UM ESTUDO SOBRE OS MÉTODOS DE REDUÇÃO DE VIESES EM SISTEMAS DE INTELIGÊNCIA ARTIFICIAL.

ANTÔNIO KLÉCIO MARQUES DIAS[1]
MARIÂNGELA FERREIRA FUENTES MOLINA[2]

RESUMO

A busca por uma inteligência artificial justa exige que os desenvolvedores considerem as implicações sociais e morais das tecnologias que criam. Nesse sentido, a mitigação de vieses é um passo fundamental para garantir que as decisões tomadas por sistemas automatizados não reproduzam ou amplifiquem as desigualdades existentes, mas, ao contrário, contribuam para um ambiente mais inclusivo e igualitário. Objetivou-se, assim, averiguar os métodos de redução de vieses em sistemas de Inteligência Artificial (IA), buscando compreender as estratégias mais eficazes para mitigar discriminação e promover a equidade nos processos automatizados. Metodologicamente, tratou-se de uma revisão bibliográfica exploratória e comparativa em harmonia com uma pesquisa qualitativa. Como resultado, salientou-se que a redução de vieses em IA é uma tarefa complexa, que envolve tanto ajustes nos dados utilizados para treinar os modelos quanto a aplicação de técnicas específicas durante e após o processo de treinamento. A utilização de abordagens como aprendizado supervisionado com regularização, auditoria de modelos e o emprego de algoritmos justos emergem como práticas promissoras na mitigação de vieses e promoção de equidade. Conclui-se, de modo complementar, que, embora desafios persistam, a contínua pesquisa e a implementação de práticas éticas são essenciais para o desenvolvimento de sistemas de IA mais transparentes, inclusivos e socialmente responsáveis.

Palavras-chave: Ética em IA; Inteligência Artificial; Justiça Algorítmica; Vieses.

ABSTRACT

The pursuit of fair artificial intelligence requires developers to consider the social and moral implications of the technologies they create. In this sense, bias mitigation is a fundamental step to ensure that decisions made by automated systems do not reproduce or amplify existing inequalities, but rather contribute to a more inclusive and egalitarian environment. The aim of this study was to investigate methods for reducing bias in Artificial Intelligence (AI) systems, seeking to understand the most effective strategies for mitigating discrimination and promoting equity in automated processes. Methodologically, this was an exploratory and comparative literature review in harmony with qualitative research. As a result, it was highlighted that reducing bias in AI is a complex task, which involves both adjustments to the data used to train the models and the application of specific techniques during and after the training process. The use of approaches such as supervised learning with regularization,

[1]Graduando em Análise e Desenvolvimento de Sistemas pela Faculdade de Tecnologia de Mogi das Cruzes – FATEC MC. Mogi das Cruzes – SP. E-mail: antonio.dias@fatec.sp.gov.br
[2]Docente, Faculdade de Tecnologia de Mogi das Cruzes FATEC MC. Mogi das Cruzes – SP.

model auditing and the use of fair algorithms emerge as promising practices in mitigating biases and promoting equity. It is concluded, in a complementary way, that, although challenges persist, continuous research and the implementation of ethical practices are essential for the development of more transparent, inclusive and socially responsible AI systems.

Key words: AI ethics; Algorithmic Justice; Artificial intelligence; Biases.

INTRODUÇÃO

Os vieses em Inteligência Artificial (IA) são distorções sistemáticas que podem ocorrer durante o desenvolvimento e o funcionamento dos sistemas computacionais, resultando em decisões ou previsões que favorecem ou discriminam certos grupos ou indivíduos. Esses vieses podem ser introduzidos de diversas formas, como nos dados de treinamento, na escolha dos algoritmos ou nas decisões de modelagem (Cozman; Kaufman, 2022). De fato, muitas vezes, os dados utilizados para treinar modelos de IA refletem desigualdades históricas ou sociais, o que, ao ser replicado nos sistemas, pode acentuar ou perpetuar esses mesmos vieses. Por exemplo, um sistema de IA projetado para selecionar candidatos a vagas de emprego pode reproduzir o viés implícito de preferir candidatos de determinado gênero ou etnia, caso esses padrões estejam presentes nos dados históricos utilizados para treiná-lo (Faustino; Bugalho, 2024).

A presença de vieses em sistemas de IA é problemática, pois pode resultar em decisões automatizadas que são injustas e discriminatórias. Em áreas como recrutamento, justiça criminal, saúde e finanças, esses vieses podem ter impactos profundos e duradouros, afetando a vida de indivíduos de maneira negativa. Estudar a redução de vieses em IA se apresenta como uma tarefa elementar diante do crescente uso dessas tecnologias em áreas fundamentais da sociedade moderna. Com a ascensão da IA, sistemas automatizados estão cada vez mais envolvidos em decisões que afetam diretamente a vida das pessoas, como a concessão de crédito, o diagnóstico médico, a seleção de candidatos a empregos e até mesmo a determinação de penas judiciais. Nesse contexto, a redução de vieses aprimora a

qualidade das decisões tomadas ao passo que também assegura que os sistemas operem de maneira justa e equitativa para todos, independentemente de características como gênero, etnia ou classe social.

O objetivo é averiguar os métodos de redução de vieses em sistemas de IA buscando compreender as estratégias mais eficazes para mitigar discriminação e promover a equidade nos processos automatizados, identificar as principais fontes de vieses em sistemas de IA incluindo os dados de treinamento e os algoritmos utilizados e investigar as abordagens existentes para a redução de vieses em IA destacando os métodos mais aplicados e seus resultados em diferentes contextos.

MATERIAL E MÉTODOS

A abordagem metodológica adotada neste estudo consistiu em uma revisão bibliográfica exploratória e comparativa, fundamentada em uma pesquisa qualitativa. Para complementar a análise, recorreu-se a bases documentais focadas em obras publicadas na última década. Como critérios de inclusão, foram descartados artigos que não apresentavam relevância para o tema, ou que eram de acesso restrito, que ultrapassavam a janela temporal definida, ou que estavam em desacordo com o idioma estabelecido (português e inglês). Além disso, foram excluídos fragmentos ou publicações incompletas.

Dentro desse contexto, o método comparativo foi utilizado para analisar duas séries ou fenômenos semelhantes de diferentes contextos sociais ou áreas do conhecimento, com o objetivo de identificar elementos comuns entre eles. Este método é amplamente utilizado em diversas disciplinas científicas, em especial nas ciências sociais, pois permite investigar grandes grupos humanos em populações distintas, localizadas em diferentes regiões geográficas (Fachin, 2005).

Além disso, destaca-se a utilização de escalas qualitativas, pois essa abordagem possibilita que a criatividade e a imaginação dos pesquisadores conduzam a elaboração de estudos que explorem novas perspectivas. A pesquisa documental, por exemplo, é reconhecida como uma metodologia inovadora que

pode fornecer importantes contribuições para o estudo de determinados temas. Os documentos, frequentemente considerados são fontes valiosas de dados, merecem atenção especial, pois podem enriquecer outros tipos de investigações qualitativas (Godoy, 1995).

Conforme Neves (1996), a pesquisa documental envolve a análise de informações que ainda não passaram por um tratamento analítico ou que podem ser "reexaminadas com o objetivo de uma nova ou complementar interpretação". Essa metodologia oferece uma base sólida para outros tipos de estudos qualitativos e possibilita que a criatividade do pesquisador oriente a investigação a partir de enfoques diferenciados.

A pesquisa se apoiou em obras e artigos científicos de autores nacionais e internacionais, com ênfase em publicações posteriores a 2015. Os dados foram extraídos de bases de dados como a *Scientific Electronic Library Online (Scielo)* e Google Acadêmico. A pesquisa foi pontuada com os seguintes termos-chave: “inteligência artificial”, “vieses”, “ética em IA” “redução de vieses” e “justiça algorítmica”.

REFERENCIAL TEÓRICO

A Inteligência Artificial (IA) tem se consolidado como uma das tecnologias mais transformadoras da atualidade, com uma crescente aplicação em diversos setores, como saúde, educação, finanças e justiça. No entanto, um dos desafios mais críticos que essa tecnologia enfrenta é a presença de vieses em seus sistemas. Os vieses em IA não são apenas um problema técnico, mas têm implicações sociais e éticas profundas, podendo comprometer a justiça e a equidade nas decisões automatizadas. A seguir, exploraremos a definição e os tipos de vieses, suas origens e os impactos que geram na sociedade (Martins et al., 2024).

Vieses em Inteligência Artificial referem-se a distorções sistemáticas nas decisões ou previsões feitas por sistemas de IA, que podem resultar em resultados injustos ou discriminatórios. Em termos simples, um viés ocorre quando o sistema

favorece ou prejudica certos grupos ou indivíduos com base em características que não deveriam influenciar a decisão, como etnia, gênero ou classe social. Esses vieses podem se manifestar de várias formas, sendo alguns dos mais comuns: o viés de seleção, em que dados tendenciosos são escolhidos para treinamento; o viés de amostragem, quando os dados não representam adequadamente toda a população; e o viés de algoritmo, que ocorre quando os próprios algoritmos replicam ou amplificam os preconceitos presentes nos dados de treinamento (Kaufman, 2022).

É importante salientar a diferença entre vieses humanos e vieses de sistemas automatizados. Enquanto os vieses humanos são muitas vezes inconscientes e podem ser atribuídos a preconceitos pessoais ou sociais, os vieses de IA surgem a partir de decisões algorítmicas baseadas em dados que refletem as desigualdades presentes na sociedade. Embora os sistemas de IA sejam projetados para serem objetivos, os algoritmos, por sua natureza, podem inadvertidamente perpetuar ou até intensificar os preconceitos humanos que existem nos dados (Alff, 2020).

Para além disso, as decisões de design tomadas pelos desenvolvedores desempenham um papel crucial na introdução de vieses. Quando as escolhas de modelagem e as abordagens de treinamento não são cuidadosamente ponderadas, é possível que o sistema reflita preconceitos preconcebidos, seja pela seleção inadequada de variáveis ou pela falta de uma análise crítica das fontes de dados. Em alguns casos, os desenvolvedores podem não estar plenamente conscientes de que suas decisões podem resultar em discriminação, já que os vieses podem ser sutis e não facilmente detectados em fases iniciais do processo de desenvolvimento (Medeiros et al., 2023).

Em casos mais extremos, como na justiça criminal, sistemas de IA que auxiliam na decisão de penas podem replicar discriminação racial, com consequências fatais para minorias raciais. Por exemplo, algoritmos usados para prever a reincidência criminal podem ser tendenciosos, já que muitas vezes são treinados com dados históricos que refletem práticas discriminatórias de policiamento e encarceramento.

Isso não apenas afeta o indivíduo em questão, mas também gera um ciclo vicioso de marginalização e exclusão (Cozman; Kaufman, 2022).

Os impactos em grupos marginalizados são particularmente graves. Minorias raciais, mulheres e pessoas com deficiência, frequentemente em desvantagem em vários aspectos sociais, podem ver suas oportunidades ainda mais restringidas pela IA enviesada. Esses grupos são, muitas vezes, os mais vulneráveis aos efeitos adversos de sistemas automatizados, pois os dados históricos utilizados para treinar as máquinas podem não refletir adequadamente suas realidades (Kaufman, 2022).

Existem diversos tipos de vieses que podem ser mitigados na etapa de pré-processamento. Por exemplo, o viés de amostragem ocorre quando a distribuição de dados no conjunto de treinamento não representa adequadamente a população geral. A sub-representação de determinados grupos, como minorias raciais ou de gênero, pode ser corrigida através de técnicas de balanceamento de dados. Além disso, o viés de confirmação, que acontece quando os dados favorecem certas crenças ou padrões preconceituosos, também pode ser atenuado ao garantir que os dados estejam devidamente diversificados e representem de forma justa todos os grupos envolvidos no processo de tomada de decisão (Pecego; Teixeira, 2024).

O processo de limpeza de dados visa identificar e remover dados que possam ser incorretos ou enviesados antes de iniciar o treinamento do modelo. A identificação de dados enviesados pode ser feita através da análise cuidadosa dos dados, observando padrões que indicam a presença de vieses, como a falta de representação de certos grupos ou a prevalência de características distorcidas. A remoção de tais dados é essencial, uma vez que a manutenção de registros enviesados no conjunto de dados pode influenciar negativamente a performance do modelo (Cozman; Kaufman, 2022).

Inclusive, técnicas como *oversampling* e *undersampling* são amplamente utilizadas para balancear dados. O *oversampling* envolve a replicação de dados de classes sub-representadas para garantir que essas classes tenham um número adequado de instâncias durante o treinamento. O *undersampling*, por sua vez, reduz

o número de instâncias das classes super-representadas, equilibrando assim o conjunto de dados. Ambas as técnicas ajudam a garantir que o modelo não desenvolva preconceitos para uma classe em detrimento de outras, resultando em um aprendizado mais equitativo (Toledo; Pessoa, 2024).

RESULTADOS E DISCUSSÃO

A redução de vieses durante o treinamento do modelo tem como intuito garantir que o sistema de IA não aprenda padrões discriminatórios ou injustos a partir dos dados de treinamento. Durante essa fase, os algoritmos de aprendizado de máquina são ajustados de modo a não apenas maximizar a precisão do modelo, mas também a garantir que as previsões não favoreçam injustamente certos grupos em detrimento de outros. O treinamento é o momento em que o modelo pode internalizar as tendências e vieses presentes nos dados, seja através de variáveis sensíveis ou da própria distribuição dos dados. Portanto, um controle rigoroso nessa etapa é essencial para prevenir que os vieses se perpetuem e prejudiquem os resultados do modelo.

Ao incorporar técnicas específicas de redução de vieses durante o treinamento, é possível ajustar os algoritmos de aprendizado de máquina para que eles se tornem mais conscientes dos potenciais vieses nos dados. Tais ajustes podem envolver mudanças no processo de otimização, na função de custo e nas técnicas de regularização, de forma a penalizar os modelos que gerem previsões enviesadas. Com isso, é possível promover a imparcialidade no modelo, tornando-o mais justo e equitativo ao longo do processo de aprendizado.

No contexto do aprendizado supervisionado, o modelo é treinado com dados rotulados, e seu objetivo é minimizar o erro entre as previsões feitas e os valores reais. Para reduzir vieses durante esse processo, é possível aplicar técnicas de regularização, que visam evitar que o modelo se ajuste excessivamente aos dados de treinamento, o que pode levar à internalização de padrões enviesados. As técnicas de regularização, como L1 (*Lasso Regression*) e L2 (*Ridge Regression*),

ajudam a controlar a complexidade do modelo, forçando-o a generalizar melhor em relação a novos dados, em vez de se ajustar a ruídos ou a distribuições desequilibradas nos dados (Cozman; Kaufman, 2022).

A regularização L1 e L2 são comumente utilizadas para evitar o sobreajuste (*overfitting*), penalizando coeficientes grandes ou não significativos nas variáveis de entrada. No caso da redução de vieses, essas penalizações também podem ser direcionadas para características que possam contribuir para a discriminação ou para a amplificação de padrões prejudiciais. Além disso, o uso de penalizações específicas para viés pode ser implementado durante a fase de treinamento, de modo a garantir que o modelo seja penalizado toda vez que uma decisão injusta ou enviesada for tomada, o que incentiva o modelo a modificar sua abordagem de forma a ser mais justo e imparcial.

Por conseguinte, outra estratégia dentro do aprendizado supervisionado envolve o uso de algoritmos de aprendizado de máquina que favorecem imparcialidade, como o *fairness-aware learning*. Esses algoritmos são projetados para considerar a equidade como um critério adicional ao lado da precisão, otimizando a função de custo de modo a levar em conta a necessidade de reduzir disparidades nas previsões feitas para diferentes grupos de indivíduos, com base em características sensíveis como raça, gênero ou idade.

Existem modelos de aprendizado de máquina que foram especificamente projetados para mitigar os vieses durante o processo de treinamento. Esses modelos implementam técnicas que permitem identificar e corrigir distorções nas previsões, de forma a garantir que o sistema de IA não favoreça injustamente um grupo em relação a outro. Entre as abordagens mais eficazes, destaca-se o uso de redes neurais que são adaptadas para detectar e corrigir vieses, promovendo a imparcialidade em suas previsões.

As redes neurais podem ser modificadas para incluir camadas ou algoritmos de ajuste que penalizam as previsões enviesadas. Por exemplo, pode-se aplicar técnicas de *adversarial debiasing*, onde um modelo adversário é treinado

simultaneamente com o modelo principal para identificar e corrigir qualquer viés presente nas previsões do modelo. Além disso, modelos como o *fairness-aware neural networks* são projetados para integrar restrições de justiça diretamente na função de perda da rede neural, de modo a otimizar simultaneamente a precisão e a imparcialidade do modelo (Medeiros et al., 2023).

A adaptação de modelos de aprendizado de máquina também pode envolver a implementação de detecção de viés em tempo real, durante o treinamento, para identificar e corrigir qualquer viés à medida que ele surge. Essa abordagem envolve monitoramento contínuo dos resultados do modelo, com base em métricas de equidade, e ajustes automáticos nos parâmetros do modelo para evitar a amplificação de padrões discriminatórios.

Uma das abordagens mais promissoras para reduzir vieses durante o treinamento de modelos é a incorporação de restrições de equidade diretamente na fase de treinamento. A ideia central é garantir que o modelo atenda a critérios de justiça e imparcialidade ao mesmo tempo em que é treinado para realizar tarefas como classificação, previsão ou decisão. Essas restrições podem ser adicionadas de diversas formas, dependendo do tipo de problema e do modelo utilizado.

Técnicas como o *reweighting*, onde as instâncias de dados são ponderadas de acordo com sua importância para a equidade do modelo, são utilizadas para ajustar a forma como o modelo aprende a partir dos dados. O *reweighting* busca garantir que as classes sub-representadas ou discriminadas tenham uma maior influência no treinamento, de forma a compensar desigualdades nas distribuições dos dados. Já o *adversarial debiasing*, mencionado anteriormente, pode ser utilizado para treinar modelos adversários que, ao identificar previsões enviesadas, ajustam o modelo principal para removê-las.

A redução de vieses não deve ser vista como um processo limitado às fases iniciais de desenvolvimento e treinamento de modelos de inteligência artificial. Após a implementação de um modelo, a avaliação e o ajuste dos resultados tornam-se igualmente essenciais para garantir que os sistemas de IA operem de maneira justa

e imparcial. O pós-processamento desempenha um papel crucial nesse contexto, proporcionando a oportunidade de revisar e ajustar as previsões do modelo, assim como detectar e corrigir possíveis distorções que possam persistir após o treinamento. Este capítulo explora as metodologias de redução de vieses após a implementação do modelo, destacando a importância de auditorias, ajustes pós-processamento e a avaliação contínua do impacto de vieses nas decisões automatizadas.

A avaliação dos resultados pós-processamento é uma etapa crucial para garantir que os modelos de IA continuem a produzir resultados justos após a sua implementação. Muitas vezes, embora os algoritmos sejam ajustados para reduzir vieses durante o treinamento, é apenas na fase de implementação que se pode observar plenamente os efeitos dos vieses presentes nos dados ou na definição do modelo. A identificação e correção de vieses após o treinamento são, portanto, essenciais para mitigar impactos indesejáveis que possam surgir quando o modelo for utilizado em contextos reais.

Existem diversas estratégias que podem ser aplicadas para ajustar as previsões feitas por modelos enviesados após seu treinamento. Uma dessas estratégias envolve a modificação das decisões do modelo com base em regras de equidade ou por meio da introdução de ajustes de pós-processamento, onde as previsões do modelo são ajustadas para atender a critérios de justiça estabelecidos. Essas modificações podem incluir a alteração dos limiares de decisão ou a aplicação de ponderações adicionais para garantir que os grupos mais vulneráveis ou marginalizados não sejam injustamente desfavorecidos. Além disso, técnicas de recalibração de modelos podem ser empregadas para garantir que as previsões do modelo atendam a requisitos éticos e sociais, ajustando a distribuição de erros de previsão entre os diferentes grupos de interesse.

A auditoria de modelos de IA é uma prática fundamental para identificar e corrigir vieses após a implementação de sistemas automatizados. Auditorias periódicas ajudam a garantir que os modelos de IA não apresentem discriminação

ou resultados inconsistentes, permitindo a correção de eventuais falhas que possam ter surgido após o treinamento. As auditorias podem ser realizadas utilizando diversas técnicas e ferramentas, como a análise de desempenho do modelo em diferentes subgrupos de dados, a detecção de viés por meio de métricas específicas e a comparação de resultados para garantir que não haja discriminação sistemática.

As métricas de *fairness* são ferramentas essenciais para avaliar a equidade dos modelos de IA. Entre as métricas mais utilizadas, destaca-se a igualdade de oportunidade, que analisa se diferentes grupos têm as mesmas probabilidades de obter resultados positivos, e a precisão equilibrada, que verifica se o modelo apresenta precisão semelhante para todos os grupos analisados. Essas métricas permitem uma avaliação mais profunda dos modelos, identificando possíveis distorções que podem ser corrigidas. A implementação de auditorias regulares também é uma prática importante para evitar vieses persistentes ao longo do tempo. A monitorização contínua das previsões do modelo e o uso de métricas de equidade são fundamentais para garantir que os sistemas automatizados não se tornem obsoletos ou prejudiquem determinados grupos à medida que novos dados são incorporados.

Após a implementação do modelo, é possível que algumas distorções ainda persistam, seja devido a vieses em características sensíveis ou falhas no equilíbrio das previsões. O ajuste pós-processamento e a recalibração de modelos são, portanto, passos essenciais para corrigir essas falhas. A recalibração envolve a modificação dos parâmetros do modelo para alinhar melhor suas previsões com as expectativas de equidade, sem sacrificar sua eficácia em termos de precisão. Esse processo pode incluir o ajuste das probabilidades de decisão ou a modificação dos limiares de classificação para assegurar que o modelo apresente desempenho similar entre os diferentes grupos.

A avaliação de impacto e a medição de vieses são componentes basilares para garantir a justiça contínua dos modelos de IA. Ferramentas e métodos quantitativos podem ser empregados para medir como os vieses afetam diferentes grupos,

especialmente aqueles que historicamente enfrentam discriminação ou marginalização. Para isso, é importante realizar uma avaliação contínua do desempenho do modelo, acompanhando as métricas de equidade e monitorando quaisquer mudanças nos padrões de desempenho ao longo do tempo.

A medição de vieses deve ser um processo iterativo, em que o modelo é periodicamente testado quanto à sua imparcialidade em relação a diferentes grupos. A implementação de *feedback* contínuo também é uma prática essencial para aprimorar a equidade dos modelos. Com isso, é possível ajustar os modelos com base nos resultados das auditorias e nas observações do impacto real das previsões, garantindo que os sistemas de IA evoluam de maneira justa e sem perpetuar desigualdades. O ciclo de avaliação e iteração contínua é, portanto, uma estratégia crucial para assegurar que a IA atenda aos princípios de justiça e equidade, ao mesmo tempo em que mantém sua precisão e funcionalidade.

A IA Explicável (XAI) emerge como uma das abordagens mais promissoras para enfrentar os desafios relacionados à transparência e imparcialidade dos modelos de IA. A explicabilidade torna-se fundamental, pois permite que os desenvolvedores, usuários e auditores compreendam como as decisões dos modelos são tomadas, facilitando a identificação e correção de vieses presentes. Em modelos complexos, como redes neurais profundas, a falta de compreensão do funcionamento interno dos algoritmos pode ocultar vieses, resultando em decisões potencialmente injustas ou discriminatórias. Portanto, a explicabilidade oferece uma janela para avaliar os processos internos do modelo, essencial para detectar e mitigar vieses que, de outra forma, poderiam passar despercebidos.

Por conseguinte, modelos explicáveis promovem a transparência ao passo que também contribuem para a imparcialidade. Ao proporcionar uma melhor compreensão das decisões automatizadas, torna-se mais fácil identificar as áreas onde os vieses estão sendo aplicados. Com essa abordagem, usuários e reguladores podem realizar uma auditoria mais eficaz dos sistemas, promovendo maior confiança e aceitação. O uso de ferramentas de IA explicável, como o LIME

(*Local Interpretable Model-Agnostic Explanations*) e o SHAP (*Shapley Additive Explanations*), permite que os pesquisadores analisem quais variáveis influenciam as previsões e como as mudanças nesses fatores afetam os resultados. Essas ferramentas têm um impacto direto na auditoria e na mitigação de vieses, fornecendo uma camada adicional de verificabilidade e controle sobre os modelos.

As técnicas exploradas nesta pesquisa para a mitigação da ocorrência de vieses possuem, cada uma, um conjunto de vantagens e de limitações, como observado na tabela 1.

Tabela 1. Comparação das técnicas exploradas.

TÉCNICA	FASE APLICADA	VANTAGENS	LIMITAÇÕES
Reweighting	Pré-processamento	Balanceia amostras para melhorar a representatividade.	Pode induzir variância elevada.
Undersampling / ***Oversampling***	Pré-processamento	Melhora distribuição das classes minoritárias.	Não resolve viés estrutural.
Regularização L1 / L2	Durante o treinamento	Simples de implementar; reduz o *overfitting.*	Pode não eliminar totalmente o viés nos dados.
Adversarial Debiasing	Durante o treinamento	Corrige viés durante o aprendizado via modelo adversário.	Alta complexidade computacional.
Fairness-aware Learning	Durante o treinamento	Considera equidade como critério de otimização.	Pode reduzir a precisão do modelo.
Pós-processamento com Recalibração	Pós-processamento	Corrige resultados sem retreinar o modelo.	Pode suavizar demais as previsões.
Auditoria	Pós-processamento	Monitora continuamente resultados enviesados.	Demanda métricas e estrutura avaliativa constante.
IA Explicável (XAI)	Pós-processamento	Aumenta transparência e facilita auditoria.	Requer ferramentas e conhecimento técnico especializado.

Fonte: Elaborada pelos autores, (2025).

Fica a critério, portanto, do utilizador, escolher a melhor técnica ou conjunto de técnicas a serem utilizadas em seu projeto, de acordo com seu propósito e capacidade.

CONCLUSÃO

A redução de vieses em IA é uma questão crucial tanto do ponto de vista técnico quanto em relação às suas implicações sociais, econômicas e éticas. Técnicas como L1 e L2, além de abordagens mais inovadoras como o aprendizado de máquinas conscientes de imparcialidade (*fairness-aware learning*), foram identificadas como essenciais para criar modelos mais equitativos. Em paralelo, a análise de estratégias pós-processamento, como auditorias de modelos e a recalibração, revelou-se fundamental para ajustar e readequar as previsões realizadas por modelos enviesados. Além disso, o papel das tecnologias emergentes, como a IA explicável (XAI), representa um avanço crucial na transparência e no entendimento dos processos internos dos modelos de IA, onde a explicabilidade facilita a auditoria, permitindo que os desenvolvedores e usuários identifiquem e corrijam vieses, contribuindo para um uso mais ético da inteligência artificial. Diante dessas abordagens, é evidente que a construção de uma IA mais justa e imparcial demanda a implementação de práticas rigorosas de auditoria, transparência e regulação.

REFERÊNCIAS BIBLIOGRÁFICAS

ALFF, H. P. **A inteligência artificial como forma de repensar o ingresso em juízo: o novo paradigma pelo abrandamento dos vieses.** Derecho público y privado ante las nuevas tecnologías, p. 49-58, 2020. Disponível em: https://www.torrossa.com/. Acesso em: 22 mar. 2025.

COZMAN, F. G.; KAUFMAN, D. **Viés no aprendizado de máquina em sistemas de inteligência artificial: a diversidade de origens e os caminhos de mitigação.** Revista USP, n. 135, p. 195-210, 2022. Disponível em: https://www.revistas.usp.br/. Acesso em: 16 mar. 2025.

FACHIN, O. **Fundamentos de Metodologia.** 5ª edição. Revista e atualizada pela norma da ABNT 14724, de 30/12/2005 Ed. Hora Saraiva. Disponível em: http://maratavarespsictics.pbworks.com/. Acesso em: 28 fev. 2025.

FAUSTINO, G.; BUGALHO, A. C. **Inteligência Artificial: Da Análise do Viés Discriminatório à Supremacia da Eficiência Humana sobre Numérica no Judiciário.** In: Anais do Congresso Brasileiro de Processo Coletivo e Cidadania. 2024. p. 906-931.Disponível em: https://revistas.unaerp.br/. Acesso em: 22 mar. 2025.

GODOY, A. S. **Introdução à pesquisa qualitativa e suas possibilidades.** RAE - Revista de Administração de Empresas, São Paulo, v. 35, n. 2, p. 57-63, 1995. Disponível em: https://www.scielo.br/. Acesso em: 01 mar. 2025.

KAUFMAN, Dora. **Desmistificando a inteligência artificial.** Autêntica Editora, 2022.

MARTINS, A. C. T. et al. **Como a inteligência artificial pode auxiliar na tomada de decisões de processos dentro do tribunal de justiça.** 2024. Disponível em: https://ric.cps.sp.gov.br/. Acesso em: 25 mar. 2025.

MEDEIROS, A. C. B. et al. **Grace: Sistema de recomendação de currículos com inteligência artificial.** In: Simpósio Brasileiro de Banco de Dados (SBBD). SBC, 2023. p. 8-14. Disponível em: https://sol.sbc.org.br/. Acesso em: 11 mar. 2025.

NEVES, J. L. **Pesquisa Qualitativa – Características, Usos e Possibilidades.** Caderno de Pesquisas em Administração, São Paulo, v.1, nº 3, 2º Sem./1996. Disponível em: https://www.hugoribeiro.com.br/. Acesso em: 05 mar. 2025.

PECEGO, D. N.; TEIXEIRA, R. L. C. **Inteligência Artificial no Judiciário: Da Opacidade à Explicabilidade das Decisões Judiciais.** Revista da Faculdade de Direito da Uerj, n. 43, 2024. Disponível em: https://search.ebscohost.com/. Acesso em: 18 mar. 2025.

TOLEDO, C.; PESSOA, D. **O uso de inteligência artificial na tomada de decisão judicial.** Revista de Investigações Constitucionais, v. 10, p. e237, 2024. Disponível em: https://www.scielo.br/. Acesso em: 24 mar. 2025.