo, a conversão das interações cotidianas em dados não resulta em um repositório neutro de dados, mas sim em uma infraestrutura corporativa que tende a replicar e automatizar assimetrias estruturais. Sob a ótica da Epistemologia Social, que investiga o conhecimento a partir de suas dimensões coletivas, políticas e institucionais (Steup; Neta, 2020), as dinâmicas de credibilidade, autoridade epistêmica e confiança sofrem um deslocamento profundo no ambiente digital. Ao delegarmos a curadoria da realidade a modelos preditivos baseados em lógicas opacas, a confiança que antes sustentava as relações entre sujeitos humanos é artificialmente transferida para a suposta neutralidade das máquinas.

Nesse cenário, cabe indagar: quem passa está sendo legitimado como sujeito cognoscente e quem tem o poder de produzir conhecimento válido? O controle assimétrico das *Big Techs* sobre as infraestruturas técnicas dita o que é reconhecido como saber, convertendo o privilégio de dados em poder informacional. Como consequência, as bases de dados e os padrões probabilísticos do Norte Global passam a ditar a verdade, enquanto os saberes, dialetos e vivências de grupos marginalizados são sistematicamente deslegitimados e classificados como "ruído" ou erro estatístico (Khanna; Li, 2025). Longe de ser neutra, essa estrutura tende a aprofundar desigualdades e infligir novos danos a grupos vulnerabilizados, sob uma falsa ideia de infalibilidade tecnológica (O'Neil, 2020).

Esses danos materializam-se em problemas como a estigmatização de povos tradicionais, a ampliação de disparidades de gênero e a simplificação distorcida de identidades e heranças culturais. Problemas que evidenciam que a IA, em especial os modelos de linguagem de larga escala (LLMs), transforma injustiças preexistentes em processos automatizados de exclusão, operando como uma engrenagem ativa de opressão estrutural (Noble, 2021). Nesse cenário de contradições, a literatura tem se debruçado sobre conceitos que buscam abarcar a gravidade dos danos gerados quando padrões de desigualdade social são reproduzidos pelas IAs.

Movidos pelos interesses comerciais das grandes empresas de tecnologia, buscadores e sistemas de recomendação que incorporam IA, assim como os LLMs, tendem a priorizar e validar as narrativas do Norte Global. Os buscadores funcionam como organizadores de conteúdo que, sob uma falsa aura de neutralidade pública, decidem quais páginas aparecem primeiro ou somem da nossa vista com base em critérios de indexação de mercado. Já as ferramentas de recomendação tentam adivinhar o nosso comportamento futuro através de

modelos matemáticos preditivos, traçando perfis de usuários para antecipar suas interações. Por fim, os LLMs cruzam volumes imensos de texto para criar respostas probabilísticas que imitam a escrita humana.

Apesar de funcionarem de formas diferentes, essas três ferramentas têm uma mesma raiz que justifica sua análise conjunta: todas são treinadas com dados passíveis de viés. Na prática, enquanto os sistemas de recomendação podem aprofundar desigualdades ao direcionar anúncios que fomentem consumo negativo, ofertas de emprego desvantajosas ou conteúdos preconceituosos para perfis vulneráveis, os buscadores podem esconder o conhecimento e as informações de minorias sociais para mostrar apenas visões dominantes no topo das pesquisas, e os LLMs repetem e naturalizam preconceitos de gênero, raça ou religião como se estivessem dando respostas técnicas, neutras e corretas.

Por trás disso, a IA acaba funcionando como uma engrenagem para o que Benjamin (2019) conceitua como “*New Jim Code*”¹, uma roupagem nova para velhas discriminações, agora codificadas em parâmetros invisíveis de engenharia de *software*. Como os sistemas são desenhados para parecerem imparciais e puramente objetivos, as distorções são naturalizadas. A tecnologia, então, passa a atuar como uma ferramenta de triagem social punitiva, decidindo de forma invisível quem tem direito aos recursos públicos, à credibilidade e à visibilidade, afetando sistematicamente os sujeitos mais desfavorecidos (Benjamin, 2019; Eubanks, 2018).

Ao conectarmos essa engrenagem à filosofia, nos deparamos com a ideia de que essas tecnologias operam como reproduzoras digitais de injustiças epistêmicas, que prejudicam especificamente a nossa condição como sujeitos de conhecimento. Para delimitar analiticamente esse fenômeno, faz-se necessário distinguir conceitos que, embora próximos na literatura, não são equivalentes. A injustiça epistêmica reside no prejuízo moral e ético causado a alguém em sua capacidade de conhecedor (Fricker, 2024). Já a violência epistêmica evoca um processo de apagamento e silenciamento ativo de saberes subalternizados por discursos hegemônicos coloniais (Judijanto, 2025) e a apropriação hermenêutica automatizada ocorre quando a IA estabelece significados e recursos interpretativos sem supervisão humana (Pozzi, 2023). A injustiça algorítmica ou de plataforma, por sua vez, reflete a materialização dessas opressões codificadas no *design* técnico dos sistemas (Hoffmann, 2021; Kay; Kasirzadeh; Mohamed, 2024), enquanto a exclusão epistêmica aponta para a barreira que impede que comunidades inteiras sequer acessem ou façam parte das bases de dados (Oronoz; Pérez; Peña-

¹ O termo faz referência a Jim Crow, personagem principal de um espetáculo de minstrel de 1832 que ridicularizava os negros como forma de entretenimento.

Martínez, 2025). Em um nível extremo, o epistemicídio denota a destruição sistemática e a morte cultural de modos inteiros de povos colonizados produzirem saberes. Embora guardem suas especificidades, esses conceitos convergem ao evidenciarem que as assimetrias no ecossistema de dados perpetuam dinâmicas estruturais de dominação.

Para fins desta revisão, adota-se o termo "Injustiça Epistêmica" como um guarda-chuva para mapear como as metodologias científicas capturam a violação da condição do sujeito como portador de conhecimentos válidos diante dos algoritmos. Classicamente, Fricker (2024) divide a injustiça epistêmica em duas instâncias. A injustiça testemunhal ocorre quando o testemunho de um sujeito epistêmico é injustamente descartado, rejeitado ou colocado em dúvida devido a preconceitos do interlocutor que minam sua credibilidade. O principal fator que compromete a credibilidade dos indivíduos dessa maneira é o preconceito de identidade, que se manifesta em diferentes dimensões - como gênero, raça, idade, deficiência, classe, hereditariedade e outros marcadores sociais. A injustiça hermenêutica, por sua vez, acontece quando preconceitos de identidade estruturais nos recursos epistêmicos compartilhados fazem com que um indivíduo ou grupo tenha parte significativa de suas experiências sociais obscurecida pela compreensão coletiva, pela ausência de mecanismos semânticos que permitam sua adequada interpretação (Fricker, 2024, p. 17, 205).

Embora a formulação original de Fricker (2024) pressuponha a agência e a responsabilidade moral humana como requisitos para a ocorrência da injustiça epistêmica, a aplicação desse conceito aos sistemas de IA justifica-se por meio de uma necessária transição da moralidade individual para a justiça sociotécnica. A inteligência artificial, nesse contexto, deve ser compreendida não como um agente consciente, mas como um vetor de ampliação da agência epistêmica humana. Ao codificarem preconceitos históricos em parâmetros matemáticos invisíveis, os algoritmos operam como engrenagens que materializam e estendem a autoridade das corporações de tecnologia sobre a curadoria da informação (Floridi, 2024). Ademais, a dimensão hermenêutica da injustiça epistêmica, conforme teorizada pela própria autora, prescinde de um perpetrador individual imediato, caracterizando-se como um dano estrutural. Ao estabelecerem lógicas de classificação que apagam saberes de grupos vulnerabilizados, as IAs aprofundam lacunas interpretativas coletivas de forma autônoma (Hopster, 2024).

Desse modo, ainda que destituídos de intencionalidade moral, os sistemas automatizados exercem uma agência funcional distribuída, simulando e automatizando déficits de credibilidade e silenciamentos semânticos em escala global. Essa reconfiguração conceitual

fornece o alicerce metodológico necessário para operacionalizar as categorias de análise adotadas na presente revisão, permitindo classificar e examinar como diferentes metodologias científicas isolam, medem e tratam os vieses de IA não apenas como falhas matemáticas de calibração, mas como injustiças epistêmicas sistêmicas.

3 METODOLOGIA

Esta pesquisa seguiu as diretrizes do *Joanna Briggs Institute* (JBI) e do PRISMA-ScR (*Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews*) para elaboração de revisões de escopo. Esse procedimento permite identificar os principais conceitos, as correntes teóricas e as lacunas de conhecimento existentes na literatura (Peters *et al.*, 2020; Tricco *et al.*, 2018), garantindo a transparência e a replicabilidade da pesquisa.

Seguindo o modelo de referência, a metodologia foi organizada em sete etapas interdependentes, a saber: 1) definição dos critérios de elegibilidade; 2) escolha das fontes de informação; 3) busca nas bases de dados; 4) seleção dos estudos; 5) extração de dados; 6) avaliação crítica a partir dos dados extraídos; 7) síntese dos resultados.

3.1 Definição dos critérios de elegibilidade

O delineamento dos critérios de elegibilidade desta revisão teve início com a criação da questão de pesquisa, desenvolvida a partir do acrônimo PCC (População, Conceito e Contexto) indicado pelo JBI (Peters *et al.*, 2024). Destrinchando os elementos da questão, compreende-se que a população a ser estudada seria constituída por grupos marginalizados (P); o conceito investigado seria a injustiça epistêmica (C); e o contexto colocado em perspectiva seria o universo das aplicações de IA (C).

A fim de facilitar o processo de seleção dos estudos a serem incluídos, os critérios de elegibilidade desta revisão foram aplicados através dos seguintes critérios de inclusão e exclusão:

1. Critério de Inclusão: artigos publicados em revistas científicas, ensaios teóricos, estudos de caso e trabalhos publicados em anais de conferências que respondessem diretamente à pergunta de pesquisa, independentemente do desenho metodológico (qualitativo, quantitativo ou misto).

2. Critérios de Exclusão: a) editoriais, cartas ao editor, resenhas de livros e resumos de eventos sem o trabalho completo publicado; b) estudos focados genericamente em vieses algorítmicos e discriminação técnica (justiça distributiva ou erros materiais), sem interface explícita com a dimensão epistêmica (silenciamento ou desvalorização do conhecimento dos grupos afetados); c) estudos sobre injustiça epistêmica aplicados a contextos tradicionais (saúde, direito, educação) que não envolvessem sistemas de IA; d) trabalhos sem substância analítica ou componente metodológico claro (revisões teóricas, postagens em blogs de tecnologia ou artigos de opinião curtos).

Com o objetivo de garantir a abrangência do mapeamento, não foram aplicados filtros de idioma ou restrições temporais.

3.2 Escolha das fontes de informação

A busca bibliográfica foi realizada de forma sistemática entre 21 de janeiro de 2026 e 25 de fevereiro do mesmo ano. Para garantir a amplitude e a diversidade da coleta, foram consultadas oito fontes de informação eletrônicas, englobando bases multidisciplinares, bases específicas da ciência da computação e o portal de periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Foram consultadas: *Scopus*, *Web of Science* (WoS), *Dimensions AI*, *ACM Digital Library* (Association for Computing Machinery Digital Library), *Redalyc* (*Red de Revistas Científicas de América Latina y el Caribe, España y Portugal*), as bases de dados disponíveis no portal da EBSCO (*Elton B. Stephens Company*), *SciELO* (*Scientific Electronic Library Online*) e o Portal de Periódicos CAPES.

Todas as bases de dados privadas citadas acima foram acessadas através do acesso CAFE (Comunidade Acadêmica Federada), com uso do login institucional da Universidade Federal da Paraíba (UFPB). As extrações das referências na *Scopus* foram realizadas com uso do login institucional da Universidade Federal do Rio de Janeiro (UFRJ).

3.3 Busca nas bases de dados

Uma vez definidas as fontes de informação a serem consultadas, foi elaborada uma estratégia de busca compreendendo os descritores relacionados aos elementos da questão de pesquisa (inteligência artificial, injustiça epistêmica e metodologias). Para tanto, foram

realizadas duas consultas distintas para identificar sinônimos e outros termos relacionados aos descritores principais desta pesquisa: primeiro, fez-se um levantamento no referencial teórico disponibilizado pelo curso de Especialização em Ética em Inteligência Artificial da UFPB; segundo, recorreu-se ao material utilizado na formulação da proposta de projeto científico submetida ao processo seletivo do curso. Complementarmente, utilizou-se o modelo de linguagem de grande escala *Gemini 1.5 Pro*, do *Google*, para revisar a cobertura dos termos encontrados e propor palavras-chave adicionais que não haviam sido mapeadas anteriormente.

Após o levantamento dos descritores, a estratégia de busca foi estruturada pelo uso dos operadores booleanos AND e OR, refinada e adaptada conforme as especificidades técnicas e sintaxes de cada fonte de informação, conforme detalhado na Tabela 1.

Tabela 1 - Estratégias de busca detalhadas por base de dados

Escopo de aplicação	Estratégia	Base de dados	Abrangência Temática
Título	("Inteligência Artificial" AND "Injustiça Epistêmica" OR "Violência Epistêmica")*	SciELO	Multidisciplinar
Título, Resumo e Descritores (Title, Abstract and Keywords)	("Artificial Intelligence" OR "AI" OR "Algorithm*" OR "Machine Learning" OR "Automated Decision-Making" OR "LLM") AND ("Epistemic Injustice" OR "Epistemic Violence" OR "Epistemic Oppression" OR "Epistemicide" OR "Hermeneutical Injustice" OR "Testimonial Injustice" OR "Epistemic Exclusion")	Web of Science	Multidisciplinar
		Dimensions	Multidisciplinar
		Redalyc	Multidisciplinar
		Portal de Periódicos CAPES	Multidisciplinar
Acréscimo do terceiro descritor (Metodologias)	("Artificial Intelligence" OR "AI" OR "Algorithm*" OR "Machine Learning" OR "Automated Decision-Making" OR "LLM") AND ("Epistemic Injustice" OR "Epistemic Violence" OR "Epistemic Oppression" OR "Epistemicide" OR "Hermeneutical Injustice" OR "Testimonial Injustice" OR "Epistemic Exclusion") AND "Methodology" OR "Analytical Framework" OR "Operationalization" OR "Case Study" OR "Auditing" OR "Measurement")	EBSCO	Multidisciplinar
		Scopus	Multidisciplinar
		ACM Digital Library	Ciências da Computação

* Estratégias de busca com mais termos não retornaram nenhum resultado.
Fonte: Dados da pesquisa (2026).

3.4 Seleção dos estudos

Os resultados das buscas realizadas foram exportados no formato RIS, XLS ou CSV, a depender da disponibilidade das fontes. Posteriormente, todas as extrações foram carregadas na plataforma Rayyan², ferramenta *online* que auxilia pesquisadores na triagem, organização e análise de artigos científicos em revisões sistemáticas de literatura e meta-análises. Todo o processo de remoção de registros duplicados e seleção de estudos potenciais foi conduzido na plataforma.

² Disponível em: <https://www.rayyan.ai/>. Acesso em: 16 jun. 2026.

A triagem inicial foi feita por meio da leitura dos títulos e resumos dos estudos encontrados e da aplicação dos critérios de elegibilidade. Em seguida, os textos integrais dos artigos pré-selecionados foram recuperados e submetidos à leitura do texto completo para confirmação final de sua inclusão. O fluxo completo de busca, triagem e seleção será apresentado visualmente no fluxograma recomendado pelo PRISMA-ScR na seção de resultados.

3.5 Extração dos dados

A extração dos dados da amostra final foi executada por meio de uma matriz de síntese estruturada no *Excel*, contemplando as seguintes categorias de análise:

- Identificação do estudo (Autores, Gênero dos autores, ano de publicação, país de afiliação, área de conhecimento e idioma do documento);
- IA analisada;
- Injustiça Epistêmica infligida;
- Metodologia;
- Lacunas/Limitações.

3.6 Avaliação crítica e abordagem reflexiva complementar

Os dados tabulados a partir da amostra final foram submetidos a um processo de avaliação crítica individualizada, com o objetivo de analisar o caminho metodológico e a consistência teórica dos estudos selecionados. Essa etapa consistiu no exame minucioso de como cada artigo operacionalizou ou identificou o conceito de injustiça epistêmica no desenvolvimento, na auditoria ou na aplicação de sistemas de Inteligência Artificial. A análise buscou mapear as forças conceituais, as limitações metodológicas apontadas pelos autores e a solidez das evidências apresentadas pela literatura para responder à questão norteadora da pesquisa.

3.7 Síntese dos resultados

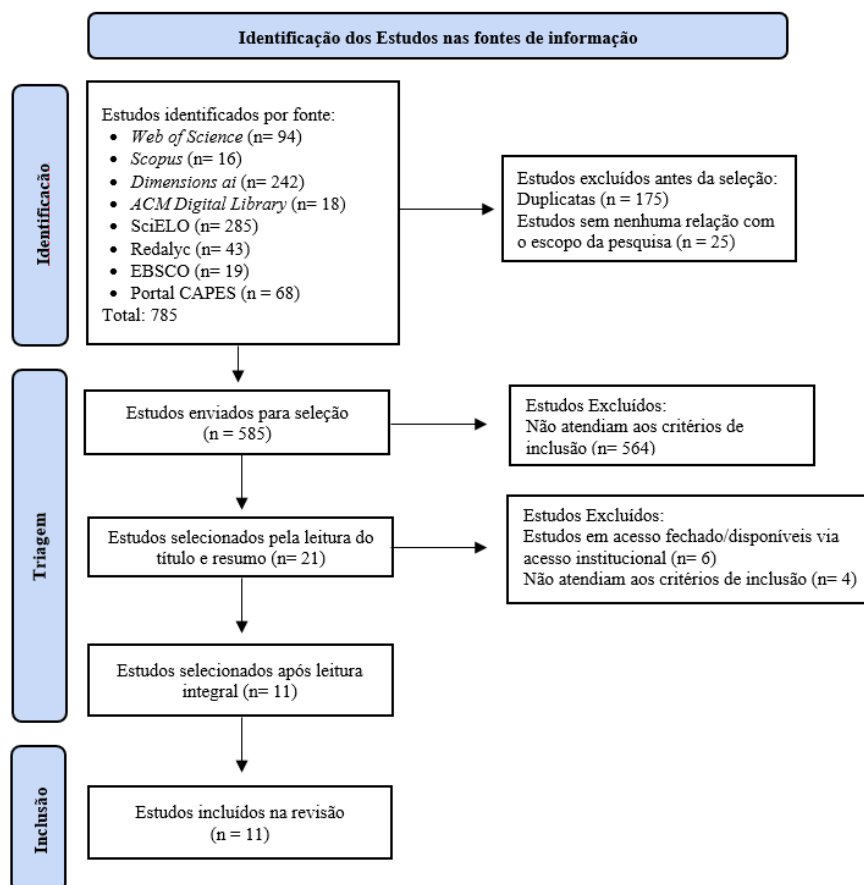
O conjunto de dados extraídos foi submetido a uma análise qualitativa-descritiva, estruturada a partir do agrupamento dos estudos por afinidade metodológica e conceitual. A

síntese narrativa foi organizada de modo a evidenciar os padrões recorrentes, as convergências teóricas e as principais lacunas identificadas na literatura sobre os impactos da IA na exclusão ou no silenciamento de grupos marginalizados. Por fim, os resultados dessa sistematização foram integrados e apresentados para fundamentar a discussão e conclusão desta pesquisa.

4 RESULTADOS

A seleção de estudos estruturada nesta revisão resultou em uma amostra final composta por 11 estudos/artigos que atendem estritamente aos critérios de inclusão delimitados. De maneira geral, a literatura científica mapeada reflete um debate predominantemente internacional e recente, concentrado de forma expressiva nos últimos quatro anos e pulverizado entre periódicos e conferências das ciências computacionais, ciências sociais aplicadas e filosofia. O fluxograma PRISMA-ScR, apresentado na Figura 1, ilustra os números obtidos nas etapas de identificação, triagem e inclusão dos estudos, demonstrando as exclusões feitas no decorrer do percurso e o quantitativo final de artigos considerados nesta pesquisa.

Figura 1 - Fluxograma de seleção dos estudos de acordo com as recomendações do PRISMA-ScR



Fonte: Dados da Pesquisa (2026).

De modo geral, a literatura recuperada demonstra que a preocupação em se investigar sistematicamente ocorrências de injustiça epistêmica em modelos de IA é um fenômeno recente e de natureza interdisciplinar. Para facilitar uma visão panorâmica dos artigos selecionados antes do detalhamento de suas variáveis metodológicas, a Tabela 2 apresenta o perfil demográfico da amostra, incluindo o identificador atribuído de cada estudo, sua autoria, ano de publicação, país de afiliação dos autores e a área científica de inserção de cada documento.

Tabela 2 - Perfil e caracterização demográfica dos estudos incluídos na revisão

ID	Sobrenome dos Autores	Ano de Publicação	Idioma	País de Afiliação dos Autores	Áreas do Conhecimento Identificadas
E01	Barry; Stephenson	2025	Inglês	Austrália	Política, Filosofia e Economia
E02	Kraft; Soulier	2024	Inglês	Alemanha	Ciências da Computação, Psicologia e Filosofia
E03	Walle	2025	Inglês	Etiópia	Gestão e Governança da Tecnologia e Sistemas de Informação
E04	Khanna; Li	2025	Inglês	Países Baixos; Reino Unido	Ciências da Comunicação e Ciência da Informação
E05	Zermeño-Flores; Tufte; Navarrete-Veja	2024	Inglês	México; África do Sul	Ciências Sociais e Ciências da Comunicação
E06	Mbalaka	2023	Inglês	África do Sul	Ciências Sociais
E07	Helm; Bella; Gertraud; Giunchiglia	2024	Inglês	Países Baixos; França; Alemanha; Itália	Ciências da Computação
E08	Gao; Yousuf; Thebault-Spieker	2025	Inglês	Estados Unidos	Ciência da Informação e Mídias Digitais
E09	Baeyaert	2025	Inglês	Portugal	[Ciências da Computação]*
E10	Oronoz; Miró Pérez; Peña-Martínez	2025	Inglês	Estados Unidos	Ciências da Educação e Economia
E11	Das; Guha; Semaan	2025	Inglês	Canadá; Estados Unidos	Ciências da Comunicação e Ciência da Informação

* Pesquisador independente, sem afiliação direta a nenhum departamento científico específico.

Fonte: Dados da Pesquisa (2026).

Após definido o perfil da amostra, os dados foram organizados em uma síntese descritiva para responder à questão central da pesquisa. As categorias utilizadas foram: sistema/aplicação de IA investigada; injustiça epistêmica identificada; metodologia aplicada; e limitações observadas. Optou-se por destacar os procedimentos metodológicos e suas restrições, em vez dos resultados e conclusões de cada estudo, em consonância com o objetivo da investigação. A síntese construída se encontra na Tabela 3.

Tabela 3 - Síntese descritiva dos estudos incluídos na revisão

ID	Sistema/aplicação de IA investigada	Injustiça Epistêmica identificada	Metodologia aplicada	Limitações observadas
E01	<i>Wordplay.ai</i> (Wordplay LTD); GPT-2 e ChatGPT (OpenAI); <i>Llama 2</i> (Meta)	Injustiça Testemunhal e Injustiça Hermenêutica	Estudo de caso comparativo, fundamentado em teorias feministas e epistemológicas, para analisar se e como os vieses da IAG contra mulheres podem perpetuar injustiças epistêmicas. As autoras aplicaram a teoria de Fricker sobre injustiça epistêmica aos resultados textuais e visuais gerados por diferentes plataformas de IA generativa. Foram considerados três estudos acadêmicos revisados por pares que simularam <i>prompts</i> relacionados a mulheres ou a questões de gênero em sentido mais amplo.	Análise teórico-crítica realizada a partir de um escopo limitado de casos dentro de uma temática específica (questões de gênero-desigualdade).
E02	<i>Kepler</i> (THU-KEG), <i>CoLAKE</i> (Universidade de Fudan) e <i>RoBERTa</i> (Meta)	Injustiça Testemunhal que pode vir a causar Injustiça Hermenêutica	Avaliação dos modelos <i>KEPLER</i> e <i>CoLAKE</i> em comparação ao RoBERTa. O objetivo foi verificar o efeito do aprimoramento do conhecimento por meio do <i>LAMA Probe</i> , um conjunto de tarefas intensivas em conhecimento. A análise se concentrou em dois tipos de viés: estereotípico, medido com três métricas: SEAT, que avalia associações entre grupos e atributos; CrowS-Pairs, que calcula a preferência por descrições estereotípicas de grupos historicamente desfavorecidos; e StereoSet, que compara respostas estereotípicas, antiestereotípicas e irrelevantes usando o índice ICAT de desempenho, avaliado na sub tarefa <i>T-REx</i> do <i>LAMA</i> , composta por 600 relações e 11 milhões de triplas do <i>Wikidata</i> . Foram selecionadas apenas triplas ligadas a entidades humanas, divididas em subconjuntos masculino e feminino, permitindo apenas comparação binária. A avaliação utilizou duas métricas: DP, razão entre acertos em exemplos femininos e masculinos, e <i>Mean P@1</i> , proporção de respostas corretas mais prováveis em cada grupo. Assim, a metodologia combinou análise de vieses estereotípicos e de desempenho em tarefas de conhecimento, oferecendo uma visão abrangente sobre desigualdades de representação nos modelos.	O estudo apresenta restrições no que tange à atualidade e à quantidade de modelos de Inteligência Artificial avaliados em seus testes. Adicionalmente, os autores apontam como lacuna a escassez de investigações empíricas recentes sobre como os diferentes desenvolvedores e atores de IA concebem o próprio conceito de "conhecimento".
E03	Plataformas de denúncias corporativas (<i>Whistleblowing Communities</i>)	Injustiça Testemunhal	Estudo conceitual e qualitativo que articula a teorização interpretativa à revisão problematizadora. O autor questiona o pressuposto de neutralidade dos canais digitais de denúncia ao transpor as teorias de injustiça epistêmica e de injustiça estrutural para o contexto dos sistemas de Inteligência Artificial.	Como artigo conceitual, este estudo é limitado pela própria metodologia. Seu objetivo é propor um novo arcabouço teórico e vocabulário, sem oferecer evidências quantitativas sobre a prevalência

			Para sustentar essa síntese, foram coletados e examinados de forma detalhada entre 40 e 50 casos documentados de denúncias mediadas por tecnologia, selecionando-se uma amostra focalizada - como os casos de moderadores de conteúdo no Quênia e trabalhadores de plataformas - com base em critérios de elegibilidade, como mediação digital, identidade social dos denunciante e a evidência de barreiras algorítmicas estruturais.	do viés algorítmico de credibilidade entre plataformas.
E04	<i>Hugging Face (Hugging Face, Inc.), GPT-4 (OpenAI), Claude (Anthropic) e Gemini (Google)</i>	Injustiça Testemunhal e Injustiça Hermenêutica (Injustiça Epistêmica Digital)	Estudo de abordagem mista (quanti-qualitativa) baseada no cruzamento de dados demográficos reais com a presença digital de todas as 7.613 línguas documentadas no <i>Ethnologue</i> . A vitalidade real das línguas foi calculada pelo número de falantes e pela escala EGIDS; a digitalidade foi quantificada pelo volume de dados de cada língua presentes no <i>Common Crawl</i> , <i>Wikipedia</i> , <i>Hugging Face</i> e <i>OLAC</i> . Após processamento estatístico para redução de dimensionalidade utilizando PCA, os autores criaram uma matriz que divide as línguas em quatro classes. Esse mapeamento quantitativo foi, então, interpretado qualitativamente à luz da teoria pós-colonial e dos conceitos de injustiça epistêmica.	A medição da presença digital das línguas considerou apenas dados públicos da internet, o que pode ter subestimado seu uso real em plataformas de mensagens privadas. O recurso a divisões baseadas na mediana para classificar as quatro subdivisões propostas estabeleceu fronteiras rígidas que podem mascarar variações internas. Por se tratar de uma análise transversal, o estudo não avalia a evolução histórica nem a causalidade dos déficits de representação ao longo do tempo. Por fim, a pesquisa foi conduzida por pessoas externas, sem o envolvimento direto das comunidades afetadas no desenho do estudo, carecendo também de análises cruzadas com outras formas de opressão, como gênero, idade e línguas de sinais.
E05	<i>SciSpace (Calaver Technologies Private Limited) e Elicit (Elicit Research, PBC)</i>	Injustiça Epistêmica Sistêmica	Estudo de abordagem qualitativa-interpretativa, adotando estratégia reflexiva de autoetnografia e observação participante fundamentada nas epistemologias feminista e decolonial. Como objeto de análise prático, a equipe de pesquisa realizou uma revisão de literatura sobre o tema do ativismo juvenil utilizando um ecossistema tecnológico misto composto por bases de dados hegemônicas (<i>Web of Science</i>), alternativas <i>open-access</i> do Sul Global (<i>Redalyc</i>), ferramentas de IA especializadas (<i>SciSpace</i> e <i>Elicit</i>) e <i>softwares</i> de mapeamento bibliométrico (<i>VOSviewer</i>). Para avaliar esse processo, os pesquisadores aplicaram a técnica de posicionamento crítico da "Flor de Perguntas" (<i>Flower of Questions</i>), avaliando de forma auto-exploratória três categorias de compromissos do sujeito pesquisador: biográficos-afetivos,	O trabalho relata que as disparidades econômicas e de financiamento entre as universidades parceiras (uma no Reino Unido e outra no México) ditaram limitações severas de acesso, visto que a equipe do Sul Global não possuía acesso à <i>Web of Science</i> devido ao seu alto custo de assinatura, restringindo-se ao uso de tecnologias gratuitas. Além disso, as políticas de segurança digital institucionais impediram o compartilhamento de documentos de trabalho entre os servidores das universidades, forçando a equipe a

			ético-políticos e pragmáticos. A grande inovação metodológica do artigo consistiu na proposta e adição de um quarto componente à técnica, batizado de Compromisso Procedimental (<i>Procedural Commitment</i>), desenhado especificamente para registrar de forma sistemática em anotações, <i>logs</i> e reuniões gravadas quais foram os erros, as orientações e as falhas de triagem algorítmica ocorridas estritamente durante a interação prática com as ferramentas de IA e tecnologias digitais.	migrar dados para contas pessoais, ao mesmo tempo em que houve incompatibilidade técnica total entre os formatos: os metadados gerados pela plataforma Redalyc não puderam ser exportados ou lidos pelo <i>software</i> VOSviewer, gerando um apagamento técnico imediato daqueles artigos na plotagem final dos mapas conceituais. Por fim, a própria capacidade preditiva e analítica da IA mostrou-se incipiente, falhando em inferir e identificar quadros teóricos e metodológicos em textos que não seguiam rigidamente a nomenclatura e a divisão padrão linear do formato ocidental IMRD, o que atesta que os processos de inferência e disambiguação de ferramentas de mineração de texto e grandes quantidades de dados ainda se encontram em estágio inicial e demandam supervisão humana contínua.
E06	<i>DALLE-E 2</i> (OpenAI) e <i>Starry AI</i> (Starryai, Inc.)	Violência Epistêmica	Estudo de abordagem mista, que combinou uma “pesquisa de mesa” (<i>desktop study</i>) para fundamentação da literatura com a realização de experimentos textuais e visuais práticos em plataformas de IA. No vetor quantitativo, o processamento visual foi realizado em <i>Python</i> para medir o ruído, a dispersão de pixels e a qualidade das imagens geradas a partir do cálculo do desvio padrão e da média dos elementos de cor RGB (<i>Red, Green, Blue</i>) dentro de uma região de interesse (ROI) predefinida. Essa análise numérica foi conjugada a um teste visual para verificar a precisão anatômica e a acurácia etnográfica da representação cultural dos sujeitos renderizados pelas IAs.	A principal limitação metodológica do foi o escopo escolhido para a abordagem experimental, o que reduz a capacidade de generalização dos resultados obtidos. Adicionalmente, o autor reconhece que as métricas quantitativas de imagem (desvio padrão e média) possuem limitações severas, pois não isolam perfeitamente variáveis como saturação, brilho e contraste, exigindo obrigatoriamente a validação visual de um especialista humano. Por fim, o estudo ressalta que a análise estritamente matemática da qualidade computacional não é um reflexo direto ou suficiente para mensurar a precisão e a violência epistêmica contida

				nos traços faciais e nos detalhes culturais gerados pelos algoritmos de IA.
E07	<i>DeepL</i> (DeepL SE), <i>ChatGPT</i> (OpenAI), <i>Google Translate</i> (Google), <i>English Princeton WordNet</i> (Universidade de Princeton)	Injustiça Hermenêutica	<i>Desktop study</i> , revisão de literatura e coleta de dados a partir dos geradores de texto-para-imagem <i>DALL-E 2</i> (OpenAI) e <i>Starry AI</i> . O modelo de análise se baseia na demonstração de diversos casos de lacunas lexicais - mapeamento de termos de parentesco e culinária em línguas como Maori, Kiswahili e Japonês, traduzidos para inglês - aliados a uma análise crítico-genealógica das assimetrias de poder na ciência da computação.	Foco estrito em dimensões linguísticas, não considerando os vieses presentes no <i>design</i> das plataformas exploradas e suas metodologias. Não aprofunda uma discussão mais explícita sobre as dimensões ético-políticas do problema.
E08	GPT-5 (OpenAI)	Injustiça Epistêmica	Estudo construído a partir de três oficinas de mapeamento participativo com 24 membros de uma comunidade. Foram desenvolvidos métodos de elicitación e um esquema capaz de preservar a riqueza narrativa, permitindo ao mesmo tempo a extração, validação e controle de procedência de entidades, tempo e lugares. Para delimitar o problema, foi realizada uma auditoria em um <i>benchmark</i> de nível municipal com 14.782 pares de perguntas e respostas locais, na qual lacunas factuais, incompreensões culturais, confusões geográficas e desalinhamentos temporais representaram 76,7% dos erros. Em um conjunto participativo de QA derivado das oficinas, o GPT-5 respondeu corretamente a menos de 21% das questões sem contexto adicional, evidenciando a necessidade de fundamentação local.	Devido à abordagem escolhida, os resultados podem ter baixa generalização, sendo recomendado pelos autores a validação adicional das proposições. A conversão de narrativas em representações computacionais pode gerar perda de significado e, sem garantias de origem e moderação, corre o risco de apropriação indevida. <i>Corpora</i> narrativos tendem a amplificar vozes com maior acesso ou respaldo institucional, enquanto perspectivas minoritárias permanecem pouco documentadas. As comunidades não são homogêneas, e sua legitimidade pode ser contestada por grupos com valores e prioridades distintos. Há uma tensão entre anonimização e utilidade: quanto mais específica a narrativa, mais útil e também mais identificável. O consentimento é contextual e temporal; o que parecia compartilhável em um <i>workshop</i> pode se tornar arriscado quando indexado e reutilizado por modelos, sobretudo para grupos sob vigilância ou estigma. Embora modelos de governança comunitária descentalizem o poder em

				relação a plataformas e corporações, eles ainda podem reproduzir assimetrias em escala local.
E09	21 estudos anteriores que usavam sistemas generativos e preditivos (<i>DALL-E 2</i> , triagens clínicas e sistemas de tutoria)	Injustiça Testemunhal, Injustiça Hermenêutica e Injustiça Distributiva	Estudo de abordagem mista utilizando: construção teórica conceitual; elaboração de casos ilustrativos e síntese de evidências empíricas. Foi conduzida uma revisão iterativa da literatura em epistemologia, ética da IA e HCI sociotécnica, mapeando três tipos de injustiça epistêmica sobre as quatro etapas do fluxo de aprendizado da IAG. Os resultados obtidos resultaram em uma taxonomia MECE e fundamentaram a formulação de quatro hipóteses teóricas. Para demonstrar concretamente como as injustiças identificadas se manifestam em sistemas de IA generativa, foram desenvolvidos dois cenários narrativos fictícios sustentados em evidências: um <i>chatbot</i> diagnóstico em saúde e um gerador de notícias multilíngue. Além disso, foi realizada uma revisão sistemática em quatro bases (<i>Scopus</i> , <i>Web of Science</i> , <i>arXiv</i> e <i>ACM Digital Library</i>), com data de corte em 31 de maio de 2025. Dos 42 registros inicialmente recuperados, 21 atenderam aos critérios de inclusão pré-registrados ($N \geq 40$, método claro, estatísticas de dispersão reportadas). Os estudos quantitativos foram convertidos em tamanhos de efeito de Cohen's, seguindo protocolo formal.	Falta de estudos fatoriais cruzados: Nenhuma das pesquisas empíricas analisadas testou simultaneamente a interação entre a opacidade do sistema e a calibração das explicações (um design fatorial 2×2 para as hipóteses H3 e H4), deixando essa co-dependência estatisticamente subespecificada. Desequilíbrio de evidências entre as hipóteses: Enquanto as hipóteses comportamentais (H3 e H4) foram validadas diretamente por tamanhos de efeito estatístico (Cohen's d), as hipóteses estruturais sobre silenciamento de inputs (H1) e deriva de conceitos (H2) sustentam-se apenas em mapeamentos conceituais, simulações e análises prévias, carecendo de estimativas meta-analíticas próprias dentro do estudo.
E10	<i>ChatGPT</i> (OpenAI)	Exclusão Epistêmica	Investigação teórica-conceitual sobre o histórico, práticas culturais, características genéticas e organização social do grupo étnico Merchera e avaliação do desempenho do <i>ChatGPT</i> em face de indagações similares. Foi realizada uma revisão exploratória da literatura em diversas bases acadêmicas, utilizando uma estratégia de busca inclusiva que considerou todos os documentos contendo as raízes “merchera” ou “merchero”, em qualquer idioma, sem restrições temporais ou linguísticas. Para ampliar a cobertura, os termos foram traduzidos e expandidos, como “ <i>merchera ethnic group</i> ” em inglês e “ <i>groupe ethnique merchera</i> ” em francês. Três critérios de exclusão garantiram rigor acadêmico: ausência de qualidade científica, irrelevância para o foco da pesquisa e indisponibilidade do texto completo.	Este estudo apresenta limitações, sobretudo pela dificuldade de acesso a informações abrangentes e confiáveis sobre o grupo étnico Merchera. A exclusão histórica dessa comunidade do discurso acadêmico gerou lacunas significativas, agravadas por vieses e estereótipos presentes em muitas fontes, o que compromete uma compreensão precisa e contextualizada. O uso de conteúdo gerado por IA, embora inovador, traz restrições metodológicas. O modelo <i>ChatGPT</i> , treinado em grandes bases de dados, pode reproduzir preconceitos ou informações não

			<p>A coleta de dados consistiu em diálogos estruturados com o <i>ChatGPT</i>, guiados por <i>prompts</i> elaborados para obter informações abrangentes sobre o grupo, incluindo contexto histórico, práticas culturais, herança genética e organização social. As respostas geradas foram avaliadas sistematicamente por meio de comparação com informações verificadas em bases acadêmicas, fontes arquivísticas e literatura especializada. A análise seguiu o modelo qualitativo, confrontando os conteúdos produzidos pela IA com o conhecimento acadêmico estabelecido, com foco na precisão, profundidade e representação do grupo.</p>	<p>verificadas, reforçando injustiças epistêmicas. Isso evidencia a necessidade de aplicar uma leitura crítica aos resultados da IA e integrá-los a métodos acadêmicos convencionais.</p> <p>A revisão de literatura, ainda que extensa, não é exaustiva. Futuras pesquisas devem ampliar o escopo, incorporando perspectivas de áreas como estudos romani, justiça epistêmica, teoria pós-colonial e ética da IA. Um enquadramento multidisciplinar favoreceria uma compreensão mais completa da identidade Merchera e das dinâmicas sociopolíticas que moldam sua representação.</p> <p>O estudo também se beneficiaria de validação contextual e especializada. Metodologias qualitativas — como entrevistas com especialistas, trabalho etnográfico e interação direta com membros da comunidade — fortaleceriam a profundidade e credibilidade dos achados, garantindo a representação fiel das experiências vividas por grupos marginalizados.</p> <p>Em síntese, embora este trabalho ofereça contribuições iniciais sobre o potencial da IA generativa na pesquisa de comunidades marginalizadas, suas conclusões devem ser interpretadas à luz das limitações metodológicas, epistemológicas e éticas apontadas. Pesquisas futuras devem buscar superar essas lacunas por meio de abordagens interdisciplinares, participativas e criticamente reflexivas.</p>
--	--	--	---	---

E11	<i>mBERT</i> (Google) e <i>BanglaBERT</i> (Universidade de Engenharia e Tecnologia de Bangladesh)	Injustiça Testemunhal e Injustiça Hermenêutica	<p>Auditoria algorítmica conduzida com os modelos <i>mBERT</i> e <i>BanglaBERT</i>, ajustados com os conjuntos de dados BSA, resultando em 38 modelos refinados. Esses modelos foram testados quanto a vieses de gênero, religião e nacionalidade, utilizando pares de sentenças idênticas do BIBED, que representam categorias identitárias bengalis de forma explícita e implícita.</p> <p>Na comparação estatística entre grupos, aplicou-se o teste qui-quadrado para avaliar a relação entre categorias identitárias e classificação de sentimento. As médias dos escores de sentimento foram comparadas par a par entre categorias de gênero (feminino-masculino), religião (hindu-muçulmano) e nacionalidade (bangladense-indiano).</p> <p>Para responder à segunda questão de pesquisa, investigou-se se a direção dos vieses dos modelos estava relacionada às categorias identitárias dos desenvolvedores dos conjuntos BSA, aplicando novamente testes estatísticos para verificar a hipótese nula.</p> <p>Na quantificação dos vieses de grupo, buscou-se entender como diferentes combinações de modelos pré-treinados e conjuntos de dados influenciam os vieses nos modelos ajustados. Para comparar classificações nominais, seguiram-se diretrizes de paridade demográfica, que avaliam taxas iguais de classificação positiva entre grupos. Quando dois modelos apresentavam taxas semelhantes, utilizou-se uma métrica secundária de comparação, PCM.</p>	Devido ao uso do BIBED, o estudo desconsiderou identidades não binárias, minorias religiosas, nacionalidades da diáspora e normas linguísticas regionais. Este foi o único conjunto de dados bengali a identificar vieses, razão principal para a adoção da classificação binária. A prática comum de binarificação em conjuntos de PLN limita análises algorítmicas e revela fragilidades do campo. Apesar dos esforços dos autores, o estudo foi restringido pelas ontologias dos recursos disponíveis.
-----	---	--	--	---

Legenda: **BIBED** – Bengali Identity Bias Evaluation Dataset; **BSA** – Bangla Sentiment Analysis; **CoLAKE** – Contextualized Language and Knowledge Embedding; **CrowS-Pairs** – Crowdsourced Stereotype Pairs; **DP** – Demographic Parity; **EGIDS** – Expanded Graded Intergenerational Disruption Scale; **HCI** – Human-Computer Interaction; **IAG** – Inteligência Artificial Generativa; **ICAT** – Idealized Context Association Test; **IMRD** – Introdução, Metodologia, Resultados e Discussão; **KG** – Knowledge Graph; **Kepler** – Knowledge Embedding and Pre-trained Language Representation; **LAMA** – Language Model Analysis; **MECE** – Mutually Exclusive and Collectively Exhaustive; **Mean P@1** – Mean Precision at 1; **OLAC** – Open Language Archives Community; **PCA** – Principal Component Analysis; **PCM** – Pairwise Comparison Metric; **PLN** – Processamento de Linguagem Natural; **QA** – Questions and Answers; **RoBERTa** – Robustly Optimized BERT Pretraining Approach; **SEAT** – Sentence Embedding Association Test; **T-Rex** – Template-based Relation Extraction; **Triplas** – Sujeito, predicado, objeto .

Fonte: Dados da Pesquisa (2026).

A partir dos resultados expostos, destaca-se que todos os trabalhos incluídos na amostra foram publicados em inglês, e apenas três deles (27,3%) apresentaram afiliação a instituições do Sul Global ^(E03, E05, E06), o que sugere uma sub-representação de perspectivas locais e de países em desenvolvimento na discussão sobre injustiça epistêmica em IA. Nesse cenário, quatro estudos (36,4%) são multinacionais, com autores vinculados a instituições de diferentes países ^(E04, E05, E07, E11), indicando uma tendência relevante de colaboração internacional, embora a maior parte da amostra ainda esteja concentrada em redes acadêmicas do Norte Global. Observa-se também diversidade disciplinar nas áreas de conhecimento envolvidas - Ciências da Computação, Filosofia, Comunicação, Ciências Sociais, Educação, Economia, entre outras -, porém não foi identificada relação direta entre a escolha metodológica dos estudos e as áreas de atuação dos autores.

De acordo com os resultados apresentados na Tabela 3, é possível observar semelhanças entre os procedimentos metodológicos adotados para mensurar os impactos das ferramentas de IA sobre populações marginalizadas. A maioria dos estudos seguiu uma abordagem qualitativa (N = 6; 54,5%), estabelecendo comparações crítico-conceituais entre a teoria e os contextos analisados ^(E01, E03, E05, E07, E08, E10). Dentre os quais destacam-se os estudos E01 e E05 que evidenciam a predominância de referenciais feministas, decoloniais e pós-coloniais, indicando uma tendência à vinculação da IA a debates sociais e epistemológicos. O segundo grupo mais representativo na amostra corresponde aos estudos de caráter misto (N = 3; 27,3%), que combinam análises quantitativas e qualitativas com o objetivo de aprofundar conceitos e confirmar hipóteses de forma prática e objetiva ^(E04, E06, E09). Dois estudos (18,2%) recorreram exclusivamente à análise quantitativa ao aplicarem métricas estatísticas e testes comparativos para elucidar possíveis vieses algorítmicos existentes nas ferramentas investigadas ^(E02, E11).

No que tange às limitações, cinco aspectos recorrentes foram identificados nos estudos analisados. O primeiro refere-se ao escopo restrito e à falta de generalização, apontados em cinco trabalhos ^(E01, E03, E06, E08 e E10), que destacam como o recorte temático reduz a capacidade de ampliar conclusões. O segundo diz respeito à dependência de dados incompletos ou enviesados, observada em três estudos ^(E04, E05, E11), causada pela baixa qualidade dos dados disponíveis e por incompatibilidades técnicas entre plataformas. Em terceiro lugar, a ausência de participação direta das comunidades afetadas, evidenciada em E04 e E10, compromete a representatividade dos resultados. O quarto aspecto envolve limitações técnicas e metodológicas ^(E05, E06 e E09), incluindo falhas nas métricas de imagem, lacunas estatísticas e barreiras institucionais que dificultam coleta e análise. Por fim, os estudos quantitativos ^{(E02 e}

E11) apontaram atualidade e abrangência insuficientes, revelando restrições quanto ao número de modelos avaliados e à incapacidade de contemplar identidades não binárias ou minorias religiosas.

Um último detalhe a ser apontado nesta seção, embora não registrado nas Tabelas 2 e 3, é que seis dos 11 estudos incluídos nesta revisão têm mulheres na posição de primeira autoria (E01, E02, E05, E07, E08, E10). Esse dado revela uma presença expressiva de pesquisadoras liderando investigações sobre injustiça epistêmica em IA, reforçando a diversidade de perspectivas e indicando maior sensibilidade às questões de gênero e marginalização no campo.

5 DISCUSSÃO

A análise dos estudos que compõem a amostra final desta revisão revelou um panorama de efervescência metodológica e conceitual, marcado pelo esforço multidisciplinar em capturar, mensurar e evidenciar a injustiça epistêmica em modelos de IA. Para responder à questão norteadora desta pesquisa, os dados foram organizados em três eixos de reflexão que demonstram como a literatura recente vem investigando estas ferramentas. O primeiro eixo compreende os métodos de análises algorítmicas, através do isolamento quantitativo de variáveis e do teste dos modelos para identificar desvios de paridade estatística e confirmar a existência de vieses.

O segundo abarca o mapeamento de assimetrias estruturais, direcionado à identificação de lacunas léxicas, discrepâncias de representação e exclusão de dados oriundos de comunidades marginalizadas. No contexto desta investigação, as lacunas léxicas referem-se à ausência ou à tradução empobrecida de vocábulos e termos de identidade específicos de línguas minoritárias ou dialetos periféricos, gerando barreiras de comunicação e apagamento cultural. As discrepâncias de representação correspondem à distribuição estatística desigual e estereotipada de identidades de gênero, raça ou religião nos conjuntos de dados massivos, perpetuando visões depreciativas e reduzindo corpos e tradições a traços homogêneos. Por sua vez, a exclusão de dados consiste na completa omissão ou invisibilização de repertórios teóricos, históricos e linguísticos de grupos vulnerabilizados nos dados de treinamento, o que impossibilita que suas vozes e conhecimentos sejam sequer processados ou recuperados pelos sistemas inteligentes.

Por fim, o terceiro eixo apresenta metodologias participativas, que buscam transformar os membros de comunidades marginalizadas em atores ativos nos processos de validação e

regulação de sistemas automatizados, devolvendo-lhes a agência semântica e o controle sobre a produção de seus saberes.

5.1 Análises algorítmicas

A condução de análises algorítmicas com o intuito de medir estatisticamente as disparidades existentes em modelos de IA quanto à interpretação de dados de populações menos ou mais bem representadas, ocorre primariamente em estudos que empregam análises quantitativas em suas estratégias metodológicas. Nesse cenário, percebe-se o predomínio de abordagens que tentam interrogar a IA a partir de sua arquitetura e de como ela interage com os dados captados pelo sistema ou nele imputados.

Em respeito à identificação da injustiça epistêmica, verifica-se que o uso de métodos quantitativos ocorre pela operacionalização de estratégias de pareamento de grupos e atributos^{E02} - *probing semântico* - ou pela medição de pontuações de confiança em conteúdos factuais, neutros ou fictícios - análise de sentimento^{E11}. Estudos que utilizam qualquer uma destas estratégias isolam o comportamento do sistema ao injetar pares de frases com estruturas sintáticas e significados idênticos, modificando apenas termos específicos que funcionam como marcadores de identidade. O objetivo é extrair desvios de paridade estatística, evidenciando como a IA penaliza sistematicamente enunciados de grupos minorizados, atribuindo a eles pontuações de confiança inferiores ou categorias de sentimento negativas, mesmo quando o conteúdo alimentado é factual ou neutro. Tal ocorrência corrobora Benjamin (2019) ao declarar que, uma vez que os sistemas baseados em aprendizado de máquina dependem de grandes conjuntos de dados “naturalmente ocorrentes”, repletos de viéses raciais, econômicos e de gênero, os dados brutos que a IA utiliza para aprender e tomar decisões sobre o mundo refletem, direta ou indiretamente, hierarquias estruturais e preconceitos culturais profundamente enraizados na sociedade.

O estudo E02 confirma a colocação da autora ao demonstrar que, mesmo em bases de conhecimento estruturadas como Wikidata e Wikipedia, utilizadas para treinar modelos de linguagem natural de grande escala, há viés representativo: apenas cerca de 20% das entradas referem-se a mulheres e menos de 1% a outros gêneros (não masculinos/binários). Além disso, essas bases permitem diferenças narrativas impróprias que perpetuam estereótipos e violência. Com resultados similares, o estudo E11 mostrou que os modelos de análise de sentimentos em bengali apresentam viés significativo em relação a gênero, religião e nacionalidade. A maioria

dos conjuntos de dados testados pelos autores favorecia narrativas masculinas (61%), muçulmanas (61%) e bangladeshis (50%), enquanto apenas uma parcela menor favorecia mulheres, hindus ou indianos. Embora os conjuntos de dados tenham sido desenvolvidos principalmente por homens, muçulmanos e bangladeshis, não foi encontrada relação estatística direta entre o perfil dos criadores e os vieses dos modelos. Ao fim da análise ficou claro que nenhum dos conjuntos de dados estava livre de viés.

Além do uso majoritário de análises quantitativas, as análises algorítmicas em IA têm sido aplicadas especialmente em sistemas fechados, de empresas privadas, sendo por vezes chamadas de “auditorias de caixa-preta”. Isso porque a opacidade dos sistemas proprietários limita até onde é possível verificar e testar as engrenagens que atuam por trás do seu funcionamento. Um exemplo disso é apresentado no estudo E06, que investigou geradores de imagem como DALL-E 2 e *Starry AI* através dos níveis de ruído e dispersão de *pixels* em regiões faciais pré-definidas, calculando o desvio padrão e variações nos canais RGB.

O cruzamento desses indicadores numéricos com a inspeção visual demonstrou estatisticamente a degradação severa na qualidade gráfica e na fidelidade geométrica das imagens sempre que os algoritmos eram solicitados a renderizar rostos e corpos de famílias racializadas ou minorias culturais, em contraste com a nitidez conferida aos grupos dominantes (caucasianos)^{E06}. Benjamin (2019) explica este fenômeno ao demonstrar que os geradores digitais de imagem herdaram os antigos padrões analógicos de calibração visual, que historicamente fixaram a pele branca como o padrão normativo de nitidez e luz. A autora declara que a incapacidade técnica de processar com precisão os canais de cor e os *pixels* de corpos negros é, fundamentalmente, uma forma de discriminação engenheirada que reproduz a supremacia racial na própria estética da IA (Benjamin, 2019).

Esses achados empíricos dialogam diretamente com o conceito de opressão algorítmica de Noble (2018) ao reforçarem sua teoria de que os sistemas computacionais de representação visual não são neutros, mas operam como espelhos que refletem e amplificam as prioridades comerciais e os preconceitos raciais de seus criadores. Esse fenômeno resulta na desumanização e na degradação sistemática de identidades periféricas e corpos racializados na infraestrutura digital. Viéses como os apresentados nos estudos E02, E06 e E11 são mais difíceis de detectar, mas potencialmente mais prevalentes, reforçando a necessidade de análises que vão além da paridade estatística ou da análise de pixels em sistemas de IA.

Análises críticas sobre a opacidade dos sistemas proprietários, a falta de explicabilidade dos processos inferenciais, as fontes de dados utilizadas em seu treinamento e os estilos de

representação adotados são necessárias para prevenir a ocorrência de injustiças epistêmicas. Quando um sistema fechado oculta os critérios para gerar informação ou conhecimento, sua barreira técnica atua como multiplicador de injustiças, silenciando a contestação e minando a agência dos sujeitos afetados (Mittelstadt *et al.*, 2016). Conforme destacado por Benjamin (2019) e Noble (2018), o isolamento estatístico de variáveis tende a reduzir a violência algorítmica a um mero “erro de calibração de código”, alimentando a ilusão tecnocêntrica de neutralidade e infalibilidade dos sistemas e induzindo o pressuposto de que correções pontuais em software seriam suficientes para sanar a injustiça epistêmica.

Diante do exposto, verifica-se que os três estudos^(E02, E06, E11) possuem limitações que, quando olhadas juntas, mostram diferentes pontos cegos da pesquisa em IA. O estudo E02 sofre com o número reduzido de modelos analisados e versões mais antigas, tornando seus resultados preliminares, requerendo continuidade e aprofundamento. Em paralelo, a abordagem quantitativa do estudo E06 não captura aspectos visuais e culturais mais detalhados, exigindo validação humana dos resultados e maior atenção à precisão. Por fim, o estudo E11 se restringe ao contexto bengali em apenas três dimensões identitárias, além de depender de conjuntos de dados públicos pouco representativos e ter acesso limitado aos algoritmos das ferramentas testadas, o que limita a generalização e a acurácia dos resultados.

Para superar essas barreiras, seria importante ampliar o escopo das categorias e idiomas estudados, diversificar e balancear os conjuntos de dados analisados, incluir mais modelos e versões recentes, combinar métodos quantitativos e qualitativos e garantir análises internas e externas. Contudo, é essencial reconhecer as limitações de recursos e acesso que marcam pesquisas no âmbito das ciências sociais aplicadas, em que frequentemente há escassez de dados de qualidade, falta de infraestrutura tecnológica e dificuldade de acesso a modelos mais recentes.

Sob uma perspectiva crítica, essa disparidade de recursos não é um acidente histórico, mas uma manifestação direta do poder informacional concentrado nas mãos de corporações tecnológicas do Norte Global. A barreira técnica imposta funciona como uma estratégia de blindagem corporativa que impossibilita a auditoria independente e o escrutínio democrático externo, alimentando a falsa aura de neutralidade e infalibilidade desses sistemas (O’Neil, 2020). Conforme denunciam Benjamin (2019) e Eubanks (2018), ao monopolizarem o acesso à infraestrutura computacional e ao sonegarem a transparência de suas bases de treinamento, as *Big Techs* estabelecem uma barreira que desautoriza e deslegitima a capacidade investigativa de cientistas localizados fora desses círculos hegemônicos, configurando uma violência

epistêmica estrutural que silencia visões dissidentes e pesquisas de interesse público vindas do Sul Global (Judijanto, 2025; Khanna; Li, 2025).

Investir em parcerias institucionais, fomentar iniciativas de coleta de dados culturalmente representativos e democratizar o acesso a ferramentas avançadas pode ajudar a reduzir essas desigualdades e conferir mais robustez aos processos de análise. Contudo, para além de meras reformas técnicas ou ajustes de calibração de código, a verdadeira superação dessas barreiras exige a descentralização de poder e o desenvolvimento de infraestruturas acadêmicas que permitam às comunidades e universidades periféricas validarem seus próprios recursos e narrativas.

Dando continuidade à discussão, o próximo eixo apresenta as abordagens metodológicas voltadas para o mapeamento de assimetrias estruturais. Essas abordagens aprofundam a compreensão das raízes dos vieses existentes e demonstram como lacunas léxicas, padrões de exclusão e hierarquias culturais são incorporados nos dados que alimentam a IA, indo um pouco mais além da identificação de padrões de violência epistêmica revelados nos vieses de interpretação e representação.

5.2 Mapeamento de assimetrias estruturais

A algoritmização das interações sociais e comerciais sustenta falácias como a “vida acessível” e a “visibilidade igualitária” para aqueles que dispõem de recursos tecnológicos capazes de criar e expor identidades digitais. Entretanto, ao desviarmos a atenção para o que acontece nas redes, ignoramos os impactos concretos dessa virtualização em massa. A ilusão de que todos estamos igualmente conectados, com os mesmos recursos disponíveis, somada às bolhas de interesse moldadas pelo comportamento online, reforça a crença de que nos aproximamos de uma utopia inclusiva, quando, na realidade, aprofundamos desigualdades e invisibilidades (Eubanks, 2018).

Entre as abordagens metodológicas identificadas nos estudos incluídos, destaca-se o Mapeamento de assimetrias estruturais, utilizado para avaliar como estas desigualdades políticas, econômicas e culturais se reproduzem nos ecossistemas digitais e nos sistemas de IA. Presente nos estudos E01, E03, E04, E07 e E10, essa estratégia analítica, consiste em cruzar dados demográficos reais com representações digitais em repositórios massivos, revelando padrões de invisibilidade e distorção.

Os resultados obtidos no estudo E01 apontam a persistência de vieses de gênero nos sistemas analisados nos estudos de caso investigados, principalmente na representação de mulheres retratadas mais negativamente, estereotipadas em papéis feminilizados e associadas a características femininas. O estudo E03 demonstrou que a própria arquitetura das plataformas perpetua desigualdades estruturais, ao invalidar denúncias de grupos marginalizados em sistemas de *compliance* anônimo, e que o uso sistemático do inglês como língua padrão atua como mecanismo de redução cultural, obscurecendo termos locais e experiências minoritárias. Essa invisibilidade linguística foi reforçada pelos estudos E04 e E07, que apontaram o apagamento de línguas faladas por milhões de pessoas e a erosão conceitual em bases multilíngues, enquanto E10 evidenciou a estigmatização de comunidades específicas nos conteúdos gerados por IA.

Diferentemente dos estudos E01, E03, E04, E07 e E10, que aplicam o cruzamento entre dados demográficos e representações digitais para evidenciar apagamentos e desigualdades, os estudos E08 e E05 concentram-se em discutir a dimensão política e cultural da produção de conhecimento mediada por IA. Ambos ressaltam que os apagamentos epistêmicos não são apenas efeitos técnicos, mas também processos de exclusão social e cultural, e enfatizam que práticas coletivas de resistência - como contribuições comunitárias e estratégias de visibilidade - podem ampliar o espaço de grupos marginalizados. Assim, enquanto o mapeamento de assimetrias estruturais fornece evidências empíricas e comparativas sobre invisibilidade e distorção, os estudos E08 e E05 ampliam o debate ao destacar que a disputa por representação é também uma questão de poder, agência e resistência, situando a injustiça epistêmica algorítmica em um campo mais amplo de lutas sociais e culturais.

Embora o mapeamento de assimetrias estruturais tenha se mostrado eficaz para expor padrões de invisibilidade e exclusão nos estudos supracitados, as limitações apontadas pelos próprios autores revelam fragilidades metodológicas e operacionais que precisam ser consideradas na interpretação dos resultados e na condução de pesquisas similares. Dentre as limitações apontadas, destacam-se: em E01, o escopo restrito a análise de estudos de caso sobre questões de gênero e desigualdade, limitando a generalização dos resultados; em E03, a falta de evidências quantitativas, que o torna dependente de futuras pesquisas empíricas e etnográficas; em E04, o uso apenas de dados da *web* pública, com categorização rígida, ausência de dados longitudinais e falta de abordagem interseccional e participativa; em E07, barreiras operacionais e técnicas, como desigualdade de acesso a bases de dados privadas, incompatibilidade de *softwares* e limitações semântica da ferramenta de mineração de texto

adotada; e em E10, o foco exclusivo em dimensões linguísticas, pouca atenção ao *design* e às implicações ético-políticas dos sistemas, além do risco de generalização e perda de significado em narrativas convertidas para uso algorítmico.

Ainda assim, sua aplicação pode ser fortalecida se forem adotadas medidas que substituam classificações rígidas por métricas graduais, incluam dados longitudinais e combinem análises quantitativas com validação qualitativa. Estas sugestões são factíveis e podem ser adotadas em curto a médio prazo, sem necessidade direta de mobilização institucional ou muitos recursos. Outras sugestões que envolvem maior coordenação institucional, seriam a ampliação do acesso a bases de dados privadas e não-hegemônicas e a interoperabilidade entre *softwares* e infraestruturas acadêmicas. Por último, mas de maneira alguma menos importante, a adoção de abordagens interseccionais e participativas com envolvimento direto das comunidades marginalizadas estudadas, talvez demande mais tempo, recursos e mudanças epistemológicas, mas é fundamental para legitimar os resultados e reduzir o silenciamento epistêmico.

Em resumo, todas as recomendações propostas são factíveis, mas variam em grau de viabilidade: as técnicas e metodológicas podem ser implementadas rapidamente, enquanto as estruturais e participativas requerem esforços institucionais e políticos mais amplos, configurando um horizonte progressivo de aplicação que fortalece a capacidade crítica e inclusiva dessa abordagem.

Ao cultivarmos deliberadamente uma abordagem solidária ao *design*, precisamos considerar que a tecnologia que pode estar funcionando muito bem para alguns de nós pode prejudicar ou excluir outros e, eventualmente, poderemos ser os próximos excluídos. Como declarado por Eubanks (2018) e reforçado por Benjamin (2019), O’Neil (2020) e Noble (2021), a opressão algorítmica não é apenas uma falha no sistema, ela já se tornou fundamental para o sistema operacional da rede.

Considerando o que foi exposto nos eixos 4.1 e 4.2, o último eixo desta seção aborda metodologias participativas que, além de analisar o contexto e os efeitos da IA na produção de injustiças epistêmicas, buscam evidenciar como a inclusão dos grupos afetados pode tornar as novas soluções mais efetivas e eficazes.

5.3 Metodologias participativas

As metodologias participativas colocam comunidades historicamente marginalizadas no centro da produção de conhecimento e da governança tecnológica. Em vez de tratá-las como objetos passivos, essas abordagens reconhecem sua agência e transformam experiências locais em insumos essenciais para a avaliação e o desenho de sistemas. Ao articular narrativas coletivas, práticas reflexivas e estratégias de resistência digital, elas ampliam a robustez das investigações quantitativas e reposicionam a pesquisa em IA como um processo ético e político.

A aplicação mais concreta dessa abordagem foi adotada na construção do protocolo “*Collective Narrative Grounding*”, desenvolvido no estudo E08, no qual os autores realizaram oficinas presenciais de cartografia social combinadas com técnicas de processamento de linguagem natural para transformar histórias de moradores sobre seus territórios em unidades narrativas estruturadas e validadas pela própria comunidade. Esse processo não apenas preserva a memória coletiva, mas também a torna indexável e reutilizável em sistemas automatizados de recuperação de informações. Ao integrar os saberes locais a partir da perspectiva de um grupo de residentes, os autores mostraram que narrativas contextualizadas podem corrigir falhas recorrentes e pontos cegos presentes nos LLMs generalistas disponíveis no mercado.

Essa transformação das dinâmicas de poder metodológico é contrastada na abordagem autorreflexiva proposta pelo estudo E05, que defende que qualquer auditoria séria sobre a injustiça algorítmica deve começar pelo escrutínio do próprio cientista e de seu ambiente institucional. Ao adaptarem a técnica metodológica da “Flor de Perguntas”, os autores constroem diretrizes para mapear seus próprios compromissos biográficos, ético-políticos e as assimetrias pragmáticas e financeiras que dividem os que estão localizados no Norte e no Sul Global, argumentando que a própria escolha das ferramentas de busca e bases de dados utilizadas para análise é condicionada por restrições econômicas de acesso que enviesam os resultados desde o começo da pesquisa.

Assim como os estudos anteriormente discutidos, os estudos E05 e E08 também possuem limitações que evidenciam que, embora as metodologias participativas tragam avanços, elas ainda enfrentam barreiras significativas. No estudo E05, os entraves foram práticos: 1) a desigualdade de recursos entre universidades parceiras restringiu o acesso a bases de dados proprietárias; 2) regras de segurança digital dificultaram a comunicação entre os autores, fazendo com que grande parte da pesquisa fosse compartilhada e discutida fora das plataformas institucionais; 3) incompatibilidades técnicas tornaram parte dos resultados ilegíveis. Além disso, as ferramentas de IA analisadas mostraram pouca capacidade de

interpretar textos fora do padrão ocidental, exigindo supervisão e intervenções humanas constantes.

No estudo E08, os desafios apontados pelos autores circundaram os campos ético e social, a partir dos quais verificou-se que transformar histórias em dados pode reduzir seu sentido, privilegiar vozes mais visíveis e gerar tensões entre a anonimização e a utilidade dos dados. Também ficou claro que comunidades não são homogêneas e que o consentimento sobre o uso das narrativas pode mudar com o tempo, especialmente em contextos de vigilância ou estigma.

Para superar esses limites, algumas medidas poderiam ser adotadas. No caso de E05, ampliar o uso de bases abertas e ferramentas interoperáveis ajudaria a reduzir desigualdades e evitar perdas técnicas, enquanto protocolos de colaboração mais flexíveis poderiam proteger a privacidade sem dificultar o trabalho conjunto entre membros de diferentes instituições. Já no estudo E08, mecanismos de moderação comunitária poderiam garantir espaço para vozes menos documentadas, e práticas de consentimento contínuo assegurariam aos participantes o direito de manter o controle sobre suas narrativas. Além disso, tecnologias mais sensíveis às nuances culturais e linguísticas poderiam preservar o significado das falas. Com esses ajustes, as metodologias participativas ganhariam mais força tanto no rigor acadêmico quanto na legitimidade ética.

A pulverização de metodologias participativas e de adição de contexto encontra sua justificativa teórica definitiva em Hopster (2024), cuja perspectiva sobre tecnologias socialmente disruptivas direciona as conclusões desta revisão. O autor argumenta que, uma vez que os sistemas baseados em IA atuam como agentes que alteram as próprias estruturas de como a sociedade produz, acessa e confia no conhecimento, os métodos tradicionais de análise baseados exclusivamente em métricas tornam-se insuficientes e obsoletos.

Nesse ponto, a contribuição de Fricker (2024) é decisiva: ao conceituar injustiça testimonial e hermenêutica, a autora mostra que metodologias participativas são instrumentos de redistribuição de credibilidade e de reconstrução dos recursos coletivos de interpretação, permitindo que grupos historicamente silenciados recuperem voz e reconhecimento. Em paralelo, Floridi (2024) reforça que a ética da IA exige responsabilidade moral e justiça contextual, legitimando a descentralização da governança e o *co-design* como caminhos necessários.

Em última instância, é apenas quando o poder de controle, a agência e a validação semântica das ferramentas retornam às mãos das próprias comunidades que se torna possível

reverter casos de injustiça e garantir que seus saberes sejam reconhecidos em pé de igualdade pelos sistemas automatizados. Essa transferência não é apenas técnica, mas política e ética, perpassa a devolução da voz, da credibilidade e do protagonismo a grupos historicamente silenciados e invisibilizados, assegurando que a IA não se sobreponha às experiências humanas, mas se torne um instrumento de fortalecimento coletivo. Ao recolocar os saberes locais no centro do desenvolvimento tecnológico, abre-se o caminho para uma ciência mais justa, plural e comprometida com a dignidade das pessoas que são as impactadas diretamente pela tecnologia.

6 CONCLUSÃO

Esta revisão teve como objetivo mapear e sistematizar as estratégias analíticas e metodológicas utilizadas na literatura científica para identificar e avaliar a ocorrência de injustiça epistêmica em sistemas baseados em IA. A partir dos resultados obtidos, foi possível verificar que os fenômenos da injustiça algorítmica e do silenciamento de grupos marginalizados tem sido capturado por meio de três modelos metodológicos complementares: o primeiro voltado para a realização de análises quantitativas baseadas em *probing* semântico e estresse computacional; o segundo relacionado ao mapeamento ecossistêmicos de assimetrias em conjuntos de dados de grande escala, através de análise quanti-qualitativa; e o terceiro com foco na participação ativa dos membros das comunidades afetadas pela generalização das ferramentas de IA amplamente utilizadas.

Nesse cenário, a principal conclusão que emerge desta pesquisa é que abordagens focadas exclusivamente em métricas de paridade estatística, apesar de necessárias, são insuficientes para responder à complexidade do problema. Os achados indicam que o combate eficaz à opacidade algorítmica e à consequente deferência cega dos usuários requer uma transição urgente em direção a *frameworks* de governança orientados à reparação social. Isso implica descentralizar o poder das grandes corporações tecnológicas e desenvolver infraestruturas que garantam soberania de dados e protagonismo metodológico das comunidades locais na validação de suas próprias narrativas. Nesse sentido, a inclusão de vozes diversas deve ser fomentada por estruturas dedicadas e equilibradas, já que tecnologias criadas para ampliar transparência podem se converter em instrumentos sofisticados de supressão, exigindo análises obrigatórias centradas na equidade e acessibilidade.

Para além dos ajustes nos modelos e bancos de dados institucionais, esta revisão demonstra a necessidade de expandir os horizontes de investigação futura em direção a ações

emancipatórias, com ênfase no fomento à conscientização e no desenvolvimento de tecnologias educacionais assistivas. Uma vez que a opacidade dos sistemas penaliza de forma desproporcional sujeitos com menor capital técnico e instrucional, o *design* de interfaces multimodais, didáticas e adaptadas a dialetos periféricos ou necessidades específicas torna-se indispensável para descentralizar a capacidade de escrutínio técnico. Sob essa ótica, o letramento digital crítico deixa de ser um complemento formativo e passa a ser compreendido como salvaguarda ética e direito humano fundamental na era da automação.

Por fim, recomenda-se que pesquisas futuras busquem cada vez mais superar o viés de amostras homogêneas ocidentalizadas e conduzir estudos que testem de maneira combinada os impactos da opacidade e da explicabilidade em ambientes de alta vulnerabilidade, como serviços públicos. Sugere-se ainda que comitês de governança e agências de financiamento estabeleçam avaliações prévias de impacto como requisito obrigatório no ciclo de vida de qualquer sistema inteligente em desenvolvimento. Nesse cenário, a principal contribuição desta pesquisa reside na sistematização inédita de um arcabouço metodológico tripartite - que integra a análises algorítmicas quantitativas, o mapeamento de assimetrias estruturais e as abordagens participativas -, oferecendo à comunidade científica uma ferramenta analítica robusta que supera a insuficiência das métricas de paridade estatística e instrumentaliza de forma prática a detecção do silenciamento epistêmico na era digital. É a partir desse mapeamento que se torna viável garantir a paridade de reconhecimento dos saberes humanos e transformar os usuários em agentes politicamente alfabetizados diante dos sistemas automatizados, orientando o avanço da IA como vetor de inclusão democrática e preservação da diversidade cultural.

7 LIMITAÇÕES

As limitações desta revisão de escopo decorrem, primeiramente, do seu próprio formato metodológico, que, por se tratar de um mapeamento amplo, privilegia a identificação de tendências e lacunas sem avaliar em profundidade a qualidade interna ou o risco de viés dos trabalhos incluídos. Isso torna os resultados estruturalmente mais descritivos do que conclusivos, funcionando como um ponto de partida para futuras investigações. No entanto, para além das barreiras procedimentais, este estudo enfrentou limitações estruturais e geopolíticas que refletem o próprio estado de desenvolvimento do campo de pesquisa em IA.

A primeira delas reside na escassez de estudos empíricos que investiguem a injustiça epistêmica em ambientes práticos de aplicação, uma vez que a literatura científica recente ainda

se encontra majoritariamente concentrada em debates teóricos ou ensaios conceituais. Esse cenário é agravado pela severa predominância de produções e afiliações institucionais do Norte Global, o que limita a diversidade cultural e geográfica das análises e deixa de fora realidades locais do Sul Global, onde os impactos da opacidade algorítmica são frequentemente mais devastadores. Adicionalmente, a própria flutuação e dificuldade conceitual do termo "injustiça epistêmica", que se desdobram em terminologias correlatas, mas não sinônimas, desafiaram a definição de um vocabulário de busca unificado.

Como consequência direta, a estratégia de busca dependeu fortemente de descritores específicos que, embora rigorosos, podem ter omitido pesquisas cruciais que utilizavam jargões distintos. Essa barreira foi intensificada pelas limitações inerentes às grandes bases de dados acadêmicas utilizadas, que historicamente priorizam periódicos de língua inglesa e indexadores comerciais do Norte Global, impondo barreiras invisíveis à recuperação de publicações e saberes gerados por universidades periféricas. Assim, os achados desta revisão de escopo não devem ser interpretados como um diagnóstico definitivo, mas como uma cartografia inicial e um convite necessário para investigações posteriores que sejam mais sistemáticas, descentralizadas e metodologicamente inclusivas.

REFERÊNCIAS

BAEYAERT, Joffrey. Epistemic injustice in Generative AI: a pipeline taxonomy, empirical hypotheses, and stage-matched governance. **EthAlca**, [S. l.], v. 4, n. 417, p. 1-16, 2025. DOI: <https://doi.org/10.56294/ai2025417>

BARRY, Isobel; STEPHENSON, Elise. The gendered, epistemic injustices of Generative AI. **Australian Feminist Studies**, [S. l.], v. 40, n. 123, p. 1-21, 2025. DOI: <https://doi.org/10.1080/08164649.2025.2480927>

BENJAMIN, Ruha. **Race after technology**: abolitionist tools for the New Jim Code. Cambridge: Polity Press, 2019.

DAS, Dipto; GUHA, Shion; SEMAAN, Bryan. How do datasets, developers, and models affect biases in a low-resourced language? *In*: INTERNATIONAL CONFERENCE ON EQUITY AND ACCESS IN ALGORITHMS, MECHANISMS, AND OPTIMIZATION (EAAMO '25), 5., 2025, Pittsburgh. **Proceedings** [...]. New York: ACM, 2025. p. 1-19. DOI: <https://doi.org/10.1145/3814890.3814910>

EUBANKS, Virginia. **Automating inequality**: how high-tech tools profile, police, and punish the poor. Nova Iorque: Picador - St. Martins' Press, 2018.

FLORIDI, Luciano. **A ética da inteligência artificial**. Curitiba: PUCPress, 2024.

FRICKER, Miranda. **Injustiça epistêmica: poder e a ética do conhecimento**. São Paulo: Editora da Universidade de São Paulo, 2024.

GAO, Zihan; YOUSUFI, Mohsin Y. K.; THEBAULT-SPIEKER, Jacob. Collective narrative grounding: community-coordinated data contributions to improve local AI systems. *In: WORKSHOP ON ALGORITHMIC COLLECTIVE ACTION (ACA@NEURIPS 2025)*, 39., 2025, Madison. **Papers** [...]. Madison: University of Wisconsin-Madison, 2025. p. 1-9.

HELM, Paula *et al.* Diversity and language technology: how language modeling bias causes epistemic injustice. **Ethics and Information Technology**, Dordrecht, v. 26, n. 8, p. 1-15, 2024. DOI: <https://doi.org/10.1007/s10676-023-09742-6>.

HOFFMANN, A. L. Terms of inclusion: data, discourse, violence. **New Media & Society**, Reino Unido, v. 23, n. 12, p. 3539-3556, 2021.

HOPSTER, J. K. G. Socially disruptive technologies and epistemic injustice. **Ethics and Information Technology**, Alemanha, v. 26, n. 14, 2024. DOI: <https://doi.org/10.1007/s10676-024-09747-9>

JUDIJANTO, Loso. Asylum processing algorithms and epistemic violence: a review of AI's role in refugee status determination. **Journal of Information, Technology and Policy**, Jakarta, v. 22, n. 2, p. 1-11, 2025. DOI: <https://doi.org/10.62836/jitp.2025.495>

KAY, Jackie; KASIRZADEH, Atoosa; MOHAMED, Shakir. Epistemic injustice in Generative AI. *In: AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY (AIES)*, 7., 2024, San Jose. **Proceedings** [...]. San Jose: Association for the Advancement of Artificial Intelligence, 2024.

KHANNA, Saurabh; LI, Xinxu. **Invisible languages of the LLM universe**. (No prelo). Amsterdam: Amsterdam School of Communication Research (ASCoR), University of Amsterdam, 2025. p. 1-26.

KRAFT, Angelie; SOULIER, Eloïse. Knowledge-enhanced language models are not bias-proof: situated knowledge and epistemic injustice in AI. *In: ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (FAccT '24)*, 2024, Rio de Janeiro. **Proceedings** [...]. New York: ACM, 2024. p. 1433-1445. DOI: <https://doi.org/10.1145/3630106.3658981>.

MBALAKA, Blessing. Epistemically violent biases in artificial intelligence design: the case of DALLE-E 2 and Starry AI. *Digital Transformation and Society*, South Africa, v. 2, n. 4, p. 376-402, 2023. DOI: <https://doi.org/10.1108/DTS-01-2023-0003>.

MITTELSTADT, B. *et al.* The ethics of algorithms: mapping the debate. **Big Data & Society**, Califórnia, v. 3, n. 2, p. 1-21, 2016. DOI: <https://doi.org/10.1177/205395171667967>

NOBLE, Safiya Umoja. **Algoritmos da opressão: como o Google fomenta e lucra com o racismo**. Santo André - SP: Editora Rua do Sabão, 2021.

O'NEIL, Cathy. **Algoritmos de destruição em massa: como o big data aumenta a desigualdade e ameaça a democracia**. Santo André - SP: Editora Rua do Sabão, 2020.

ORONÓZ, Soraya; PÉREZ, Albert Miró; PEÑA-MARTÍNEZ, Juan. Exploring the Merchera ethnic group through ChatGPT: the risks of epistemic exclusion. **Information**, [S. l.], v. 16, n. 6, 461, p. 1-16, 2025. DOI: <https://doi.org/10.3390/info16060461>

PETERS, M. D. J. *et al.* Scoping Reviews. In: AROMATARIS, E. *et al.* (ed.). **JBIM Manual for Evidence Synthesis**. JBI, 2024. DOI: <https://doi.org/10.46658/JBIMES-24-09>

STEUP, M.; NETA, R. **Epistemology**. Stanford: The Stanford Encyclopedia of Philosophy, 2020. Disponível em: <https://plato.stanford.edu/archives/spr2024/entries/epistemology/>. Acesso em: 31 maio 2026.

TRICCO, A. C. *et al.* PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. **Annals of Internal Medicine**, [Filadelfia], v. 169, n. 7, p. 467-473, 2018. DOI: <https://doi.org/10.7326/M18-0850>

TURING, Alan M. Computing machinery and intelligence. **Mind**, [S. l.], v. 59, n. 236, p. 433-460, 1950.

WALLE, Yelkal Muluaem. Breaking digital silence: AI, structural bias, and the uneven landscape of whistleblowing. **AI & SOCIETY: Journal of Knowledge, Culture and Communication**, [S. l.], v. 40, n. 15, p. 1-15, 2025. DOI: <https://doi.org/10.1007/s00146-025-02758-0>

ZERMEÑO-FLORES, Ana Isabel; TUFTE, Thomas; NAVARRETE-VEGA, Mabel Andrea. Knowledge production and epistemic injustices: the use of digital technologies and artificial intelligence. **MedieKultur: Journal of media and communication research**, Denmark, v. 40, n. 76, p. 120-138, 2024.