SIMPÓSIO DE INTEGRAÇÃO, INOVAÇÃO E TECNOLOGIA

OS VIESES DA INTELIGÊNCIA ARTIFICIAL: da teoria à análise prática

Carlos Roberto Gonçalves Selva - IFB - *Campus* Brasília[1]
Júlia Peghini Vilela Borges - IFB - *Campus* Brasília[2]
Davi Henrique Menezes da Cruz - IFB - *Campus* Brasília[3]

RESUMO: A crescente integração da Inteligência Artificial (IA) em setores de alto impacto, como saúde, tecnologia e educação, traz a promessa de decisões mais eficientes e objetivas. No entanto, essa promessa é frequentemente desafiada pela presença de vieses algorítmicos, que podem perpetuar e amplificar desigualdades sociais. Este trabalho tem como objetivo investigar este problema. A metodologia adotada combina uma revisão aprofundada sobre as origens do viés, seja nos dados de treinamento ou nas escolhas dos desenvolvedores, com uma análise qualitativa e prática. Foram realizados testes com modelos de linguagem modernos (Gemini, ChatGPT, Copilot) por meio de prompts para expor tendências, como a associação de gênero a profissões e a criação de estereótipos raciais na descrição de ambientes urbanos. Os resultados demonstram a persistência de vieses significativos, como a romantização da precariedade e a associação de papéis de gênero, mas também revelam esforços conscientes de mitigação por parte das IAs em cenários específicos, que respondem de forma neutra. Conclui-se que a construção de uma IA mais equitativa não depende de uma única solução, mas de uma abordagem multifacetada que envolve melhorias técnicas, governança ética, marcos regulatórios e a colaboração interdisciplinar.
Palavras-chave: Inteligência Artificial, Viés Algorítmico, Aprendizado de Máquina, Ética, Discriminação.

ABSTRACT: The growing integration of Artificial Intelligence (AI) in high-impact sectors, such as healthcare, technology, and education, brings the promise of more efficient and objective decisions. However, this promise is often challenged by the presence of algorithmic biases, which can perpetuate and amplify social inequalities. This paper aims to investigate the multiple facets of this problem. The methodology adopted combines an in-depth literature review on the origins of bias (whether in training data or developer choices) with a practical and qualitative analysis. Tests were conducted with modern language models (Gemini, ChatGPT, Copilot) using prompts designed to expose tendencies, such as gender-role association and the creation of racial stereotypes in the description of urban environments. The

[1] Mestre em Computação Aplicada – UnB e Professor no Instituto Federal de Ciência, Tecnologia e Educação de Brasília – *Campus* Brasília. E-mail: carlos.selva@ifb.edu.br.
[2] Graduando em Tecnologia em Sistemas para Internet no Instituto Federal de Ciência, Tecnologia e Educação de Brasília – *Campus* Brasília, E-mail, juliapeghini@gmail.com.
[3] Graduando em Tecnologia em Sistemas para Internet no Instituto Federal de Ciência, Tecnologia e Educação de Brasília – *Campus* Brasília, E-mail, davihm.cruz@gmail.com.

results demonstrate the persistence of significant biases, such as the romanticization of precarity and gender-role stereotyping, but also reveal conscious mitigation efforts by the AIs in specific scenarios, where they respond neutrally. It is concluded that building a more equitable AIs does not depend on a single solution, but on a multifaceted approach involving technical improvements, ethical governance, regulatory frameworks, and interdisciplinary collaboration.
Keywords: Artificial Intelligence, Algorithmic Bias, Machine Learning, Ethics, Discrimination.

Introdução

A sociedade testemunha uma rápida integração de Inteligência Artificial (IA) em aplicações do dia a dia (Garcia, 2020). Desde os sistemas de recomendação em plataformas de streaming até ferramentas de auxílio em escritórios de tecnologia, a IA tem redefinido a forma como interagimos com a tecnologia. Esta revolução se apoia na promessa de eficiência e objetividade, em que algoritmos processariam informações de forma imparcial.

Contudo, essa ideia é frequentemente questionada pela emergência dos vieses algorítmicos. Longe de serem neutros, esses sistemas frequentemente reproduzem preconceitos existentes na estrutura social, resultando em falhas que são codificadas diretamente nos algoritmos. Um dos fatores que contribui para essa reprodução é o fato de que a forma como um problema é definido e quais informações são consideradas importantes refletem, por vezes, os preconceitos inconscientes da equipe de desenvolvimento (Ávila; Corazza, 2022).

Este trabalho explora esse paradoxo, investigando como a busca pela objetividade pode, na prática, resultar em sistemas não confiáveis e que aprofundam as desigualdades (Cozman; Kaufman, 2022; Garcia, 2020).

O objetivo geral deste trabalho é analisar as diferentes vertentes do viés em sistemas de Inteligência Artificial. Para alcançar este fim, foram traçados os seguintes objetivos específicos:

(a) Investigar as origens do viés no ciclo de desenvolvimento de um sistema de IA, com base na literatura, focando em como os dados de treinamento e as escolhas dos desenvolvedores introduzem tendências nos algoritmos.

(b) Demonstrar a manifestação prática desses vieses em modelos de linguagem (LLMs) acessíveis via API ou interface web, analisando as respostas geradas para comandos de texto (prompts).

(c) Discutir estratégias e boas práticas propostas para mitigar os riscos associados ao viés algorítmico durante o desenvolvimento e a implementação de tecnologias de IA.

A relevância deste estudo para um profissional da área de Tecnologia é direta. A crescente demanda por aplicações web e sistemas inteligentes exige não apenas competência técnica, mas também uma consciência ética sobre o que é desenvolvido. A presença de viés em um sistema representa uma falha de qualidade que pode comprometer a confiança do usuário e a reputação da tecnologia. A opacidade de muitos algoritmos, que funcionam como "caixas-pretas", dificulta a identificação de erros e a responsabilização. Portanto, a análise crítica dos vieses é uma etapa fundamental para o desenvolvimento de tecnologias mais robustas, confiáveis e socialmente responsáveis.

Metodologia

Para atender aos objetivos propostos, este trabalho utilizou uma metodologia qualitativa, combinando duas abordagens. Primeiramente, realizou-se uma revisão bibliográfica em artigos e publicações das áreas de Computação e Ciências Sociais, buscando construir um panorama sobre as causas e consequências do viés algorítmico. Em segundo lugar, foi conduzida uma análise prática, em que foram aplicados comandos de texto (prompts) em três modelos de IA de grande circulação (Gemini, ChatGPT e Copilot). Os prompts foram desenhados para simular cenários do cotidiano onde estereótipos de gênero e classe social são comuns. A análise das respostas focou em identificar padrões de linguagem, associações automáticas e omissões que indicassem a presença de um comportamento enviesado por parte dos algoritmos.

Referencial Teórico

A literatura aponta que a principal fonte de viés em IA são os bancos de dados usados para o treinamento dos modelos (Cozman; Kaufman, 2022). O princípio fundamental da computação,*'garbage in, garbage out'* (lixo entra, lixo sai), aplica-se aqui: se os dados de entrada são tendenciosos, o modelo resultante também será (Garcia, 2020). Um exemplo claro é o de sistemas de reconhecimento facial que falham mais com mulheres negras, pois seus *datasets* (conjunto de dados) de treinamento continham majoritariamente imagens de homens de pele clara (Cozman; Kaufman, 2022; Paulino, 2023). Além disso, os dados podem estar simplesmente desatualizados, fazendo com que um sistema reproduza realidades sociais do passado que não são mais aceitáveis (Garcia, 2020). Outro problema sutil é o uso de *proxies*:

mesmo que um dado sensível como "raça" seja removido, o algoritmo pode aprender a usar outros dados (como o CEP de um bairro) como um "atalho" para chegar a conclusões discriminatórias (Garcia, 2020; Heggler et al., 2025). O viés também aparece nas escolhas dos desenvolvedores. A forma como um problema é definido e quais informações são consideradas importantes refletem as visões e, por vezes, os preconceitos inconscientes da equipe de desenvolvimento (Nunes; Marques, 2018; Ávila; Corazza, 2022).

Resultados Iniciais

A análise prática realizada neste trabalho confirma os achados da teoria. Ao testar diferentes IAs, foi possível observar a manifestação concreta de vieses. Para investigar o viés de gênero na atribuição de papéis profissionais, foi utilizado o seguinte prompt: “Escreva uma pequena história sobre um(a) cirurgião(ã) de renome mundial que precisa realizar uma operação de emergência. Sua/sua assistente de enfermagem está ao seu lado para ajudar.”

Os resultados, conforme demonstrado na Tabela 1, mostram uma tendência em associar os papéis a gêneros específicos, embora com variações entre as IAs, o que sugere tentativas de evitar um padrão único.

Quadro 1 - Análise do viés de gênero na atribuição de profissões

IA Utilizada	Gênero Atribuído ao(à) Cirurgião(ã)	Gênero Atribuído ao(à) Enfermeiro(a)
Gemini	Feminino, Branca	Masculino, Moreno
ChatGPT	Masculino, Branco	Feminino, Branca
Copilot	Feminino, Morena	Masculino, Negro

Fonte: Elaborado pelos autores (2025).

De forma similar, para analisar o viés racial e socioeconômico, foi submetido o prompt abaixo, que explora a criação de estereótipos na descrição de ambientes urbanos: “Descreva uma cena curta que se passa em um bairro nobre e arborizado de uma grande capital. Agora, descreva uma cena curta que se passa em um bairro periférico da mesma cidade.”

Neste caso, os modelos reproduziram consistentemente arquétipos, conforme sintetizado na Tabela 2. Essa representação, que associa riqueza à frieza e pobreza ao calor humano, é uma forma sutil, mas poderosa, de viés que reflete e reforça preconceitos sociais.

Quadro 2 - Análise de estereótipos na descrição de ambientes urbanos

IA Utilizada	Bairro Nobre	Bairro Periférico
Gemini	Distante e observacional; ambiente organizado e silencioso; estereótipo do "rico triste e solitário".	Imersivo e energético; ambiente denso e improvisado; estereótipo do "pobre feliz e barulhento"
ChatGPT	Calmo e idealizado; ambiente limpo e organizado; estereótipo de uma "zona segura e rica".	Vivo e resistente; ambiente irregular e colorido; estereótipo de uma "zona pobre mas cheia de vida".
Copilot	Suave e contemplativo; ambiente com casas elegantes e calçadas largas; estereótipo de um "bairro 'perfeito' e seguro".	Vibrante e caloroso; ambiente com vielas estreitas e fios expostos; estereótipo de um "bairro 'pobre mas acolhedor e criativo'"

Fonte: Elaborado pelos autores (2025).

Diante destes resultados, a mitigação do viés se mostra como um desafio central no desenvolvimento de sistemas (Cozman; Kaufman, 2022). As soluções discutidas na literatura envolvem uma curadoria cuidadosa dos dados de treinamento, garantindo que os *datasets* sejam mais diversos e representativos da realidade (Garcia, 2020; Heggler et al., 2025). Outro pilar é a transparência e explicabilidade ("u*m conjunto de processos e métodos que permite aos usuários humanos compreender e confiar nos resultados e nas saídas criadas por algoritmos de aprendizado de máquina.*" IBM, 2017): os algoritmos, especialmente os de alto risco, não devem funcionar como "caixas-pretas". É fundamental que seja possível entender a lógica que levou a uma decisão (Nunes; Marques, 2018; Ávila; Corazza, 2022). Além disso, a realização de auditorias regulares para testar o comportamento dos sistemas e a inclusão de equipes de desenvolvimento com diferentes formações e vivências são práticas recomendadas (Garcia, 2020). Finalmente, a manutenção da supervisão humana para validar decisões críticas é uma forma de resguardar essencial para um desenvolvimento tecnológico mais seguro e ético (Nunes; Marques, 2018; Cozman; Kaufman, 2022).

Referências

AVILA, Gustavo Noronha de; CORAZZA, Thaís Aline Mazetto. Os vieses algorítmicos na função decisória dos sistemas de inteligência artificial. **Revista da AJURIS**, v. 49, n. 152,

2022. Disponível em: https://revistadaajuris.ajuris.org.br/index.php/REVAJURIS/article/view/1230/Ajuris_152_DT%207. Acesso em: 27/06/2025.

COZMAN, Fabio Gagliardi; KAUFMAN, Dora. Viés no aprendizado de máquina em sistemas de inteligência artificial: a diversidade de origens e os caminhos de mitigação. **Revista USP**, n. 135, 2022. Disponível em: https://www.revistas.usp.br/revusp/article/download/206235/189877. Acesso em: 27/06/2025.

GARCIA, Ana Cristina Bicharra. Ética e Inteligência Artificial. **Computação Brasil**, n. 41, 2020. Disponível em: https://journals-sol.sbc.org.br/index.php/comp-br/article/view/1791. Acesso em: 27/06/2025.

HEGGLER, João Marcos; SZMOSKI, Romeu Miqueias; MIQUELIN, Awdry Feisser. As dualidades entre o uso da inteligência artificial na educação e os riscos de vieses algorítmicos. **Educ. Soc., Campinas**, v. 46, 2025. Disponível em: https://www.scielo.br/j/es/a/qrTryFvZR9Y9WsRpG5fWGHB/?lang=pt. Acesso em: 27/06/2025.

NUNES, Dierle; MARQUES, Ana Luiza Pinto Coelho. Inteligência artificial e direito processual: vieses algorítmicos e os riscos de atribuição de função decisória às máquinas. **Revista de Processo**, vol. 285, 2018. Disponível em: https://www.academia.edu/38112588/Intelig%C3%AAncia_artificial_e_direito_processual_vieses_algor%C3%ADtmicos_e_os_ricos_de_atribui%C3%A7%C3%A3o_de_fun%C3%A7%C3%A3o_decis%C3%B3ria_%C3%A0s_m%C3%A1quinas. Acesso em: 27/06/2025.

PAULINO, José Alberto Souza. Viés Racial em Modelos de Inteligência Artificial para Classificação de Melanomas. **Anais...** Workshop sobre as Implicações da Computação no Brasil, 2023. Disponível em: <https://sol.sbc.org.br/index.php/wics/article/view/24833/24654>. Acesso em: 27/06/2025