ISSN: 2595-5942

Mitigação de Vieses em Sistemas de Inteligência Artificial: Estratégias, Desafios e Implicações Éticas no Contexto Contemporâneo

Higor Cavallini Gallina
Faculdade de Tecnologia Dr. Thomaz Novelino
higor.gallina@aluno.cps.sp.gov.br

Tonis Roberto de Carvalho Junior
Faculdade de Tecnologia Dr. Thomaz Novelino
tonis.roberto@aluno.cps.sp.gov.br

Alexandre Gomes da Silva
Mestre em Computação Aplicada
Docente pela Faculdade de Tecnologia Dr. Thomaz Novelino
alexandre.silva09@cps.sp.gov.br

Resumo
O avanço da Inteligência Artificial (IA) tem ampliado sua aplicação em processos decisórios em diversos setores, ao mesmo tempo em que evidencia desafios relacionados à presença de vieses algorítmicos. Esses vieses podem comprometer a equidade e a justiça das decisões automatizadas, reproduzindo desigualdades sociais existentes. Diante desse cenário, o presente estudo tem como objetivo analisar as principais estratégias de mitigação de vieses em sistemas de IA considerando seus impactos sociais e limitações técnicas. A pesquisa caracteriza-se como qualitativa, de natureza exploratória e descritiva, sendo desenvolvida por meio de revisão bibliográfica sistematizada em bases acadêmicas nacionais e internacionais. Os resultados indicam que as estratégias de mitigação podem ser classificadas em três níveis: pré-processamento, processamento e pós-processamento, destacando-se a importância da governança de dados, da diversidade nas equipes de desenvolvimento e da adoção de técnicas de explicabilidade. Conclui-se que a mitigação de vieses exige uma abordagem multidimensional, integrando aspectos técnicos, éticos e regulatórios, sendo fundamental para garantir sistemas de IA mais justos, transparentes e confiáveis.

Palavras-chave: algoritmos; ética; inteligência artificial; mitigação de vieses; tomada de decisão.

Abstract
The growing use of Artificial Intelligence (AI) in decision-making processes across multiple sectors has increased efficiency and automation but has also raised concerns regarding algorithmic bias. Such biases may compromise fairness and reinforce existing social inequalities. This study aims to examine the main strategies for mitigating bias in AI systems, considering their social impacts and technical limitations. The research adopts a qualitative, exploratory, and descriptive approach, based on a systematic literature review of national and international academic sources. The results indicate that mitigation strategies operate at three levels: pre-processing, in-processing, and post-processing, highlighting the relevance of data governance, diversity in development teams, and explainability techniques. The findings suggest that effective bias mitigation requires a multidimensional approach that integrates

ISSN: 2595-5942

technical, ethical, and regulatory aspects, contributing to the development of more reliable and equitable AI systems.

Keywords: *algorithms; artificial intelligence; bias mitigation; decision-making; ethics.*

1 INTRODUÇÃO

A crescente integração da Inteligência Artificial (IA) nas estruturas da sociedade contemporânea tem reconfigurado substancialmente as arquiteturas decisórias, com especial incidência em domínios críticos como a segurança pública, o sistema financeiro e a gestão de recursos humanos. Não obstante os benefícios intrínsecos a essas tecnologias, a célere expansão da IA em contextos deliberativos de alta sensibilidade fomenta reflexões rigorosas acerca das implicações sociais de processos automatizados, notadamente no que tange à potencial cristalização de disparidades por intermédio de distorções algorítmicas.

Embora o potencial transformador da área se manifeste em setores estratégicos como a medicina e o desenvolvimento sustentável, a implementação de sistemas autônomos desprovidos de mecanismos de controle pode exacerbar a vulnerabilidade de grupos sub-representados. Nesse panorama, o enfrentamento desses vieses técnicos constitui-se como um desafio premente e imprescindível, revelando-se pilar indispensável para assegurar a equidade, a transparência e a idoneidade ética das soluções inteligentes. Ao serem fundamentadas em bases informacionais enviesadas, tais estruturas computacionais tendem a espelhar ou intensificar assimetrias históricas, operando, frequentemente, sob a opacidade característica das chamadas “caixas-pretas”.

As distorções em sistemas de IA emergem de forma multidimensional, manifestando-se desde a etapa de mineração de dados até as interfaces de interação com o usuário, abrangendo tipologias como os vieses comportamentais, de agregação e de interação. A inexistência de medidas mitigatórias consistentes favorece a continuidade e o reforço de mecanismos discriminatórios estruturados.

Sob essa ótica, a presente pesquisa orienta-se pela seguinte problemática: quais táticas revelam-se mais promissoras para a contenção de vieses em arquiteturas de Inteligência Artificial, sopesando-se seus reflexos socioculturais e as inerentes restrições de ordem técnica?

ISSN: 2595-5942

O objetivo geral deste trabalho é analisar estratégias de mitigação de vieses em sistemas de IA. Como objetivos específicos, pretende-se:

- Identificar as principais fontes de vieses algorítmicos;
- Analisar criticamente seus impactos sociais;
- Examinar técnicas de mitigação aplicáveis ao desenvolvimento de sistemas de IA.

A relevância da pesquisa justifica-se pela expansão do uso da IA em contextos decisórios, exigindo o desenvolvimento de soluções alinhadas a princípios éticos, como equidade, transparência e responsabilidade.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta os principais conceitos e discussões teóricas relacionados à Inteligência Artificial, aos vieses algorítmicos e às estratégias utilizadas para mitigação desses problemas em sistemas inteligentes.

2.1 Conceito e evolução da Inteligência Artificial

A Inteligência Artificial (IA) consolidou-se como um campo multidisciplinar voltado ao desenvolvimento de sistemas capazes de executar tarefas que, tradicionalmente, exigiriam capacidades cognitivas humanas, como aprender, inferir e tomar decisões. Conforme discutido por Russell e Norvig (2020), a área se estrutura a partir de duas orientações principais: uma centrada na simulação do comportamento humano e outra orientada pela busca de decisões racionais e ótimas. Essa distinção não é meramente conceitual, pois implica diferentes formas de modelagem e avaliação dos sistemas, o que influencia diretamente seus resultados e limitações.

As origens da IA remontam a debates filosóficos sobre a natureza da mente e da inteligência, mas seu desenvolvimento técnico ganha forma a partir das proposições de Alan Turing, especialmente com o chamado “Teste de Turing”. Ainda que esse marco tenha sido fundamental para o avanço da área, parte da literatura contemporânea questiona sua suficiência como critério de inteligência, uma vez que privilegia a imitação do comportamento humano em detrimento da compreensão efetiva dos processos cognitivos subjacentes.

Do ponto de vista histórico, a evolução da IA é marcada por mudanças significativas de paradigma. Os primeiros sistemas, baseados em regras lógicas e estruturas determinísticas, apresentavam limitações diante de problemas complexos

e dinâmicos. A transição para abordagens orientadas por dados, impulsionada pelo avanço do *Machine Learning*, ampliou consideravelmente a capacidade dos sistemas em lidar com grandes volumes de informação e identificar padrões de forma autônoma. No entanto, essa mudança também introduziu novas vulnerabilidades, especialmente relacionadas à qualidade e à representatividade dos dados utilizados.

Com o desenvolvimento do *Deep Learning*, a área alcançou níveis elevados de desempenho em tarefas como reconhecimento de imagens, processamento de linguagem natural e sistemas autônomos. Mais recentemente, a ascensão dos modelos de linguagem de grande escala e da IA generativa tem ampliado o alcance dessas tecnologias, permitindo a produção automatizada de conteúdos complexos. Apesar desses avanços, a literatura aponta que tais modelos tendem a operar como "caixas-pretas", o que dificulta a interpretação de seus resultados e levanta questionamentos quanto à transparência e à responsabilidade de suas aplicações.

Nesse contexto, o debate sobre IA ultrapassa o domínio técnico e incorpora dimensões éticas e sociais. Noble (2018) argumenta que algoritmos não são neutros, uma vez que refletem as estruturas e desigualdades presentes nos dados que os alimentam. De forma convergente, O'Neil (2016) destaca que modelos algorítmicos podem reforçar mecanismos de exclusão ao serem aplicados em contextos sensíveis, como crédito, segurança pública e recrutamento. Essa perspectiva evidencia que a aparente objetividade dos sistemas pode ocultar processos de discriminação estruturada.

A literatura também enfatiza que os vieses não se restringem a uma única etapa do desenvolvimento dos sistemas. Buolamwini e Gebru (2018) demonstram, por meio de evidências empíricas, que falhas em bases de dados podem gerar desigualdades significativas nos resultados, sobretudo em relação a grupos sub-representados. Complementarmente, Barocas, Hardt e Narayanan (2019) argumentam que o viés deve ser compreendido como um fenômeno que atravessa todo o ciclo de vida dos sistemas, desde a coleta de dados até a aplicação dos resultados, exigindo abordagens mais amplas de análise e intervenção.

Apesar dos desafios identificados, o potencial transformador da IA permanece significativo. Aplicações recentes, como as desenvolvidas na biologia molecular, evidenciam sua capacidade de acelerar descobertas científicas e ampliar fronteiras do conhecimento. No entanto, o avanço dessas tecnologias impõe a necessidade de uma abordagem crítica e responsável, na qual o desenvolvimento técnico esteja

ISSN: 2595-5942

articulado a princípios éticos e a mecanismos de regulação. Assim, mais do que aprimorar a eficiência dos sistemas, torna-se fundamental assegurar que sua utilização contribua para a redução, e não a ampliação, de desigualdades sociais.

2.2 Estratégias de mitigação de vieses

Previamente ao exame das táticas mitigatórias, revela-se imperativo delimitar os conceitos de desigualdade e equidade sob o prisma da Inteligência Artificial. A **desigualdade** em sistemas de IA compreende a manifestação de decisões automatizadas que produzem desfechos ou tratamentos sistematicamente desvantajosos para determinados indivíduos ou grupos. Tais fenômenos comumente espelham discriminações históricas cristalizadas nos repositórios informacionais que fundamentam a aprendizagem das estruturas computacionais.

Em contrapartida, a **equidade algorítmica** orienta-se para a concepção de modelos aptos a gerar conclusões justas e íntegras, minimizando segregações arbitrárias vinculadas a marcadores identitários, tais como etnia, gênero, faixa etária ou extrato social. Sob essa ótica, a relação entre esses conceitos evidencia que o enfrentamento dos vieses técnicos abordados ao longo deste estudo consolida-se como pilar indispensável para combater a desigualdade e assegurar a idoneidade ética nas arquiteturas decisórias assistidas por IA.

A mitigação de vieses em sistemas de Inteligência Artificial exige intervenções distribuídas ao longo de todo o ciclo de desenvolvimento, sendo comumente organizada em três níveis: pré-processamento, processamento e pós-processamento. Essa classificação, embora útil do ponto de vista didático, não deve ser compreendida como solução suficiente, uma vez que os vieses possuem natureza complexa e podem emergir tanto da estrutura dos dados quanto das decisões incorporadas ao longo do desenvolvimento dos sistemas (O'Neil, 2016). Nesse sentido, a literatura indica que abordagens isoladas tendem a apresentar resultados limitados quando não articuladas a uma perspectiva mais ampla.

Grande parte dos estudos converge ao reconhecer que os conjuntos de dados utilizados no treinamento frequentemente refletem desigualdades históricas de ordem social, econômica e cultural. Isso implica que os sistemas aprendem padrões que não apenas reproduzem essas distorções, mas, em determinadas circunstâncias, podem intensificá-las. O caso de sistemas aplicados ao recrutamento, discutido por Binns

ISSN: 2595-5942

(2018), ilustra como modelos treinados com dados históricos tendem a favorecer perfis já predominantes, perpetuando assimetrias de gênero e de representação.

Os vieses em sistemas de Inteligência Artificial podem emergir em diferentes etapas do ciclo de desenvolvimento, desde a coleta de dados até os impactos gerados pelas decisões automatizadas. Esse processo cíclico está representado na Figura 1.

Figura 1 – Ciclo de vieses em sistemas de Inteligência Artificial

Fonte: Elaborado pelos autores

Conforme ilustrado na Figura 1, os vieses em sistemas de Inteligência Artificial não se configuram como eventos isolados, mas como fenômenos que se retroalimentam ao longo do ciclo de desenvolvimento, reforçando desigualdades e demandando intervenções contínuas. Nesse sentido, limitar a análise à dimensão dos dados revela-se insuficiente. De acordo com Mehrabi et al. (2021), os vieses também podem ser introduzidos pelas próprias escolhas dos desenvolvedores, sobretudo quando aspectos contextuais e culturais não são devidamente considerados.

Essa perspectiva amplia o debate ao evidenciar que a mitigação de vieses não se restringe a soluções técnicas, mas envolve, igualmente, reflexões sobre os processos de desenvolvimento e os agentes envolvidos. Nesse contexto, a diversidade nas equipes deixa de ser apenas uma recomendação institucional e passa a assumir caráter estratégico, contribuindo para a ampliação da capacidade de identificação de riscos e limitações dos sistemas.

ISSN: 2595-5942

As estratégias de mitigação, quando analisadas criticamente, revelam potencialidades e limites. Técnicas de pré-processamento, como balanceamento e reamostragem de dados, contribuem para reduzir distorções iniciais, mas não eliminam vieses que emergem durante o treinamento. Abordagens de *in-processing*, por sua vez, incorporam critérios de equidade diretamente nos algoritmos, porém podem implicar *trade-offs* entre desempenho e justiça. Já o pós-processamento atua sobre os resultados, corrigindo padrões identificados, mas sem necessariamente modificar as causas estruturais do problema. Essa tensão entre precisão e equidade é recorrente na literatura e evidencia a inexistência de soluções universais.

Diante dessas limitações, autores como Raji *et al*. (2020) defendem a adoção de práticas complementares, como governança de dados, auditorias contínuas e mecanismos de responsabilização. Tais instrumentos ampliam o escopo da mitigação, permitindo o monitoramento dos sistemas em operação e a correção de distorções ao longo do tempo. De forma convergente, Jobin *et al*. (2019) destacam a importância de diretrizes internacionais que orientem o desenvolvimento de sistemas mais transparentes e alinhados a princípios éticos, embora reconheçam a heterogeneidade dessas iniciativas e os desafios de sua aplicação prática.

Outro elemento central nesse debate refere-se à explicabilidade dos modelos. Conforme apontam Doshi-Velez e Kim (2017), a capacidade de interpretar decisões algorítmicas é fundamental para identificar vieses e promover ajustes consistentes. No entanto, essa proposta enfrenta limitações quando aplicada a modelos complexos, cuja opacidade dificulta a rastreabilidade dos processos internos. Assim, a busca por transparência deve ser compreendida como um desafio técnico e, ao mesmo tempo, como uma exigência ética.

Além das dimensões técnicas, a literatura enfatiza a necessidade de formação crítica dos profissionais envolvidos no desenvolvimento de sistemas de IA. Crawford *et al*. (2019) argumentam que a incorporação de princípios éticos na formação desses agentes contribui para decisões mais conscientes e responsáveis, ainda que não elimine, por si só, os riscos associados ao uso da tecnologia.

Por fim, destaca-se o papel das políticas públicas e dos marcos regulatórios na mitigação de vieses em sistemas de Inteligência Artificial. Iniciativas como a Estratégia de Inteligência Artificial da União Europeia evidenciam a busca por diretrizes capazes de conciliar a inovação tecnológica com a proteção de direitos fundamentais. Contudo, a efetividade dessas regulações depende da articulação entre diferentes atores,

ISSN: 2595-5942

governos, setor privado e comunidade acadêmica, bem como da capacidade de adaptação a contextos específicos.

Nesse cenário, conforme sintetizado no Quadro 1, as estratégias de mitigação podem ser classificadas em diferentes níveis ao longo do ciclo de desenvolvimento dos sistemas, reforçando a necessidade de uma abordagem integrada e contínua.

Quadro 1 – Síntese das estratégias de mitigação de vieses em IA

Tipo	Descrição	Vantagens	Limitações	Autor
Pré-processamento	Ajuste e balanceamento dos dados	Reduz vieses iniciais	Não corrige vieses do modelo	Mehrabi *et al*. (2021)
In-processing	Inclusão de equidade no algoritmo	Atua diretamente no modelo	Pode reduzir desempenho	Binns (2018)
Pós-processamento	Correção dos resultados após o treinamento	Fácil aplicação	Não resolve causa estrutural	Raji *et al*. (2020)

Fonte: Elaborado pelos autores.

Por conseguinte, a mitigação de vieses em Inteligência Artificial não pode ser reduzida a um conjunto de técnicas aplicadas de forma pontual. Trata-se de um processo contínuo e multidimensional, que exige a integração entre soluções técnicas, práticas organizacionais e mecanismos regulatórios, evidenciando o caráter essencialmente sociotécnico do problema.

2.3 Ética no desenvolvimento de sistemas de IA

A incorporação crescente da Inteligência Artificial (IA) em processos decisórios amplia a necessidade de uma reflexão ética consistente, que vá além de princípios abstratos e se traduza em práticas efetivas de desenvolvimento e uso dessas tecnologias. A literatura aponta que a ética em IA não se limita à formulação de diretrizes normativas, mas envolve a análise crítica das condições concretas em que os sistemas são concebidos, treinados e aplicados. Nesse sentido, mais do que um complemento ao desenvolvimento técnico, a dimensão ética constitui um elemento estruturante, capaz de influenciar diretamente os resultados produzidos pelos sistemas.

2.3.1 Transparência

A transparência é frequentemente apontada como um dos pilares da ética em IA, sobretudo por sua relação com a confiabilidade e a possibilidade de controle social sobre sistemas automatizados. Coeckelbergh (2020) argumenta que tornar os

ISSN: 2595-5942

processos algorítmicos compreensíveis não é apenas uma exigência técnica, mas uma condição para a atribuição de responsabilidade e para a legitimação do uso dessas tecnologias. No entanto, essa noção enfrenta limites práticos, especialmente diante de modelos complexos, cuja estrutura dificulta a interpretação de seus mecanismos internos.

As diretrizes da Comissão Europeia (2019) reforçam a necessidade de transparência em diferentes níveis, incluindo o funcionamento dos algoritmos e os impactos de suas decisões. Ainda assim, a literatura aponta que a transparência, por si só, não garante justiça ou ausência de vieses, podendo inclusive gerar uma falsa sensação de controle quando não acompanhada de mecanismos efetivos de auditoria e avaliação crítica. Dessa forma, a transparência deve ser compreendida como condição necessária, mas não suficiente, para o desenvolvimento ético da IA.

2.3.2 Equidade

A equidade ocupa posição central no debate ético sobre Inteligência Artificial, especialmente diante da capacidade dos sistemas de reproduzir e, em alguns casos, intensificar desigualdades preexistentes. Russell e Norvig (2020) já indicavam que algoritmos, ao aprenderem com dados históricos, tendem a incorporar padrões sociais que refletem assimetrias estruturais. Essa constatação é aprofundada por O'Neil (2016), ao demonstrar como modelos matemáticos podem operar como instrumentos de ampliação da desigualdade, sobretudo quando aplicados em contextos sensíveis.

Contudo, a busca por equidade não é isenta de tensões. A literatura evidencia que diferentes definições de justiça podem levar a soluções conflitantes, o que torna o processo de mitigação de vieses uma tarefa complexa e contextual. Assim, mais do que alcançar uma suposta neutralidade, o desafio consiste em explicitar critérios, avaliar impactos e assumir decisões que, inevitavelmente, envolvem escolhas normativas.

2.3.3 Responsabilidade e *accountability*

A questão da responsabilidade em sistemas de IA revela um dos principais desafios éticos contemporâneos, especialmente diante do aumento da autonomia dessas tecnologias. Bostrom e Yudkowsky (2014) discutem as dificuldades de atribuir responsabilidade em contextos nos quais as decisões são mediadas por sistemas

ISSN: 2595-5942

complexos, envolvendo múltiplos agentes — desenvolvedores, organizações e usuários.

Nesse cenário, o conceito de *accountability* ganha relevância ao enfatizar a necessidade de mecanismos que permitam identificar, justificar e, quando necessário, corrigir decisões automatizadas. No entanto, a literatura aponta que a responsabilização não pode ser tratada apenas no nível individual, devendo considerar também estruturas institucionais e organizacionais que influenciam o desenvolvimento e a aplicação da IA.

Dessa forma, garantir responsabilidade em sistemas de Inteligência Artificial implica não apenas definir quem responde por eventuais falhas, mas também estabelecer processos de supervisão, auditoria e governança capazes de prevenir danos e assegurar o alinhamento dessas tecnologias a princípios éticos mais amplos.

2.4 Impactos da Inteligência Artificial na sociedade

A expansão da Inteligência Artificial (IA) nas últimas décadas tem produzido transformações profundas em diferentes dimensões da vida social, econômica e institucional. Conforme evidenciado por Antevere Filho e Conceição (2023), a IA vem sendo incorporada de forma crescente em setores estratégicos, promovendo ganhos expressivos de eficiência, precisão e capacidade analítica. Esse avanço, no entanto, não ocorre de maneira neutra. Seus efeitos envolvem simultaneamente benefícios e riscos, o que exige uma análise crítica de suas implicações.

Entre os impactos positivos mais recorrentes, destaca-se a contribuição da IA para o aumento da produtividade e para a otimização de processos. Em áreas como saúde, por exemplo, sistemas baseados em aprendizado de máquina têm possibilitado diagnósticos mais rápidos e precisos, ampliando as chances de tratamento eficaz e reduzindo custos operacionais. De modo semelhante, na indústria e na logística, a automação inteligente tem permitido maior controle sobre cadeias produtivas, reduzindo falhas humanas e melhorando o desempenho organizacional (Antevere Filho; Conceição, 2023). Essas aplicações evidenciam o potencial da IA como ferramenta de apoio à tomada de decisão, especialmente em contextos que envolvem grande volume de dados.

No campo das interações sociais e do consumo digital, a IA também desempenha papel relevante ao personalizar experiências e facilitar o acesso à informação. Sistemas de recomendação e assistentes virtuais tornam a navegação

ISSN: 2595-5942

mais eficiente, adaptando conteúdos às preferências dos usuários. Contudo, como apontado por Pariser (2012), essa personalização pode gerar efeitos indesejados, como a formação de “bolhas de filtro”, nas quais os indivíduos passam a ter contato restrito a informações que reforçam suas próprias visões, limitando o debate público e potencializando processos de polarização.

Outro aspecto crítico refere-se aos impactos no mercado de trabalho. A automação de tarefas, especialmente aquelas de natureza repetitiva, tende a substituir determinadas funções, ao mesmo tempo em que cria demandas por habilidades técnicas e cognitivas mais complexas. Esse movimento, embora associado a ganhos de eficiência, pode aprofundar desigualdades sociais, sobretudo em contextos em que não há políticas adequadas de requalificação profissional. Nesse sentido, a literatura alerta para a necessidade de adaptação das estruturas educacionais e institucionais frente às novas exigências impostas pela economia digital (Antevere Filho; Conceição, 2023).

Além das transformações econômicas, a IA também levanta preocupações relevantes no campo ético e jurídico, especialmente no que diz respeito à privacidade e ao uso de dados. A capacidade de coletar, processar e analisar grandes volumes de informações pessoais amplia os riscos de vigilância e uso indevido desses dados. Custers (2013) destaca que a análise massiva de dados pode resultar em práticas discriminatórias e em processos de tomada de decisão pouco transparentes, afetando diretamente direitos individuais.

No âmbito da segurança e do controle social, tecnologias como o reconhecimento facial exemplificam esse dilema. Embora contribuam para a prevenção de crimes e para a gestão de espaços públicos, também levantam questionamentos sobre vigilância em larga escala e consentimento no uso de dados biométricos. Estudos como o de Solon (2019) apontam que a utilização de imagens sem autorização para treinamento de algoritmos representa uma violação significativa da privacidade.

Os impactos desses vieses podem ser observados em diferentes aplicações práticas. Buolamwini e Gebru (2018) demonstraram que sistemas de reconhecimento facial apresentavam taxas de erro significativamente maiores para mulheres e pessoas negras quando comparadas a homens brancos. De forma semelhante, sistemas automatizados de recrutamento podem favorecer perfis historicamente predominantes, reproduzindo desigualdades de gênero e representação profissional.

ISSN: 2595-5942

Também há registros de algoritmos utilizados para concessão de crédito e avaliação de risco que produziram resultados desproporcionais entre diferentes grupos sociais, evidenciando os efeitos concretos dos vieses algorítmicos em contextos decisórios.

Diante desse cenário, torna-se evidente que os impactos da Inteligência Artificial não podem ser analisados apenas sob a perspectiva de inovação tecnológica. Trata-se de um fenômeno que reconfigura relações sociais, dinâmicas econômicas e estruturas institucionais. Assim, conforme destacam Antevere Filho e Conceição (2023), é fundamental que o avanço da IA seja acompanhado por políticas públicas, marcos regulatórios e iniciativas educacionais que promovam seu uso responsável.

Em síntese, a Inteligência Artificial apresenta um caráter ambivalente: ao mesmo tempo em que amplia capacidades humanas e impulsiona o desenvolvimento, também introduz novos desafios relacionados à equidade, à privacidade e à governança tecnológica. Compreender esses impactos de forma crítica é condição essencial para orientar decisões que assegurem que os benefícios da IA sejam distribuídos de maneira mais justa e sustentável na sociedade contemporânea.

3 MATERIAIS E MÉTODO

O presente estudo caracteriza-se como uma pesquisa qualitativa, de natureza exploratória e descritiva, desenvolvida por meio de revisão bibliográfica sistematizada. A coleta de dados foi realizada em bases acadêmicas como Google Scholar, SciELO e periódicos internacionais indexados, com foco em publicações relevantes nas áreas de Inteligência Artificial, aprendizado de máquina e ética algorítmica.

A triagem das fontes obedeceu a parâmetros de rigor metodológico, privilegiando tanto obras seminais no domínio da Inteligência Artificial e da equidade nos algoritmos quanto produções contemporâneas veiculadas entre os anos de 2010 e 2024. Tal procedimento visou abarcar uma pluralidade de correntes teóricas e implementações práticas concernentes às distorções algorítmicas e seus respectivos mecanismos de contenção.. Foram excluídos trabalhos sem rigor metodológico ou que não abordassem diretamente o problema de pesquisa.

A análise dos dados foi conduzida por meio de síntese temática, permitindo a categorização dos estudos em três eixos principais: (i) tipos de vieses algorítmicos, (ii) impactos sociais e (iii) estratégias de mitigação. Posteriormente, realizou-se uma análise comparativa entre os autores, buscando identificar convergências,

ISSN: 2595-5942

divergências e lacunas na literatura, com o objetivo de responder ao problema de pesquisa proposto.

4 ANÁLISE E RESULTADOS

A análise da literatura evidencia que os vieses em sistemas de Inteligência Artificial não devem ser compreendidos como ocorrências pontuais, mas como fenômenos de natureza estrutural, que atravessam diferentes fases do ciclo de desenvolvimento dos sistemas. Nesse sentido, Barocas, Hardt e Narayanan (2019) argumentam que tais distorções podem emergir desde a etapa inicial de coleta e seleção de dados até os processos finais de interpretação e aplicação dos resultados. De modo complementar, Mehrabi *et al.* (2021) aprofundam essa compreensão ao sistematizar diferentes categorias de vieses, como os de representação, interação e agregação, destacando a complexidade e a multiplicidade de suas origens.

Há consenso na literatura de que a qualidade e a representatividade dos dados constituem fatores centrais na gênese desses vieses. Estudos empíricos, como o de Buolamwini e Gebru (2018), demonstram que conjuntos de dados pouco diversos podem comprometer significativamente o desempenho de sistemas de reconhecimento facial, sobretudo quando aplicados a grupos historicamente sub-representados. No entanto, a discussão não se restringe à dimensão técnica dos dados. Binns (2018) chama atenção para o papel das decisões humanas no desenvolvimento dos algoritmos, evidenciando que escolhas metodológicas e pressupostos implícitos também influenciam os resultados produzidos pelos sistemas.

No que diz respeito às estratégias de mitigação, a literatura aponta três frentes principais de atuação. A primeira refere-se ao pré-processamento, que envolve a preparação e o balanceamento dos dados antes do treinamento, com o objetivo de reduzir distorções iniciais. A segunda abordagem, denominada *in-processing*, consiste na incorporação de critérios de equidade diretamente nos algoritmos durante o processo de aprendizagem. Por fim, o pós-processamento busca ajustar os resultados gerados, corrigindo possíveis assimetrias identificadas após a modelagem.

Apesar dos avanços nessas abordagens, observa-se que sua aplicação isolada tende a apresentar limitações. Raji *et al.* (2020) destacam que soluções estritamente técnicas não são suficientes para eliminar vieses de forma consistente, sendo necessário integrá-las a práticas mais amplas de governança, monitoramento e auditoria contínua dos sistemas. Essa perspectiva reforça a ideia de que o

ISSN: 2595-5942

enfrentamento do problema demanda uma abordagem sistêmica, que vá além do nível operacional dos algoritmos.

Outro aspecto recorrente na literatura refere-se à importância da explicabilidade dos modelos. Conforme argumentam Doshi-Velez e Kim (2017), a transparência nos processos decisórios permite não apenas identificar eventuais distorções, mas também fortalecer a confiabilidade dos sistemas perante seus usuários. Essa discussão é aprofundada por Burrell (2016), ao destacar que a opacidade característica de muitos modelos complexos representa um desafio significativo para a compreensão e o controle de seus resultados.

Adicionalmente, os estudos analisados convergem ao enfatizar a necessidade de abordagens interdisciplinares. Crawford *et al.* (2019) ressaltam a relevância da formação ética dos profissionais envolvidos no desenvolvimento dessas tecnologias, enquanto Jobin *et al*. (2019) evidenciam o papel das diretrizes e regulações internacionais na promoção de práticas mais responsáveis. Tais contribuições indicam que a mitigação de vieses não se limita ao domínio técnico, exigindo também reflexões de natureza social, ética e institucional.

A análise comparativa da literatura permite identificar importantes convergências entre os autores. Mehrabi et al. (2021), Barocas, Hardt e Narayanan (2019) e Raji et al. (2020) concordam que os vieses não possuem uma única origem, podendo surgir em diferentes etapas do ciclo de vida dos sistemas. Da mesma forma, observa-se consenso quanto à necessidade de abordagens multidimensionais para mitigação, combinando intervenções técnicas e mecanismos de governança.

Em relação às estratégias mais eficazes, os estudos indicam que soluções baseadas exclusivamente em pré-processamento apresentam resultados limitados quando utilizadas isoladamente. As abordagens que combinam técnicas de balanceamento de dados, critérios de equidade durante o treinamento e monitoramento contínuo após a implantação tendem a apresentar resultados mais consistentes. Embora existam divergências quanto à melhor forma de implementação dessas estratégias, a literatura converge para a importância da integração entre aspectos técnicos, éticos e regulatórios.

ISSN: 2595-5942

Em decorrência do exposto, a resolução da problemática central evidencia a inexistência de uma tática exclusiva apta a dirimir as assimetrias algorítmicas de modo autônomo. A literatura converge ao apontar que o êxito metodológico reside na articulação sinérgica entre procedimentos de pré-processamento informacional, calibrações no curso do aprendizado de máquina e retificações subsequentes à modelagem, invariavelmente pautadas por diretrizes de governança, auditoria sistemática e preceitos éticos. Tal arranjo multidimensional potencializa a contenção de distorções técnicas e sociais, viabilizando a consolidação de arquiteturas computacionais dotadas de maior equidade e confiabilidade.

CONSIDERAÇÕES FINAIS

O presente estudo teve como objetivo analisar estratégias de mitigação de vieses em sistemas de Inteligência Artificial, considerando seus impactos sociais e limitações técnicas. A partir da revisão bibliográfica realizada, foi possível identificar que os vieses algorítmicos estão presentes em diferentes etapas do desenvolvimento dos sistemas, sendo influenciados tanto por dados quanto por decisões humanas.

Os resultados evidenciaram que, embora existam diversas técnicas de mitigação, sua eficácia depende da aplicação integrada ao longo de todo o ciclo de vida da IA, envolvendo desde o tratamento dos dados até a implementação de mecanismos de governança e auditoria. Nesse contexto, destacou-se a relevância da diversidade nas equipes de desenvolvimento, da transparência dos algoritmos e da adoção de diretrizes éticas.

Entre as principais limitações do estudo, destaca-se a dependência de fontes secundárias, o que reforça a necessidade de pesquisas empíricas futuras que avaliem a aplicação prática das estratégias discutidas. Como trabalhos futuros, sugere-se a realização de estudos de caso e experimentos aplicados que permitam mensurar a efetividade das técnicas de mitigação em diferentes contextos.

Conclui-se que a construção de sistemas de Inteligência Artificial mais justos e confiáveis exige uma abordagem multidisciplinar, capaz de integrar aspectos técnicos, sociais e éticos, contribuindo para o desenvolvimento de tecnologias alinhadas aos princípios de equidade e responsabilidade.

REFERÊNCIAS

ISSN: 2595-5942

ANTEVERE FILHO, Luiz Carlos; CONCEIÇÃO, Gislaine Cristina da. **Impactos da inteligência artificial na sociedade**. Revista Interface Tecnológica, Taquaritinga, SP, v. 20, n. 2, p. 134–145, 2023. DOI: 10.31510/infa.v20i2.1777.

BAROCAS, S.; HARDT, M.; NARAYANAN, A. **Fairness and machine learning**. Cambridge: MIT Press, 2019.

BINNS, R. Fairness in machine learning: lessons from political philosophy. **Proceedings of Machine Learning Research**, v. 81, p. 1–11, 2018.

BOSTROM, Nick; YUDKOWSKY, Eliezer. The ethics of artificial intelligence. In: FRANKISH, Keith; RAMSEY, William M. (org.). **The Cambridge handbook of artificial intelligence.** Cambridge: Cambridge University Press, 2014.

BUOLAMWINI, J.; GEBRU, T. Gender shades: intersectional accuracy disparities in commercial gender classification. **Proceedings of Machine Learning Research**, v. 81, p. 1–15, 2018.

BURRELL, J. How the machine 'thinks': understanding opacity in machine learning algorithms. **Big Data & Society**, v. 3, n. 1, 2016.

COECKELBERGH, Mark. Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. **Science and Engineering Ethics**, v. 26, n. 4, p. 2051-2068, 2020.

COLUMBIA UNIVERSITY. **AI for Social Good and Society Initiative (AI4SGS)**. Columbia School of Social Work, 2024.

COMMISSION EUROPEIA. **Ethics guidelines for trustworthy AI.** Brussels: Commission Europeia, 2019.

COMMISSION EUROPEIA. **White paper on artificial intelligence: a European approach to excellence and trust**. Brussels: Commission Europeia, 2020.

CRAWFORD, K. et al. **AI Now 2019 Report**. New York: AI Now Institute, 2019.

CUSTERS, Bart. Data mining and discrimination. In: CUSTERS, Bart et al. (org.). **Discrimination and privacy in the information society**. Berlin: Springer, 2013.

DOSHI-VELEZ, F.; KIM, B. Towards a rigorous science of interpretable machine learning. **arXiv preprint**, 2017.

JOBIN, A.; IENCA, M.; VAYENA, E. The global landscape of AI ethics guidelines. **Nature Machine Intelligence**, v. 1, p. 389–399, 2019.

MEHRABI, N. et al. A survey on bias and fairness in machine learning. **ACM Computing Survey*s***, v. 54, n. 6, 2021.

NOBLE, Safiya Umoja. **Algorithms of oppression: how search engines reinforce racism**. New York: New York University Press, 2018.

ISSN: 2595-5942

O'NEIL, C. **Weapons of math destruction**. New York: Crown, 2016.

PARISER, Eli. **The filter bubble: what the Internet is hiding from you**. New York: Penguin Press, 2012.

RAJI, I. D. et al. **Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing.** In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT '20). New York: Association for Computing Machinery, 2020. p. 33–44.

RUSSELL, Stuart; NORVIG, Peter. **Inteligência Artificial: uma abordagem moderna.** 4. ed. São Paulo: Pearson Education do Brasil, 2020.

SOLON, Olivia. Facial recognition's "dirty little secret": millions of online photos scraped without consent. **NBC News**, 2019. Disponível em: https://www.cnbc.com/2019/03/12/millions-of-photos-scraped-without-consent-for-facial-recognition.html. Acesso em: 29 abr. 2026.