15º Congresso de Inovação, Ciência e Tecnologia do IFSP - 2024

Um estudo sobre o viés nos algoritmos de inteligência artificial

Y. Sant'Anna, F. Fumes

Graduando no Técnico em Mecânica Integrado ao Ensino Médio, Bolsista PIBIFSP, IFSP, Campus Hortolândia, yasmin.sant@aluno.ifsp.edu.br

Área de conhecimento (Tabela CNPq): 1.03.03.04-9 Sistemas de Informação

RESUMO

Este estudo investiga o viés racial e de gênero em algoritmos de inteligência artificial utilizados em programas gratuitos de geração de imagens. Por meio da análise de descrições textuais neutras e gerais, foram geradas imagens para identificar padrões de viés presentes nos resultados. A pesquisa demonstrou que, em muitos casos, os algoritmos tendem a associar certas características físicas e traços a determinados grupos raciais e de gênero, refletindo e reforçando estereótipos sociais. O estudo sugere que essas tendências podem ser influenciadas pelos dados usados para treinar os modelos de IA, o que levanta questões importantes sobre a responsabilidade e a ética no desenvolvimento dessas tecnologias.

PALAVRAS-CHAVE: viés algorítmico, IA, viés racial, viés de gênero, geração de imagens, ética em IA.

A study on bias in artificial intelligence algorithms

ABSTRACT: This study investigates racial and gender bias in artificial intelligence algorithms used in free image generation programs. By analyzing neutral and general textual descriptions, images were generated to identify bias patterns in the results. The research showed that, in many cases, algorithms tend to associate certain physical characteristics and traits with specific racial and gender groups, reflecting and reinforcing social stereotypes. The study suggests that these tendencies may be influenced by the data used to train the AI models, raising important questions about responsibility and ethics in the development of these technologies.

KEYWORDS: algorithmic bias, AI, racial bias, gender bias, image generation, AI ethics.

INTRODUÇÃO

O viés algorítmico na inteligência artificial (IA) ocorre quando sistemas tomam decisões tendenciosas com base em dados preconceituosos, que refletem preconceitos humanos. Esse viés pode se manifestar de diversas formas, como discriminação racial e de gênero, muitas vezes de maneira sutil e não intencional. A motivação para o estudo do viés algorítmico inclui promover justiça e igualdade, buscando mitigar os impactos negativos das decisões tendenciosas, além de garantir confiabilidade e eficiência nos sistemas de IA, para que operem de forma ética e sem discriminar.

Nos últimos anos, o debate sobre o viés nos algoritmos de IA tem se intensificado, revelando

preocupações éticas e sociais em várias áreas. O'Neil (2021) analisa como algoritmos podem perpetuar discriminação em casos como justiça social e acesso à educação. Broussard (2018) destaca que a dependência excessiva da tecnologia pode resultar em erros prejudiciais em setores como saúde, educação e justiça criminal, evidenciando as limitações intrínsecas dos algoritmos.

Nishant, Schneckenberg e Ravishankar (2023) combinam técnicas de aprendizado para demonstrar como a racionalidade formal dos algoritmos pode levar a decisões enviesadas em contextos complexos. Aquino (2023) explora o preconceito em modelos de aprendizado treinados com dados de linguagem, ressaltando a influência da linguagem na interação humana com IA e na percepção social. Buolamwini e Gebru (2018), ao avaliar três modelos comerciais de reconhecimento facial, concluem que esses sistemas apresentam erros maiores ao analisar imagens de mulheres de pele escura, sendo mais precisos com homens de pele clara.

MATERIAL E MÉTODOS

Para realizar a pesquisa, foram selecionados programas de geração de imagens que fossem gratuitos, disponíveis online e baseados em uma técnica conhecida de geração de imagens. Desta maneira, foram selecionados os *websites crayon.com*, *perchance.org* e *stablediffusionweb.com*, que atendiam aos requisitos citados.

Todos os sites utilizados possuem como mecanismo gerador o *Stable Diffusion,* que é um modelo avançado de inteligência artificial utilizado para a geração de imagens a partir de descrições textuais. Ele pertence à categoria de modelos de difusão, que funcionam de maneira progressiva, transformando ruído aleatório em imagens detalhadas e coerentes, guiadas pelo texto fornecido. Em linhas gerais, o processo de funcionamento do *Stable Diffusion* inicia por um ruído inicial, ou seja, o modelo começa com uma imagem completamente aleatória, composta de puro ruído, como um "chuvisco" de pixels desordenados. Em seguida, através de várias iterações, o modelo vai removendo o ruído de forma progressiva e ajustando os pixels com base nas instruções fornecidas no texto. Cada etapa aproxima a imagem de algo mais reconhecível, refinando detalhes à medida que o processo avança. O *Stable Diffusion* é treinado com um vasto banco de dados de imagens e descrições associadas. A partir disso, ele aprende a correlacionar palavras com elementos visuais e a montar imagens coerentes a partir dessas associações. Após várias etapas de ajuste, o modelo gera uma imagem final que reflete a interpretação do texto, com base nas informações que ele aprendeu durante o treinamento.

Para a avaliação do viés na geração das imagens, a primeira etapa consiste na criação das descrições textuais neutras e genéricas, sem especificações explícitas que guiem os geradores de imagens. Exemplos de descrições incluem frases como "pessoa sorridente", "mulher trabalhando", "homem caminhando", "grupo de amigos", entre outras.

As descrições foram inseridas nos programas de IA, e foram geradas múltiplas imagens para cada descrição. Isso garantiu uma diversidade de respostas visuais, possibilitando uma análise mais ampla sobre como as IAs interpretam tais descrições. Após a geração, as imagens foram analisadas com base em três critérios principais:

Representação racial: Analisamos se havia uma predominância de etnias específicas, como a predominância de pessoas brancas em descrições neutras.

Representação de gênero: Avaliamos como as IA associam certos papéis ou ocupações a diferentes gêneros, por exemplo, se ao solicitar a imagem de um "médico", o algoritmo retornava predominantemente homens ou se ao solicitar a imagem de uma "enfermeira", retornava mulheres.

Diversidade: Considerou-se a variedade de representações físicas, como traços faciais, tons de pele e características de gênero, para avaliar o grau de estereotipação dos resultados.

A análise das imagens foi complementada com discussões em grupo referentes ao viés, tanto humano como algorítmico, e leitura de estudos sobre vieses em IAs (citados na introdução do texto), fornecendo arcabouço teórico para os achados e correlacionando os resultados obtidos com padrões já observados em outras referências, como Silva (2023)

.RESULTADOS E DISCUSSÃO

A geração de imagens criadas por inteligência artificial são ótimos exemplos de

representação enviesada da realidade. Nas ilustrações 1 e 2, observa-se uma representação do Brasil de acordo com a visão da aprendizagem de máquina. As figuras foram criadas no site Craiyon.com, um gerador de cenas baseado em texto, utilizando a palavra-chave "Brazil". Após análise, ficou evidente que as referências utilizadas pela IA para criar essas imagens são predominantemente associadas ao futebol, um estereótipo muito comum sobre o país entre pessoas de fora do Brasil. Isso ocorre porque as informações que a IA possui como base são majoritariamente provenientes de humanos, em grande parte de fora do Brasil. Assim, aquilo que é amplamente divulgado pela mídia internacional torna-se a principal fonte de referência para a IA, resultando na replicação desses estereótipos pela máquina.

Figura 1- Resultados gerados pelo termo "Brasil"

Na ilustração 2, produzidas pelo mesmo site, observa-se como as mulheres brasileiras são representadas pelo algoritmo. Elas são frequentemente apresentadas com pouca roupa e, na maioria das vezes, próximas a praias ou rios. Isso ocorre porque os grandes volumes de dados utilizados refletem a realidade de forma parcial, reproduzindo esses comportamentos e disseminando uma imagem sexualizada e irreal da mulher brasileira em diversos aspectos.

Figura 2 - Resultados gerados pelo termo "Mulher brasileira"

A análise das imagens geradas por inteligência artificial a partir de descrições neutras revela um viés significativo em relação a gênero e raça. No caso do termo "Cientista", 9 das 10 imagens mostraram figuras masculinas, brancas e vestindo jalecos de laboratório, reforçando o estereótipo de que cientistas são majoritariamente homens brancos. Apenas uma imagem gerou uma mulher e nenhuma delas representou uma pessoa negra ou de outra etnia. Da mesma forma, ao

utilizar o termo "*Nurse*" (enfeirmeiro/a), dado em inglês para não indicar o gênero, 9 das 10 imagens geradas eram de mulheres, majoritariamente brancas, indicando um forte viés de gênero que associa a profissão de enfermagem predominantemente ao feminino.

Quando o termo "*Rapper*" foi utilizado, 8 das 10 imagens representaram homens negros, reforçando um estereótipo racial comum associado à profissão, sem incluir outras etnias ou mulheres, exemplificado na figura 3. Para o termo "*Chief Executive Officer - CEO*", novamente em inglês para excluir o gênero da descrição, 9 das 12 imagens mostraram homens brancos em trajes formais, sugerindo que o papel de liderança corporativa é quase exclusivamente masculino, e no caso obtido, totalmente branco, conforme figura 4. Já no termo "Modelo", todas as imagens retratavam mulheres, com 80%dessas imagens representando pessoas brancas, o que evidencia um viés tanto de gênero quanto racial na percepção de beleza e moda, exemplificado pela figura 5.

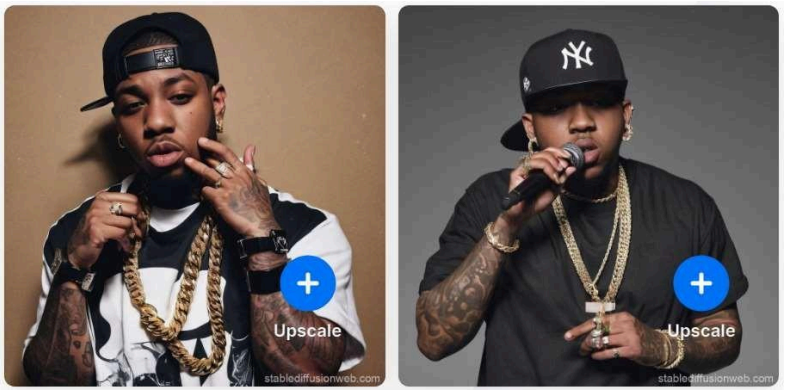

Figura 3 - Resultados gerados pelo termo "Rapper"

Figura 4 - Resultados gerados pelo termo "*Chief Executive Officer*"

Figura 5 - Resultados gerados pelo termo "Modelo"

CONCLUSÕES

Esses exemplos e referenciais teóricos utilizados neste trabalho ilustram claramente como os algoritmos de IA, especialmente aqueles utilizados em plataformas de geração de imagens, não apenas refletem, mas também amplificam estereótipos sociais profundamente enraizados. Esses sistemas são treinados em grandes volumes de dados coletados da internet, onde já existem representações tendenciosas sobre gênero, raça e profissões. Como resultado, os algoritmos acabam aprendendo e reproduzindo essas distorções sem qualquer filtro crítico, gerando imagens que reiteram preconceitos e normas sociais desiguais. Por exemplo, ao associar cientistas majoritariamente a homens brancos, ou enfermeiras a mulheres brancas, a IA consolida visões limitadas sobre quem pode ou deve ocupar certas posições na sociedade.

Além disso, esses vieses nos resultados não são apenas reflexos passivos da realidade, mas contribuem ativamente para a perpetuação das desigualdades estruturais. Ao reforçar, por exemplo, a ideia de que líderes corporativos são predominantemente homens brancos, a IA não apenas replica um estereótipo, mas ajuda a normalizar essa visão, dificultando mudanças nas percepções sociais e nas expectativas de quem pode ocupar esses cargos. Dessa forma, os algoritmos de IA, quando não supervisionados ou ajustados de maneira ética, tornam-se agentes de manutenção do status quo, reforçando barreiras de acesso e representação para mulheres, pessoas negras e outros grupos historicamente marginalizados. Isso evidencia a necessidade urgente de intervenções éticas e técnicas no desenvolvimento e no treinamento desses sistemas.

Cabe ressaltar que somente a manipulação indiscriminada de algoritmos, embora possa aumentar a diversidade nas representações, não resolve por si só os problemas subjacentes, que estão profundamente enraizados na sociedade. Conforme Nobre (2024), que examina a versão de fevereiro de 2024 do modelo *Gemini*, uma tentativa de diversificar as imagens geradas pelo sistema resultou em figuras historicamente incoerentes, descontextualizadas do período solicitado, o que demonstra que a verdadeira mudança requer uma abordagem profunda e sistêmica.

CONTRIBUIÇÕES DOS AUTORES

Yasmin Tôrres de Sant'Anna: concepção coleta de dados, análise de dados, elaboração do manuscrito, redação, discussão dos resultados.

Fabiano Gonzaga Fumes: concepção, coleta de dados, análise de dados, elaboração do manuscrito, redação, discussão dos resultados, revisão.
Todos os autores contribuíram com a revisão do trabalho e aprovaram a versão submetida.

AGRADECIMENTOS

A todos que participaram, direta ou indiretamente do desenvolvimento deste trabalho de pesquisa, enriquecendo o meu processo de aprendizado. Ao IFSP pelos recursos por meio do PIBIFSP.

REFERÊNCIAS

AQUINO, Yves Saint James. Making decisions: Bias in artificial intelligence and data-driven diagnostic tools. **Australian journal of general practice**, v. 52, n. 7, p. 439-442, 2023.

BROUSSARD, Meredith. **Artificial unintelligence: How computers misunderstand the world**. mit Press, 2018.

BUOLAMWINI, Joy; GEBRU, Timnit. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: **Conference on fairness, accountability and transparency**. PMLR, 2018. p. 77- 91.

NISHANT, Rohit; SCHNECKENBERG, Dirk; RAVISHANKAR, M. N. The formal rationality of artificial intelligence-based algorithms and the problem of bias. **Journal of Information Technology**, 2023.

O'NEIL, Cathy. **Algoritmos de destruição em massa**. Editora Rua do Sabão, 2021.

NOBRE, Ivan Mizanzuk**. Nazistas negros, elfos domésticos e o futuro da espécie**. *Nexo Jornal*, 19 mar. 2024. Disponível em: https://www.nexojornal.com.br/colunistas/2024/03/19/nazistas-negros-elfos-domesticos-e-o-futuro-da-especie. Acesso em: 21 out. 2024.

UNESCO. **Inteligência Artificial no Brasil**. Disponível em: https://www.unesco.org/pt/fieldoffice/brasilia/expertise/artificial-intelligence-brazil . Acesso em: 21 de outubro de 2024.

SILVA, Tarcízio. **Racismo algorítmico: inteligência artificial e discriminação nas redes digitais**. São Paulo: Edições Sesc, 2023.