Racismo Algorítmico em Sistemas de Inteligência Artificial: Uma Revisão Sistemática da Literatura

Adriele Mesquita de Souza Campos, Natanael da Silva dos Santos, Douglas Tadeu Andrade da Silva, Libia de Souza Boss Cunha

[1]Instituto Federal de Mato Grosso (IFMT) - Campus Campo Verde
78841-006 – Campo Verde – MT – Brasil

diely60@gmail.com, natanaels2cv@gmail.com douglastadeu333@gmail.com, libia.boss@ifmt.edu.br

Abstract. *This article presents a systematic literature review on algorithmic racism in Artificial Intelligence (AI) systems, analyzing its social impacts. The study, based on the protocol proposed by Kitchenham and Charters (2007), examined 38 studies that supported the answers to the research questions. The results indicate that algorithmic racism manifests in different contexts, such as healthcare, public security, and digital platforms, and is influenced by biased data and institutional factors. Impacts such as discrimination and the amplification of inequalities are observed, highlighting the need for technical, ethical, and regulatory approaches for mitigation.*

Resumo. *Este artigo apresenta uma revisão sistemática da literatura sobre racismo algorítmico em sistemas de Inteligência Artificial (IA), analisando seus impactos sociais. A pesquisa, baseada no protocolo de Kitchenham e Charters (2007), analisou 38 estudos fundamentaram as respostas às questões de pesquisa. Os resultados indicam que o racismo algorítmico se manifesta em diferentes contextos, como saúde, segurança pública e plataformas digitais, sendo influenciado por dados enviesados e fatores institucionais. Observam-se impactos como discriminação e ampliação de desigualdades, evidenciando a necessidade de abordagens técnicas, éticas e regulatórias para mitigação.*

1. Introdução

O racismo pode ser compreendido como um sistema social de poder que produz e mantém desigualdades baseadas em hierarquias raciais, marginalizando grupos historicamente classificados como “inferiores”. Esse fenômeno não se limita a ações individuais, mas em estruturas sociais, políticas e econômicas [Almeida, 2019]. Com o avanço das tecnologias digitais, tais desigualdades passaram também a se manifestar em sistemas computacionais, especialmente em aplicações baseadas em Inteligência Artificial (IA).

Nesse contexto, emerge o conceito de racismo algorítmico, que se refere às práticas discriminatórias resultantes do uso de algoritmos capazes de reproduzir ou amplificar preconceitos historicamente presentes na sociedade. Esses sistemas são construídos a partir de grandes volumes de dados, modelos matemáticos e decisões humanas durante seu desenvolvimento e implementação, fatores que podem contribuir para a geração ou intensificação de desigualdades sociais [Silva, 2022].

O racismo algorítmico, portanto, não constitui apenas um problema técnico, mas sim uma manifestação significativa de estruturas históricas de opressões raciais. Munanga (1999) argumenta que, apesar da desconstrução do conceito de raça pela comunidade científica, sua estrutura social continua presente e se manifesta por meio de novas tecnologias.

Algoritmos de IA são treinados a partir de conjuntos de dados que frequentemente refletem os vieses existentes na sociedade. Como consequência, esses vieses podem se manifestar em diferentes aplicações tecnológicas, como sistemas de reconhecimento facial, classificação de imagens, análise de textos ou processos automatizados de tomada de decisão. Quando os dados utilizados no treinamento apresentam distorções ou representações desiguais, os modelos resultantes tendem a reproduzir e amplificar tais distorções.

Nesse sentido, o racismo algorítmico refere-se ao conjunto de mecanismos pelos quais desigualdades sociais e preconceitos estruturais influenciam o desenvolvimento e o funcionamento de tecnologias baseadas em algoritmos. Essas tecnologias podem, portanto, reforçar, intensificar ou até mesmo ocultar dinâmicas racialmente discriminatórias presentes na sociedade.

A análise de trabalhos anteriores evidencia que a discussão sobre vieses raciais e preconceitos em sistemas de Inteligência Artificial tem sido amplamente investigada na literatura científica recente. Estudos de revisão sistemática, como os conduzidos por Peres et al. (2021) e por Intahchomphoo e Gundersen (2020), demonstram que algoritmos de aprendizagem de máquina frequentemente reproduzem desigualdades presentes nos dados utilizados em seu treinamento, refletindo vieses relacionados a raça, gênero, condição socioeconômica, idade e deficiência. Essas distorções podem impactar áreas sensíveis da sociedade, como saúde, segurança pública, recrutamento e serviços financeiros, evidenciando que os sistemas de IA não são neutros, mas influenciados pelos contextos sociais e institucionais nos quais são desenvolvidos e aplicados. Nesse sentido, a revisão conduzida por Nyland (2023) contribui para o debate sobre racismo algorítmico, reforçando a necessidade de investigações críticas sobre o desenvolvimento e a aplicação dessas tecnologias, bem como a adoção de bases de dados mais representativas, métodos de avaliação mais transparentes e diretrizes éticas capazes de mitigar vieses raciais em sistemas automatizados.

A literatura aponta três principais dimensões nas quais o racismo pode se manifestar em sistemas de Inteligência Artificial. A primeira refere-se ao viés nos dados de treinamento, uma vez que os conjuntos de dados utilizados frequentemente refletem desigualdades históricas e sociais, como evidenciado no estudo *Gender Shades*, que demonstrou maiores taxas de erro em sistemas de reconhecimento facial para mulheres negras devido à predominância de rostos brancos nas bases de treinamento [Buolamwini; Gebru, 2018]. A segunda dimensão está relacionada ao viés no desenho algorítmico, pois a própria construção matemática dos modelos pode incorporar preconceitos quando métricas de desempenho são definidas sem considerar os contextos sociais de aplicação [O'neil, 2016]. Por fim, destaca-se o viés na implementação, que ocorre quando sistemas de IA são aplicados em ambientes já marcados por desigualdades estruturais, podendo reforçar ou ampliar essas desigualdades existentes [Browne, 2015; Almeida, 2021].

Apesar de frequentemente serem apresentados como tecnologias neutras e objetivas, sistemas de IA podem reproduzir e amplificar vieses estruturais presentes nas

sociedades contemporâneas. Estudos recentes têm evidenciado a presença de discriminação racial em diferentes aplicações tecnológicas, como demonstrado por Noble (2018), Buolamwini e Gebru (2018), Obermeyer et al. (2019), Intahchomphoo e Gundersen (2020), Peres et al. (2021) e Silva (2022). No entanto, a investigação científica sobre racismo algorítmico ainda se encontra em estágio inicial, levantando diversos questionamentos acerca de suas manifestações e impactos.

Diante desse cenário, esta pesquisa busca responder ao seguinte problema de investigação: como o racismo algorítmico se manifesta em sistemas de Inteligência Artificial e quais estratégias têm sido propostas para mitigar seus impactos? Para responder a essa questão, este estudo tem como objetivo central analisar a presença do racismo algorítmico em sistemas de IA e compreender seus impactos sociais, por meio de uma revisão sistemática da literatura.

Embora existam revisões anteriores sobre vieses e discriminação em sistemas de Inteligência Artificial, muitos desses estudos abordam o tema de forma ampla, contemplando diferentes tipos de preconceito algorítmico, como gênero, idade, classe social e etnia, sem aprofundar especificamente as manifestações do racismo algorítmico em sistemas contemporâneos de IA. Além disso, parte dessas revisões concentra-se em contextos anteriores à expansão recente de tecnologias como modelos generativos, sistemas avançados de reconhecimento facial e modelos de linguagem de larga escala.

Nesse contexto, esta revisão busca sistematizar evidências empíricas recentes sobre racismo algorítmico em diferentes domínios da Inteligência Artificial, incluindo reconhecimento facial, sistemas de decisão automatizada e modelos de linguagem. Diferentemente de revisões anteriores, o presente estudo integra perspectivas técnicas, sociais, éticas e regulatórias, contribuindo para uma compreensão mais abrangente das formas pelas quais desigualdades raciais vêm sendo reproduzidas e ampliadas por sistemas algorítmicos contemporâneos.

A pesquisa foi conduzida utilizando o método de Revisão Sistemática da Literatura (RSL), que permite identificar, avaliar e interpretar estudos relevantes sobre uma determinada questão de pesquisa ou área temática, conforme proposto por Kitchenham e Charters (2007). O estudo foi orientado pelas seguintes perguntas de pesquisa:

- QP1: Quais são os principais contextos de aplicação da Inteligência Artificial identificados na literatura e quais fatores contribuem para a manifestação do racismo algorítmico nesses sistemas?
- QP2: Quais são os casos históricos e contemporâneos mais relevantes que evidenciam danos causados por vieses raciais na IA?
- QP3: Quais são as implicações sociais, éticas e técnicas do preconceito racial em sistemas de Inteligência Artificial?

Com base nessas questões, foram analisados artigos selecionados a partir de palavras-chave previamente definidas, possibilitando uma investigação sobre como o racismo pode se manifestar nos sistemas de Inteligência Artificial e quais são os impactos dessas ocorrências para a sociedade. A análise dos estudos permitiu compreender diferentes abordagens e evidências apresentadas na literatura acerca do tema. O presente artigo está organizado da seguinte forma: a Seção 2 descreve o percurso metodológico

adotado na revisão, a Seção 3 apresenta e discute os resultados obtidos e a Seção 4 reúne as considerações finais do estudo.

2. Metodologia

A metodologia utilizada neste estudo foi proposta por Kitchenham e Charters (2007), seguindo um protocolo específico para revisões sistemáticas da literatura. Esse protocolo organiza o processo de investigação em três etapas principais: planejamento da revisão, condução da revisão e relato dos resultados.

A etapa de planejamento consistiu na definição do objetivo da pesquisa e na formulação das questões de investigação que orientaram a revisão, descritas na Seção 1 deste documento. Para apoiar a organização do protocolo e a gestão das etapas da revisão sistemática, foi utilizada a ferramenta *Parsifal* ®, uma plataforma amplamente empregada em revisões sistemáticas na área de computação, que auxilia na definição do protocolo, seleção dos estudos e extração dos dados.

2.1. Condução da Revisão Sistemática

Na etapa de condução da revisão, foi realizada uma busca sistemática em bases de dados científicas reconhecidas nas áreas de computação e tecnologia, especificamente ACM Digital Library e Scopus. Não foi estabelecido recorte temporal para a busca, com o objetivo de abranger a evolução histórica das discussões sobre racismo algorítmico na literatura científica e identificar tanto estudos clássicos quanto pesquisas recentes sobre o tema. A estratégia de busca foi construída com base no modelo PICOC (Population, Intervention, Comparison, Outcome, Context), utilizando termos relacionados a inteligência artificial, racismo algorítmico e mitigação de vieses. A string de busca combinou palavras-chave e seus sinônimos utilizando operadores booleanos: ("*Artificial intelligence*" OR "*AI*" OR "*machine learning*" OR "*Inteligência Artificial*") AND ("*algorithmic racism*" OR "*racismo algorítmico*" OR "*racial bias*" OR "*viés racial*") AND ("*mitigation*" OR "*fairness*" OR "*ethical*" OR "*mitigação*" OR "justiça" OR "*ética*").

A busca inicial retornou 690 estudos, sendo 567 da ACM Digital Library e 123 da Scopus. Após a remoção automática de duplicatas, restaram 613 estudos. Em seguida, foi realizada a triagem por meio da leitura de títulos, resumos e palavras-chave, aplicando-se os critérios de inclusão e exclusão previamente definidos, o que resultou em 149 estudos selecionados para leitura completa.

Os critérios de inclusão e exclusão utilizados na seleção dos estudos desta revisão sistemática foram definidos com o objetivo de garantir a relevância e a qualidade dos trabalhos analisados. Foram incluídos estudos primários e secundários que abordassem o racismo algorítmico em sistemas de Inteligência Artificial, publicados nos idiomas português, inglês ou espanhol, que apresentassem evidências empíricas, análises teóricas relevantes ou discussões relacionadas aos impactos sociais, éticos e técnicos do fenômeno, desde que estivessem disponíveis integralmente para análise. Por outro lado, foram excluídos estudos fora do escopo da pesquisa, trabalhos duplicados, artigos com dados insuficientes para análise, publicações em outros idiomas e estudos sem acesso ao texto completo. Após a leitura integral dos artigos, 105 estudos foram considerados elegíveis para a etapa de extração dos dados.

Na etapa de extração e análise dos dados, os estudos selecionados foram analisados por meio de um formulário estruturado, contendo informações como identificação do estudo, autores, país de publicação, contexto de aplicação da IA, tipos de viés identificados, impactos sociais e possíveis estratégias de mitigação. Cada estudo recebeu um identificador único no formato E1, E2, E3, …, permitindo a organização e rastreabilidade das evidências analisadas.

Os 105 estudos elegíveis foram utilizados na etapa de mapeamento e contextualização da literatura, permitindo identificar tendências gerais, abordagens metodológicas e contextos de aplicação da IA relacionados ao tema. Entretanto, para responder diretamente às questões de pesquisa propostas, 38 estudos fundamentaram a etapa de análise e discussão dos resultados, pois apresentavam evidências empíricas ou discussões diretamente alinhadas aos objetivos centrais desta revisão.

Por fim, na etapa de relato da revisão, os resultados foram sintetizados e organizados, com base na metodologia de análise de conteúdo de Bardin (2011). A análise buscou identificar padrões recorrentes na literatura sobre as manifestações do racismo algorítmico, os contextos de aplicação da IA e suas implicações sociais, éticas e técnicas. Os resultados obtidos são apresentados e discutidos na seção seguinte.

3. Análise e Discussão dos Resultados

A análise e discussão dos resultados desta revisão sistemática foram fundamentadas em 38 estudos diretamente relacionados às questões de pesquisa propostas, apresentados no Quadro 1. A partir desses estudos, foram identificados padrões recorrentes sobre as manifestações do racismo algorítmico, os contextos de aplicação da Inteligência Artificial e seus impactos sociais, éticos e técnicos.

Quadro 1. Estudos Utilizados para Responder as Questões de Pesquisa

ID	Ano	Referência
E1	2025	TANKSLEY, Tiera et al. “Ethics is not neutral”: Understanding Ethical and Responsible AI Design from the Lenses of Black Youth.
E2	2024	SMITH, Julie M. “I'm Sorry, but I Can't Assist”: Bias in Generative AI.
E3	2022	HARRINGTON, Christina N. et al. “It’s Kind of Like Code-Switching”: Black Older Adults’ Experiences with a Voice Assistant for Health Information Seeking.
E4	2025	HALE, J.; KIM, P. H.; GRATCH, J. Probably fair algorithms may perpetuate racial and gender bias: a study of salary dispute resolution.
E5	2024	HAN, Jessy Xinyi et al. A Causal Framework To Evaluate Racial Bias in Law Enforcement Systems.
E7	2018	YANG, Ke et al. A Nutritional Label for Rankings.
E9	2025	HAIMSON, Oliver L. et al. AI Attitudes Among Marginalized Populations in the U.S.: Nonbinary, Transgender, and Disabled Individuals Report More Negative AI Attitudes.
E10	2022	SAPIEZYNSKI, Piotr et al. Algorithms that “Don't See Color”: Measuring Biases in Lookalike and Special Ad Audiences.
E11	2021	METAXA, Danaë et al. An Image of Society: Gender and Racial Representation and Impact in Image Search Results for Occupations.
E13	2024	HAQUE, MD Romael et al. Are We Asking the Right Questions?: Designing for Community Stakeholders’ Interactions with AI in Policing.
E14	2023	WÓJCIK, Malwina Anna. Assessing the Legality of Using the Category of Race and Ethnicity in Clinical Algorithms: the EU Anti-Discrimination Law Perspective.
E15	2025	IMANA, Basileal; KOROLOVA, Aleksandra; HEIDEMANN, John. Auditing for Bias in Ad Delivery Using Inferred Demographic Attributes.

E17	2025	SHAHID, Sumaiya Binte et al. Bias in Deep Learning Skin Cancer Detection: Parallel Residual Convolution Network Classification and Racial Bias Quantification.
E19	2021	SRINIVASAN, Ramya; CHANDER, Ajay. Biases in AI Systems: A Survey for Practitioners.
E22	2020	PÄÄKKÖNEN, Juho et al. Bureaucracy as a Lens for Analyzing and Designing Algorithmic Systems.
E23	2023	ALLAREDDY, Veerasathpurush et al. Call for algorithmic fairness to mitigate amplification of racial biases in artificial intelligence models used in orthodontics and craniofacial health.
E24	2021	MARKS, Paul. Can the biases in facial recognition be fixed; also, should they?
E31	2021	LI, Jinyang; MOSKOVITCH, Yuval; JAGADISH, H. V. DENOUNCER: Detection of unfairness in classifiers.
E32	2025	HAMID, Tarek et al. DermaGlow: Objective Quantification of Melanin, Erythema and Skin-tone Using Wearable Optical Spectroscopy.
E39	2021	ZHANG, Lingfeng; ZHANG, Yueling; ZHANG, Min. Efficient white-box fairness testing through gradient search.
E42	2022	BAEZA-YATES, Ricardo. Ethical Challenges in AI.
E45	2021	COE, J.; ATAY, M. Evaluating Impact of Race in Facial Recognition across Machine Learning and Deep Learning Algorithms.
E48	2019	KHADEMI, Aria et al. Fairness in Algorithmic Decision Making: An Excursion Through the Lens of Causality.
E49	2022	PUYOL-ANTÓN, Esther et al. Fairness in Cardiac Magnetic Resonance Imaging: Assessing Sex and Racial Bias in Deep Learning-Based Segmentation.
E50	2024	CHEN, Zhenpeng et al. Fairness Testing: A Comprehensive Survey and Analysis of Trends.
E51	2017	GALHOTRA, Sainyam; BRUN, Yuriy; MELIOU, Alexandra. Fairness testing: testing software for discrimination.
E53	2020	BLACK, Emily; YEOM, Samuel; FREDRIKSON, Matt. FlipTest: fairness testing via optimal transport.
E55	2025	JOHNSON, D. K. N. Gaslighting Ourselves: Racial Challenges of Artificial Intelligence in Economics and Finance Applications.
E68	2022	GRABOWICZ, Przemyslaw A.; PERELLO, Nicholas; MISHRA, Aarshee. Marrying Fairness and Explainability in Supervised Learning.
E70	2020	JIN, Zhongjun et al. MithraCoverage: A System for Investigating Population Bias for Intersectional Fairness.
E71	2023	ORPHANOU, Kalia; OTTERBACHER, Jahna; KLEANTHOUS, Styliani et al. Mitigating Bias in Algorithmic Systems—A Fish-eye View.
E73	2022	KOSTICK-QUENET, K. M.; COHEN, I. G.; GERKE, S. et al. Mitigating Racial Bias in Machine Learning. Journal of Law, Medicine & Ethics, 2022.
E74	2023	ABHARI, Julian; ASHOK, Ashwin. Mitigating Racial Biases for Machine Learning Based Skin Cancer Detection.
E86	2025	YUCER, Seyma et al. Racial Bias within Face Recognition: A Survey.
E100	2024	SCHAAP, A.; KITHARIDIS, S.; VAN STEIN, N. Towards Fairness in Machine Learning: Balancing Racially Imbalanced Datasets Through Data Augmentation and Generative AI.
E102	2021	ZHAO, D. et al. Understanding and Evaluating Racial Biases in Image Captioning.
E103	2024	VAN DER WAL, Oskar et al. Undesirable Biases in NLP: Addressing Challenges of Measurement.
E104	2018	STEELS, Luc. What needs to be done to ensure the ethical use of AI

Observa-se, a partir dos estudos apresentados no Quadro 1, uma concentração de publicações recentes, especialmente entre 2021 e 2025, evidenciando o crescimento das discussões sobre racismo algorítmico em sistemas de Inteligência Artificial. Esse aumento acompanha a expansão de aplicações baseadas em IA generativa,

reconhecimento facial, sistemas automatizados de decisão e modelos de linguagem, indicando que o tema tem ganhado maior relevância científica, social e regulatória nos últimos anos.

3.1. Resposta para as questões de pesquisa

RQP1: Quais são os principais contextos de aplicação da Inteligência Artificial identificados na literatura e quais fatores contribuem para a manifestação do racismo algorítmico nesses sistemas?

A análise dos estudos selecionados demonstra que o racismo algorítmico se manifesta em diferentes contextos de aplicação da Inteligência Artificial, especialmente em áreas socialmente sensíveis, como segurança pública, saúde, educação, mercado digital e sistemas governamentais automatizados. Os estudos evidenciam que, embora esses sistemas sejam frequentemente apresentados como tecnologias neutras e objetivas, seu funcionamento tende a reproduzir desigualdades históricas presentes nos dados, nas instituições e nos processos de desenvolvimento tecnológico.

Em contextos de vigilância e segurança pública, como policiamento preditivo e sistemas judiciais, esses sistemas podem reforçar padrões de criminalização de comunidades negras e socialmente vulneráveis. Na saúde, a baixa representatividade de pessoas negras em bases de dados clínicos pode resultar em diagnósticos menos precisos. Já em ambientes educacionais e corporativos, sistemas de linguagem e recomendação podem associar nomes ou características raciais a significados negativos ou restringir oportunidades.

O Quadro 2 sintetiza as principais aplicações da IA e os fatores associados à manifestação do racismo algorítmico identificados nos estudos analisados.

Quadro 2. Aplicações da IA e contextos associados ao racismo algorítmico

Aplicação de IA	Setores	Contextos associados ao racismo algorítmico	Estudos
Reconhecimento facial e vigilância	Segurança pública, escolas, espaços urbanos	Desigualdade estrutural; controle institucional sobre corpos negros	E1, E5, E13
Modelos de linguagem (LLMs)	Educação, aconselhamento estudantil	Dados de treinamento com vieses raciais e socioeconômicos; nomes como proxy racial	E17, E24, E32
Algoritmos médicos e de diagnóstico	Saúde	Dados clínicos racializados; sub-representação de pessoas com pele escura	E7, E14
Assistentes de voz e reconhecimento de fala	Tecnologia doméstica, comunicação	Racismo linguístico e hegemonia cultural branca	E17, E24
Sistemas de recomendação e anúncios	Mercado digital, redes sociais	Inferências raciais indiretas; exclusão de usuários negros	E24, E32
Algoritmos de decisão social e governamental	Políticas públicas, serviços sociais	Instituições burocráticas reproduzem desigualdades de raça e classe	E1, E5, E13

Os resultados indicam que o racismo algorítmico não decorre apenas de falhas técnicas isoladas, mas também de fatores estruturais e institucionais presentes no desenvolvimento da Inteligência Artificial, como bases de dados enviesadas, baixa diversidade nas equipes e ausência de auditorias algorítmicas. Dessa forma, o fenômeno configura-se como um problema sistêmico, no qual limitações técnicas e desigualdades sociais atuam conjuntamente na reprodução de injustiças raciais em diferentes aplicações da IA.

RQP2: Quais são os casos históricos e contemporâneos mais relevantes que evidenciam danos causados por racismo algorítmico na IA?

A análise dos estudos selecionados (E2, E4, E7, E9, E11, E15, E19, E23, E31, E42, E48) evidenciou diversos casos históricos e contemporâneos que demonstram os danos causados por racismo algorítmico em sistemas de IA. Esses casos abrangem diferentes setores, como justiça criminal, vigilância, saúde, recrutamento e publicidade digital. Entre os exemplos históricos, destaca-se o algoritmo de seleção da Escola de Medicina no Reino Unido (1988), que discriminava candidatos com nomes não europeus e mulheres, e o sistema COMPAS, amplamente criticado por classificar réus negros como mais propensos à reincidência criminal. Nos casos contemporâneos, destacam-se falhas em sistemas de reconhecimento facial que resultaram em prisões injustas de pessoas negras, além de discriminações em plataformas digitais de anúncios e classificações ofensivas em bases de dados de visão computacional. Esses episódios demonstram que sistemas de IA podem reproduzir e amplificar desigualdades raciais presentes nas bases de dados e nos contextos institucionais em que são implementados, conforme sintetizado no Quadro 3.

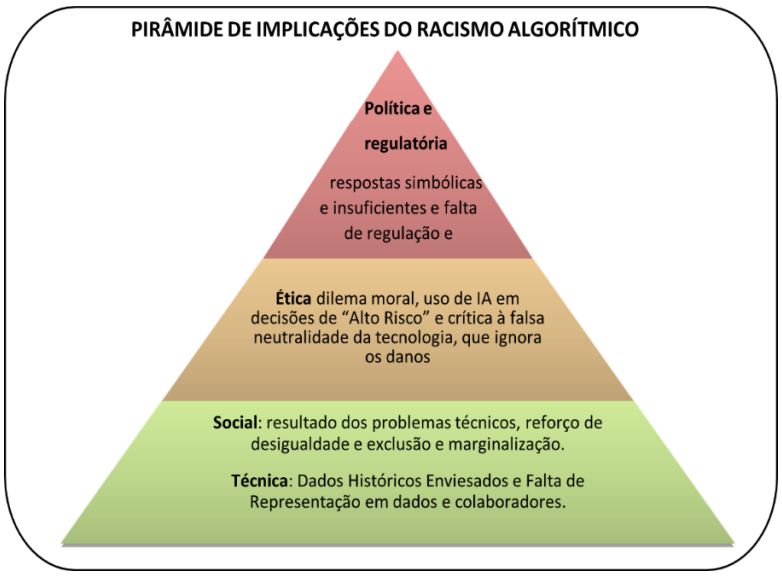
Quadro 3. Casos representativos de racismo algorítmico em sistemas de IA

Ano / Período	Sistema / Tecnologia de IA	Área de Aplicação	Impactos Identificados
1988 (Reino Unido)	Algoritmo de seleção da Escola de Medicina	Educação / Admissão	Discriminação contra candidatos com nomes não europeus e exclusão de mulheres.
2016 (EUA)	COMPAS – Avaliação de risco criminal	Justiça criminal	Réus negros classificados como mais propensos à reincidência, reforçando desigualdades no sistema judicial.
2018–2020 (EUA)	Sistemas de reconhecimento facial	Segurança pública / Vigilância	Prisões injustas de pessoas negras e maior taxa de erro para rostos negros.
2019 (Internacional)	ImageNet Roulette	Visão computacional	Rotulagem ofensiva e estereotipada associada a pessoas negras em bases de dados.
2019–2021 (EUA)	Facebook Ads / Google AdFisher	Publicidade e crédito	Exclusão de minorias raciais em anúncios de emprego, moradia e crédito.
2023–2024 (Global)	Modelos de linguagem (ex.: ChatGPT, LLaMA)	Educação / Geração de texto	Reprodução de estereótipos raciais e associações negativas a grupos minoritários.

RQP3: Quais São as Implicações Sociais, Éticas e Técnicas do Preconceito Racial em Sistemas de Inteligência Artificial?

Os estudos analisados indicam que os sistemas de Inteligência Artificial não são neutros e podem reproduzir ou até ampliar desigualdades estruturais existentes. Para sintetizar essas evidências, a Figura 1 apresenta uma pirâmide de implicações do racismo algorítmico, organizada como uma cadeia de causa e efeito. Na base da pirâmide encontram-se os fatores técnicos, considerados a origem do problema.

Figura 1. Pirâmide de Implicações do Racismo Algorítmico



Diversos estudos apontam que o racismo algorítmico surge principalmente do uso de dados históricos enviesados e da falta de representatividade nos conjuntos de dados e nas equipes de desenvolvimento. Zhao et al. (2021) (E102) demonstram um forte desequilíbrio em bases de imagens utilizadas no treinamento de sistemas de visão computacional, com até 23 vezes menos imagens de mulheres de pele escura em comparação a homens de pele clara. Além disso, Yucer et al. (2025) (E86) mostram que o racismo algorítmico pode ser introduzido em diferentes etapas do pipeline de engenharia, desde a coleta das imagens até o processamento e treinamento dos modelos. Johnson (2025) (E55) sintetiza esses desafios técnicos no acrônimo DO-GOOD, indicando que dados enviesados, opacidade do design e confiança excessiva em modelos matemáticos contribuem para decisões discriminatórias.

O segundo nível da pirâmide corresponde às implicações sociais, que representam os impactos concretos desses problemas técnicos. Quando sistemas de IA são treinados com dados que refletem desigualdades históricas, suas decisões tendem a reproduzir e amplificar essas desigualdades. Kostick-Quenet et al. (E73), por exemplo, demonstram que sistemas utilizados na área da saúde podem restringir o acesso de pacientes negros a determinados tratamentos, uma vez que os algoritmos aprendem padrões de sistemas de saúde já desiguais. De forma semelhante, Harrington et al. (2022) (E3) identificam a presença de racismo linguístico em assistentes de voz, que apresentam dificuldades em compreender o *African American Vernacular English (AAVE),* evidenciando como a falta de diversidade nos dados de treinamento pode resultar em exclusão tecnológica. Esses exemplos mostram que falhas técnicas se traduzem em processos de marginalização e reforço de desigualdades sociais.

O terceiro nível da pirâmide envolve as questões éticas geradas por esses impactos sociais. À medida que os danos se tornam evidentes, cresce o debate sobre a legitimidade do uso de IA em decisões de alto risco. Smith (2024) (E2) discute os dilemas morais associados ao uso de IA generativa em contextos sensíveis, como aconselhamento educacional, enquanto Coe e Atay (2021) (E45) questionam a legitimidade ética do uso de reconhecimento facial em segurança pública. De forma mais ampla, a automação de práticas burocráticas discriminatórias pode perpetuar o racismo estrutural ao conferir aparência de neutralidade técnica a processos excludentes (E22). Steels (E104) amplia esse debate ao alertar para os riscos éticos do uso de IA na manipulação da opinião pública e na amplificação de divisões sociais.

No topo da pirâmide encontram-se as respostas políticas e regulatórias, que representam a reação institucional aos problemas identificados. Entretanto, a literatura aponta que essas respostas ainda são limitadas. Orphanou et al. (E71) observam que, embora organizações como a ACM e a UNESCO tenham publicado recomendações éticas para o desenvolvimento responsável da IA, a implementação prática dessas diretrizes ainda é restrita. Sapiezynski et al. (2022) (E10) destacam que, em muitos casos, as respostas governamentais se limitam a ações judiciais ou acordos corporativos, sem a criação de estruturas regulatórias abrangentes. Como alternativa, Puyol-Antón et al. (2022) (E49) defendem a necessidade de abordagens regulatórias mais proativas, sugerindo que órgãos reguladores exijam auditorias de justiça algorítmica (fairness audits) antes da aprovação de sistemas de IA para uso em áreas sensíveis, como a saúde.

3.2. Discussão

A partir da análise dos estudos foi possível identificar padrões recorrentes que permitem responder às questões de pesquisa propostas. A análise dos estudos também evidencia que o racismo algorítmico não pode ser compreendido apenas como uma falha técnica isolada, mas como um fenômeno profundamente relacionado às estruturas sociais nas quais os sistemas de IA são desenvolvidos e aplicados.

Os estudos analisados apresentam evidências empíricas dos impactos e demonstram, por exemplo, que assistentes de voz amplamente utilizados, como Siri e Alexa, apresentam maior dificuldade em reconhecer a fala de pessoas negras e idosas, uma vez que os conjuntos de dados utilizados em seu treinamento são compostos majoritariamente por vozes de pessoas brancas e jovens (E39). De forma semelhante, sistemas de diagnóstico médico baseados em Inteligência Artificial podem apresentar menor precisão em pacientes negros quando os bancos de imagens utilizados no treinamento são majoritariamente compostos por indivíduos brancos (E73). Esses casos evidenciam como a ausência de diversidade nos dados compromete não apenas a precisão técnica dos sistemas, mas também a justiça e a equidade das decisões automatizadas.

Além das limitações relacionadas aos dados, os estudos também apontam que fatores institucionais e organizacionais contribuem para a persistência desses vieses. A baixa diversidade nas equipes de desenvolvimento e nos processos de coleta e seleção de dados tende a reforçar perspectivas limitadas na construção dos sistemas de IA (E50). Nesse contexto, bases de dados desbalanceadas (E70) podem resultar em decisões automatizadas injustas (E103), reforçando desigualdades já existentes na sociedade. Por outro lado, diferentes pesquisas também apontam possíveis estratégias de mitigação, como o uso de técnicas de balanceamento e aumento de dados para reduzir distorções nos conjuntos de treinamento (E100), bem como o desenvolvimento de métricas e testes de

justiça algorítmica que permitam identificar e corrigir vieses durante o processo de desenvolvimento (E50, E51, E53).

Para além das abordagens técnicas, a literatura destaca a importância de medidas institucionais e regulatórias para enfrentar o racismo algorítmico. Entre as estratégias sugeridas estão a exigência de maior transparência nos modelos utilizados (E68), a realização de auditorias independentes para avaliar a equidade dos sistemas e a implementação de regulações específicas para o uso de IA em setores sensíveis. Nesse sentido, alguns estudos defendem que órgãos reguladores devem exigir auditorias de justiça algorítmica como pré-requisito para a implementação de sistemas de IA em áreas críticas, como saúde e justiça criminal (E49), tratando o viés algorítmico não apenas como um problema ético, mas também como um risco tecnológico com implicações legais.

Os resultados desta revisão também dialogam com pesquisas anteriores que investigaram vieses em sistemas de Inteligência Artificial. Peres et al. (2021), por exemplo, identificaram a presença de preconceitos relacionados a gênero, raça, etnia e condição socioeconômica em aplicações de IA em áreas como saúde, segurança pública e recrutamento profissional. A presente revisão amplia esse diagnóstico ao analisar um conjunto mais abrangente de estudos e ao conectar esses vieses às suas raízes históricas e estruturais. De forma semelhante, a revisão conduzida por Intahchomphoo e Gundersen (2020), já apontava que sistemas de IA podem gerar oportunidades desiguais para diferentes grupos sociais. Entretanto, os resultados desta revisão mais recente indicam que esses problemas permanecem presentes e continuam sendo identificados em diferentes aplicações tecnológicas.

Outro ponto relevante é que os achados desta revisão reforçam que problemas identificados em estudos clássicos da literatura, como o Gender Shades de Buolamwini e Gebru (2018), não representam casos isolados. Pelo contrário, a literatura recente demonstra que os mecanismos de falha identificados naquele estudo continuam presentes em diferentes setores e aplicações de Inteligência Artificial. Os resultados desta revisão evidenciam a recorrência desse padrão em áreas como saúde (E17, E74) e justiça criminal (E5, E24), sugerindo que o racismo algorítmico constitui um fenômeno sistêmico e persistente.

Apesar do crescimento recente das pesquisas sobre racismo algorítmico, a literatura ainda apresenta algumas limitações relevantes. Observa-se predominância de estudos conduzidos em contextos norte-americanos e europeus, bem como forte concentração de publicações em língua inglesa, o que pode limitar a compreensão de manifestações do fenômeno em contextos sociais e culturais distintos. Além disso, muitos estudos concentram-se na identificação dos vieses, enquanto ainda são limitadas as pesquisas empíricas que avaliam longitudinalmente a efetividade de estratégias de mitigação em contextos reais de aplicação.

5. Considerações Finais

Este estudo evidenciou que o racismo algorítmico constitui um fenômeno recorrente e sistêmico nos sistemas de Inteligência Artificial, manifestando-se em diferentes contextos e áreas de aplicação. A análise realizada demonstra que esses vieses não se limitam a falhas técnicas isoladas, mas refletem desigualdades históricas e estruturais presentes na sociedade, especialmente quando incorporadas em bases de dados e processos institucionais. Diferentemente de revisões anteriores, esta pesquisa integra evidências

recentes e aborda o racismo algorítmico de forma multidimensional, considerando aspectos técnicos, sociais, éticos e regulatórios. Ademais, os resultados evidenciam impactos sociais relevantes, como discriminação, exclusão e o reforço de desigualdades raciais.

Embora existam avanços em estratégias de mitigação, como técnicas de balanceamento de dados, testes de justiça algorítmica e maior transparência nos modelos, ainda há limitações significativas na implementação prática dessas soluções. Dessa forma, torna-se essencial a adoção de abordagens integradas que articulem dimensões técnicas, éticas e regulatórias, além da promoção da diversidade nas equipes de desenvolvimento e nos conjuntos de dados utilizados.

Como limitação deste trabalho, destaca-se que a revisão foi conduzida em duas bases de dados científicas e contemplou apenas estudos publicados em português, inglês e espanhol, o que pode ter restringido o acesso a pesquisas desenvolvidas em outros contextos linguísticos e regionais.

Por fim, este trabalho contribui para o avanço das discussões sobre racismo algorítmico ao sistematizar evidências recentes da literatura, reforçando a necessidade de pesquisas futuras que aprofundem mecanismos de mitigação e proponham políticas mais efetivas para garantir o desenvolvimento de sistemas de IA mais justos e equitativos.

Referências

ABHARI, J.; ASHOK, A. (2023). Mitigating Racial Biases for Machine Learning Based Skin Cancer Detection. In: *International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc '23),* 24. New York: ACM, p. 556–561. DOI: 10.1145/3565287.3617639.

ALLAREDDY, V. et al. (2023). Call for algorithmic fairness to mitigate amplification of racial biases in artificial intelligence models used in orthodontics and craniofacial health. *Orthodontics & Craniofacial Research*, v. 26, s. 1, p. 124–130. DOI: 10.1111/ocr.12721.

BAEZA-YATES, R. (2022). Ethical Challenges in AI. In: ACM International Conference on Web Search and Data Mining, 15. New York: ACM, p. 1–2. DOI: 10.1145/3488560.3498370.

BARDIN, L. (2011) *Análise de conteúdo*. Tradução de Luís Antero Reto e Augusto Pinheiro. São Paulo: Edições 70. Tradução de: *L'Analyse de Contenu*, 1997. Disponível em: https://ia802902.us.archive.org/8/items/bardin-laurence-analise-de-conteudo/bardin-laurence-analise-de-conteudo.pdf. Acesso em: set. 2025.

BLACK, E.; YEOM, S.; FREDRIKSON, M. (2020). FlipTest: fairness testing via optimal transport. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. New York: ACM, p. 111–121. DOI: 10.1145/3351095.3372845.

BROWNE, S. (2010). "Digital Epidermalization: Race, Identity and Biometrics". *Critical Sociology*, v. 36, n. 1, p. 131–150.

BUOLAMWINI, J.; GEBRU, T. (2018). "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Conference on Fairness, Accountability and Transparency*. p. 77–91.

COE, J.; ATAY, M. (2021). Evaluating Impact of Race in Facial Recognition across Machine Learning and Deep Learning Algorithms. *Computers*, v. 10, n. 9, p. 113. DOI: 10.3390/computers10090113.

CHEN, Z. et al. (2024). Fairness Testing: A Comprehensive Survey and Analysis of Trends. *ACM Transactions on Software Engineering and Methodology*, v. 33, n. 5, p. 1–59. DOI: 10.1145/3652155.

GALHOTRA, S.; BRUN, Y.; MELIOU, A. (2017). Fairness testing: testing software for discrimination. In: *Joint Meeting on Foundations of Software Engineering*, 11. New York: ACM, p. 498–510. DOI: 10.1145/3106237.3106277.

GRABOWICZ, P. A.; PERELLO, N.; MISHRA, A. (2022). Marrying Fairness and Explainability in Supervised Learning. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. New York: ACM, p. 1905–1916. DOI: 10.1145/3531146.3533236.

HAIMSON, O. L. et al. (2025). AI Attitudes Among Marginalized Populations in the U.S.: Nonbinary, Transgender, and Disabled Individuals Report More Negative AI Attitudes. In: *ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. New York: ACM, p. 1224–1237. DOI: 10.1145/3715275.3732081.

HAQUE, M. R. et al. (2024). Are We Asking the Right Questions?: Designing for Community Stakeholders' Interactions with AI in Policing. In: *CHI Conference on Human Factors in Computing Systems (CHI '24)*. New York: ACM, p. 1–20. Art. 301. DOI: 10.1145/3613904.3642738.

HALE, J.; KIM, P. H.; GRATCH, J. (2025). Algoritmos "provavelmente justos" podem perpetuar o viés racial e de gênero: um estudo sobre a resolução de disputas salariais. *Autonomous Agents and Multi-Agent Systems*, v. 39, n. 20.

HAMID, T. et al. (2025). DermaGlow: Objective Quantification of Melanin, Erythema and Skin-tone Using Wearable Optical Spectroscopy. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, v. 9, n. 2, p. 1–28. DOI: 10.1145/3729474.

HAN, J. X. et al. (2024). A Causal Framework To Evaluate Racial Bias in Law Enforcement Systems. In: *AAAI/ACM Conference on AI, Ethics, and Society*, 7. [S.l.: s.n.].

HARRINGTON, C. N. et al. (2022). "It's Kind of Like Code-Switching": Black Older Adults' Experiences with a Voice Assistant for Health Information Seeking. In: *CHI Conference on Human Factors in Computing Systems (CHI '22)*. New York: ACM, p. 1–15. Art. 604. DOI: 10.1145/3491102.3501995.

IMANA, B.; KOROLOVA, A.; HEIDEMANN, J. (2025). Auditing for Bias in Ad Delivery Using Inferred Demographic Attributes. In: *ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. New York: ACM, p. 2640–2656. DOI: 10.1145/3715275.3732172.

INTAHCHOMPHOO, C.; GUNDERSEN, O. E. (2020). "Artificial Intelligence and Race: a Systematic Review". *Legal Information Management*, v. 20, n. 2, p. 74–84. DOI: 10.1017/S1472669620000183.

JIN, Z. et al. (2020). MithraCoverage: A System for Investigating Population Bias for Intersectional Fairness. In: *ACM SIGMOD International Conference on Management of Data (SIGMOD '20).* New York: ACM, p. 2721–2724. DOI: 10.1145/3318464.3384689.

JOHNSON, D. K. N. (2025). Gaslighting Ourselves: Racial Challenges of Artificial Intelligence in Economics and Finance Applications. The Review of Black Political Economy, v. 52, n. 3, p. 303–335. DOI: 10.1177/00346446251331615.

KHADEMI, A. et al. (2019). Fairness in Algorithmic Decision Making: An Excursion Through the Lens of Causality. In: *The World Wide Web Conference (WWW '19).* New York: ACM, p. 2907–2914. DOI: 10.1145/3308558.3313559.

KITCHENHAM, B. A.; CHARTERS, S. (2007) Guidelines for performing systematic literature reviews in software engineering. Tech. Rep. EBSE-2007-01, Keele University.

KOSTICK-QUENET, K. M.; COHEN, I. G.; GERKE, S. et al. Mitigating Racial Bias in Machine Learning. *Journal of Law, Medicine & Ethics*, v. 50, n. 1, p. 92–100, 2022. DOI: 10.1017/jme.2022.13.

LI, J.; MOSKOVITCH, Y.; JAGADISH, H. V. (2021). DENOUNCER: detection of unfairness in classifiers. *Proceedings of the VLDB Endowment*, v. 14, n. 12, p. 2719–2722. DOI: 10.14778/3476311.3476328.

MARKS, P. (2021). Can the biases in facial recognition be fixed; also, should they? Communications of the ACM, v. 64, n. 3, p. 20–22. DOI: 10.1145/3446877.

METAXA, D. et al. (2021). An Image of Society: Gender and Racial Representation and Impact in Image Search Results for Occupations.

MUNANGA, Kabengele. Rediscutindo a mestiçagem no Brasil: identidade nacional versus identidade negra. Petrópolis, RJ: Vozes, 1999.

NOBLE, S. U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. New York: NYU Press.

NYLAND, J. J. A. O. L. (2023). "Racismo algorítmico: uma revisão de literature". *Research, Society and Development*, v. 12, n. 2, e1912239907. DOI: http://dx.doi.org/10.33448/rsd-v12i2.39907.

OBERMEYER, Z. et al. (2019). "Dissecting racial bias in an algorithm used to manage the health of populations". *Science*, v. 366, n. 6464, p. 447–453.

O'NEIL, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown.

ORPHANOU, K.; OTTERBACHER, J.; KLEANTHOUS, S.; BATSUREN, K.; GIUNCHIGLIA, F.; BOGINA, V.; SHULNER TAL, A.; HARTMAN, A.; KUFLIK, T. (2023). Mitigating Bias in Algorithmic Systems—A Fish-eye View. *ACM Computing Surveys*, v. 55, n. 5, Art. 87, p. 1–37. DOI: 10.1145/3527152.

PÄÄKKÖNEN, J. et al. (2020). Bureaucracy as a Lens for Analyzing and Designing Algorithmic Systems. In: *CHI Conference on Human Factors in Computing Systems (CHI '20)*. New York: ACM, p. 1–14. DOI: 10.1145/3313831.3376780.

PERES, I. E. V. et al. (2021). “Preconceito em algoritmos de aprendizagem de máquina e suas bases de treinamento: uma revisão sistemática de literatura”. Universidade Presbiteriana Mackenzie, São Paulo.

PUYOL-ANTÓN, E. et al. (2022). Fairness in Cardiac Magnetic Resonance Imaging: Assessing Sex and Racial Bias in Deep Learning-Based Segmentation. Frontiers in Cardiovascular Medicine, v. 9. DOI: 10.3389/fcvm.2022.859310.

SAPIEZYNSKI, P. et al. (2022). Algorithms that “Don’t See Color”: Measuring Biases in Lookalike and Special Ad Audiences. In: *AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*. New York: ACM, p. 609–616. DOI: 10.1145/3514094.3534135.

SHAHID, S. B. et al. (2025). Bias in Deep Learning Skin Cancer Detection: Parallel Residual Convolution Network Classification and Racial Bias Quantification: Skin Cancer Racial Bias. In: *International Conference on Computing Advancements*, 3. New York: ACM, p. 993–1000. DOI: 10.1145/3723178.3723310.

SILVA, T. (2022). Racismo algorítmico: inteligência artificial e discriminação nas redes digitais. São Paulo: Editora Polis.

SMITH, J. M. (2024). “I'm Sorry, but I Can't Assist”: Bias in Generative AI. In: *RESPECT Annual Conference (RESPECT 2024)*. New York: ACM, p. 75–80. DOI: 10.1145/3653666.3656065.

SRINIVASAN, R.; CHANDER, A. (2021). Biases in AI Systems: A Survey for Practitioners. Queue, v. 19, n. 2, p. 1–20. DOI: 10.1145/3466132.3466134.

TANKSLEY, T. et al. (2025). “Ethics is not neutral”: Understanding Ethical and Responsible AI Design from the Lenses of Black Youth. In: *CHI Conference on Human Factors in Computing Systems (CHI '25)*. New York: ACM, p. 1–20. Art. 200. DOI: 10.1145/3706598.3713510.

WÓJCIK, M. A. (2023). Assessing the Legality of Using the Category of Race and Ethnicity in Clinical Algorithms: the EU Anti-Discrimination Law Perspective. In: *EWAF*.

YANG, K. et al. (2018). A Nutritional Label for Rankings. In: *International Conference on Management of Data (SIGMOD '18)*. New York: ACM, p. 1773–1776. DOI: 10.1145/3183713.3193568.

YUCER, S. et al. (2025). Racial Bias within Face Recognition: A Survey. ACM Computing Surveys, v. 57, n. 4, p. 1–39. Art. 105. DOI: 10.1145/3705295.

ZHANG, L.; ZHANG, Y.; ZHANG, M. (2021). Efficient white-box fairness testing through gradient search. In: *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2021)*. New York: ACM, p. 103–114. DOI: 10.1145/3460319.3464820.

ZHAO, D. et al. (2021). Understanding and Evaluating Racial Biases in Image Captioning. In: *IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal. [S.l.]: IEEE, p. 14810–14820. DOI: 10.1109/ICCV48922.2021.01456.