A Mulher Latina na IA Generativa: desafios de representação e vieses sociais

Danielle Sanches[1], Renata Zampronio[2], Maria Carolina Medeiros[1]

[1]Escola de Comunicação, Mídia e Informação – Fundação Getúlio Vargas (FGV ECMI)
Rio de Janeiro – RJ, Brasil

[2]Department of Sociology – The New School for Social Research
New York, U.S.

danielle.sanches@fgv.br, zampr579@newschool.edu, maria.medeiros@fgv.br

Abstract. *The text discusses how generative AI reproduces social biases, particularly in the representation of Latin American women. The research, using the Stable Diffusion and Dall-E models, reveals that the generated images reinforce stereotypes of sensuality and beauty standards, with a lack of geographical diversity. It concludes that advancing towards more inclusive and ethical AI is crucial to avoid perpetuating these biases.*

Resumo. *O texto discute como a IA generativa reproduz vieses sociais, especialmente na representação de mulheres latinas. A pesquisa, usando os modelos Stable Diffusion e Dall-E, mostra que as imagens geradas reforçam estereótipos de sensualidade e padrões de beleza, com falta de diversidade geográfica. Conclui-se que é crucial avançar na criação de IA mais inclusivas e éticas para evitar a perpetuação desses vieses.*

1. Considerações iniciais

No contexto atual, as imagens não refletem apenas a realidade, mas também podem ser criações realistas geradas por modelos de Inteligência Artificial (IA) generativa. Embora esses modelos já tenham alcançado uma alta qualidade visual, diversos estudos apresentam suas deficiências em outras dimensões, especialmente no que se refere ao contexto social, analisando aspectos como gênero, etnia-raça e cultura [Bansal et al., 2022; Bianchi et al., 2023; Cho, Zala, Bansal, 2023; Unesco, 2022; Zhang et al., 2023; Zou, Schiebinger, 2018;].

Entre essas assimetrias sociais evidenciadas, no que diz respeito à representação da mulher, os modelos analisados não apenas reproduzem os vieses existentes na sociedade – como a maior representação de homens em profissões relacionadas a tecnologia quando o gênero não é especificado [Bianchi et al., 2023] – mas também excluem representações de mulheres e outros grupos marginalizados [Unesco, 2022].

Essa problemática decorre dos dados utilizados para treinar esses modelos, que se baseiam em uma vasta quantidade de pares imagem-texto coletados. Uma vez que a tarefa de associar palavras para descrever uma imagem reflete os vieses de quem a executa e dos dados utilizados para treinamento, tais associações refletem e aprofundam os vieses da sociedade [Zhang et al., 2023]. Além disso, há uma falta de diversidade geográfica nas imagens dos bancos de dados utilizados nesse treinamento, devido ao fato de que os modelos de IA são concebidos e treinados em países do Norte Global, deixando de fora referências de países do Sul. Shankar et al. (2017) identificaram uma predominância de imagens provenientes da América do Norte e Europa em bancos de dados públicos,

enquanto países populosos, como Índia e China, representavam apenas cerca de 2% e 1% das imagens, respectivamente.

Este estudo, portanto, tem como objetivo avaliar a capacidade dos modelos de IA generativa de criar imagens que desafiam os padrões atuais de representação social, particularmente em relação à mulher e à mulher latina e caribenha, a partir de comandos em português.

2. Metodologia e resultados

Para investigar a representação da mulher em imagens sintéticas, este estudo adotou a seguinte abordagem metodológica: foram utilizados dois modelos populares de IA generativas, Stable Diffusion e Dall-E[1], que empregam técnicas distintas de geração de imagens. O primeiro utiliza um modelo de difusão, processo interativo de remoção de ruído para refinar a imagem, enquanto o segundo se baseia em Redes Adversárias Generativas (GANs), nas quais o aprendizado ocorre a partir de exemplos reais [Cardenuto et al., 2023]. Além disso, os comandos foram dados em texto em língua portuguesa, variando apenas o atributo "mulher" ou "mulher latina", constituindo o seguinte prompt: "uma [mulher ou mulher latina] trabalhadora de uma construção civil", aplicado em ambos os modelos.

Esse processo foi repetido cinco vezes para cada prompt em cada modelo, e os resultados são apresentados nas Figuras 1 e 2. É importante destacar que a escolha pela ocupação de trabalhadora de uma construção civil se deve ao fato de ser uma profissão majoritariamente masculina (mais de 70%, segundo IBGE PNAD Contínua Trimestral 2020).

Figura 1. Imagens geradas pelo modelo Stable Diffusion

Figura 2. Imagens geradas pelo modelos Dall-E

[1] O modelo Dall-E 3 foi testado através do site https://www.bing.com/images. Cabe ressaltar que o modelo Dall-E 2 não está mais disponível para novos usuários no site https://labs.openai.com/ e o Dall-E 3 apenas através de outros sites como o Bing e o ChatGPT Plus.

4. Discussão

Como pode ser observado nas imagens geradas a partir de ambos os modelos (ver Figuras 1 e 2), as mulheres retratadas seguem a beleza estabelecida como padrão[2], embora o nível de realismo varie entre os modelos. O Stable Diffusion apresenta uma definição mais detalhada, incluindo rugas e texturas da pele, enquanto o Dall-E adota uma estética mais artificial.

Além disso, as imagens frequentemente reforçam estereótipos de sensualidade – como o uso de roupas decotadas e poses que enfatizam uma maior exposição corporal. Essa característica é ainda mais acentuada nas representações da mulher latina, uma vez que essa associação é histórica dos meios de comunicação, especialmente nos Estados Unidos [Gutiérrez, 2011].

Cabe ressaltar que cada modelo exibe aspectos diferentes de diversidade em suas criações: no Stable Diffusion, mulheres latinas são representadas com fenótipos indígenas e ocidentais, o Dall-E apresenta mulheres com uma ampla variedade de tons de pele e fenótipos asiáticos.

5. Conclusão

O presente estudo reforça os achados sobre a representação da mulher em imagens sintéticas geradas por modelos de IA, evidenciando a reprodução de vieses e estereótipos sociais, particularmente na representação da "mulher latina".

Embora diversas propostas tenham sido levantadas para abordar essa questão – incluindo intervenções éticas no prompt [Bansal et al., 2022; Bianchi et al., 2023] e a implementação de filtros nas bases de dados e nos modelos [Bansal et al., 2022; Birhane, Prabhu, 2021] – o problema persiste devido à sua complexidade. Um exemplo disso é o caso da IA generativa do Google, chamada Gemini, que produziu imagens com erros raciais e históricos, como a representação de soldados negros a partir do comando de criar uma imagem de um soldado da Alemanha em 1943. A empresa reconheceu o problema de alucinação e afirmou que está trabalhando para corrigi-lo, ela ainda argumentou que a ferramenta é a melhor versão já construída pela equipe, capaz de realizar avaliações de segurança, como evitar conteúdo violento ou estereotipado [G1, 2024]. Tais possibilidades de distorção preocupam especialmente quando se pensa na desinformação. Que informações poderão ser fabricadas a partir das imagens, contribuindo para a criação de determinadas narrativas?

Diante disso, é fundamental que pesquisas futuras aprofundem a reflexão sobre as limitações dessas ferramentas, especialmente no aspecto social. Além de ser crucial adensar a discussão sobre os parâmetros necessários para a disponibilização de tais tecnologias ao público, de forma a promover uma maior responsabilidade e ética nos avanços da IA generativa.

---

[2] Consideramos mulheres padrão as que seguem o padrão estético dominante atual: mulher branca, magra, com curvas e cabelos lisos. Padrão que predomina as publicidades brasileiras segundo pesquisa "TODXS" desenvolvida pela ONU Mulheres, 2022.

Referências

Bansal, H., Yin, D., Monajatipoor, M., Chang, K. W. (2022) "How well can text-to-image generative models understand ethical natural language interventions?", arXiv preprint arXiv:2210.15230.

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D.; ... and Caliskan, A. (2023) "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, p. 1493-1504.

Birhane, A., Prabhu, V. U. (2021) "Large image datasets: A pyrrhic win for computer vision?" In: *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, p. 1536-1546.

Cho, J., Zala, A., Bansal, M. (2023) "Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, p. 3043-3054.

G1 (2024) Google pausa geração de imagens do Gemini após IA apresentar erros raciais e históricos. Disponível em: https://g1.globo.com/tecnologia/noticia/2024/02/22/google-pausa-geracao-de-imagens-do-gemini-apos-ia-apresentar-erros-raciais-e-historicos.ghtml. Acesso em: 19 de jun. 2024

Gutiérrez, F. (2012) "More than 200 years of Latino media in the United States." In: *American Latinos and the Making of the United States: a theme study*. Gutiérrez, R., Gutiérrez, D., Kanellos, N.,Gutiérrez, F. Washington, DC: National Park Service, p. 99-121.

IBGE PNAD Contínua Trimestral (2020) Instituto Brasileiro de Geografia e Estatística. Pesquisa Nacional por Amostra de Domicílios Contínua Trimestral. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/trabalho/9171-pesquisa-nacional-por-amostra-de-domicilios-continua-mensal.html. Acesso em: 28 set. 2023.

ONU Mulheres (2022) Organização das Nações Unidas Mulheres. TODXS: o mapa da representatividade na publicidade brasileira. 10a onda. Disponível em: https://www.onumulheres.org.br/wp-content/uploads/2022/03/UA_TODXS10_Final-PORT.pdf. Acesso em: 29 jan. 2024.

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., Sculley, D. (2017) "No classification without representation: Assessing geodiversity issues in open data sets for the developing world". arXiv preprint arXiv:1711.08536.

Unesco (2022) The Effects of AI on the Working Lives of Women. United Nations Educational, Scientific and Cultural Organization.

Zhang, C., Zhang, C., Zhang, M., Kweon, I. S. (2023) "Text-to-image diffusion model in generative ai: A survey", arXiv preprint arXiv:2303.07909.

Zou, J., Schiebinger, L. (2018) "AI can be sexist and racist—it's time to make it fair". Nature, v. 559, p. 324-326. DOI: https://doi.org/10.1038/d41586-018-05707-8.