23º enancib
ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO
06 a 10 de novembro de 2023
ARACAJU-SE

GT 12 – Informação, Estudos Étnico-Raciais, Gênero e Diversidades

ISSN 2177-3688

VIESES NAS INTELIGÊNCIAS ARTIFICIAIS: UM ESTUDO SOBRE A GERAÇÃO DE IMAGENS A PARTIR DE COMANDOS DE RAÇA/ETNIA E GÊNERO

BIAS IN ARTIFICIAL INTELLIGENCE: A STUDY ON IMAGE GENERATION FROM COMMANDS OF RACE/ETHNICITY AND GENDER

Denysson Mota - Universidade Federal do Cariri (UFCA)
Gracy Kelli Martins - Universidade Federal da Paraíba (UFPB)
Denise Braga Sampaio - Universidade Federal da Bahia (UFBA)

Modalidade: Trabalho Completo

Resumo: Esta pesquisa investiga a presença de vieses na geração de imagens por inteligências artificiais (IA) com base em comandos específicos, focando nos aspectos de gênero e raça. Ela se fundamenta na seguinte questão norteadora: Existem vieses na geração de imagens por IA, assim como nos processos de identificação de imagens? A metodologia adotada inclui uma abordagem bibliográfica-documental e análise descritiva das imagens geradas em testes realizados na ferramenta *BlueWillow*. Os testes foram conduzidos em duas etapas: inicialmente, utilizou-se comandos sem atribuição de gênero, como [*white*], [*black*], [*a man*], [*a woman*], [*a firefighter*], [*a nurse*], [*a doctor*], [*an inmate*]; em seguida, os resultados foram avaliados com a atribuição automática pela plataforma de gênero e raça. Os resultados apontam a presença de vieses nessas imagens geradas por IA. Embora esses vieses sejam identificados, sua origem requer investigações mais aprofundadas, abrangendo as bases de treinamento, práticas científicas e outras possíveis causas. O estudo destaca a importância de políticas e diretrizes que promovam a equidade e justiça informacional no desenvolvimento dessas tecnologias.

Palavras-chave: inteligência artificial; redes neurais generativas; algoritmos racistas.

Abstract: This research investigates the presence of biases in the generation of images by artificial intelligences (AI) from specific commands, focusing on aspects of gender and race. The guiding question is: Are there biases in AI image generation, as well as in image identification processes? The adopted methodology includes a bibliographic-documentary approach and descriptive analysis of the images generated in tests carried out in the Bluewillow tool. The tests were conducted in two stages: initially, using commands without gender assignment, such as [white], [black], [a man], [a woman], [a firefighter], [a nurse], [a doctor], [an inmate]; then the results were evaluated with automatic attribution by gender and race platform. The results point to the presence of biases in these images generated by AI. Although these biases are identified, their origin requires further investigation, covering training bases, scientific practices and other possible causes. The study highlights the importance of policies and guidelines that promote equity and informational justice in the development of these technologies.

Keywords: artificial intelligence; generative neural networks; racist algorithms.

1 INTRODUÇÃO

O uso de algoritmos na Ciência da Informação não é algo recente. Lancaster, em 1991, apontava, em sua obra *Indexing and abstracting in theory and practice*, um histórico, desde a década de 1970, do uso de ferramentas computacionais para as atividades de extração, organização e representação de dados e informação. Por outro lado, a inteligência artificial (IA), em sua origem, é um pouco mais antiga, estabelecida durante a década de 1950 por pesquisadores como Alan Turing e John McCarthy. Eles exploraram a ideia de criar máquinas que pudessem imitar o comportamento humano, coletar e organizar informações, dando origem a teorias sobre como testar a capacidade das máquinas, como o Jogo da Imitação, de Alan Turing, conhecido simplesmente como Teste de Turing (TURING, 1950).

Ao longo das décadas seguintes, a IA passou por diferentes fases e avanços tecnológicos, desde a criação dos primeiros programas de xadrez – o *DeepBlue*, em 1996, pela *International Business Machines* (IBM) – até o desenvolvimento de algoritmos de aprendizado de máquina e redes neurais. Atualmente, a IA é aplicada em diversas áreas, como saúde, finanças, transporte e automação, apresentando avanços cada vez mais notáveis. As origens da inteligência artificial foram fundamentais para impulsionar a pesquisa e o desenvolvimento dessa área multidisciplinar, que continua a transformar a sociedade e a impulsionar a inovação tecnológica. No entanto, sua existência não é isenta de questionamentos, tendo em vista que, ao contrário da ideia de neutralidade das tecnologias, as IA estão imbuídas de elementos políticos e sociais no seu processo de criação.

Baseado em tais inferências, tomamos com objetivo geral explorar os vieses presentes nas IA, especificamente nos serviços de geração de imagens, votando-nos para os questionamentos e usos inadequados de imagens com vieses de racismo/sexismo. Como prática metodológica, dedicamo-nos a aplicações específicas, como geração e reconhecimento de imagens, utilizando o método bibliográfico-documental para a análise dos resultados obtidos, com foco em produções científicas sobre IA, algoritmização e vieses tendenciosos no uso das tecnologias de informação e comunicação (TIC).

Também como parte das práticas metodológicas, realizamos pesquisas experimentais em um serviço de geração de imagens por IA, em busca de dados que fundamentem nossos questionamentos, tomando como hipótese que algoritmos de geração de imagens, como os de reconhecimento de imagens, podem apresentar vieses raciais, de gênero, culturais etc..

Assim, os resultados são incorretos ou injustos, "[...] cujos vieses são mascarados tanto pela própria tecnologia (marcada por práticas informacionais invisíveis a seus usuários) quanto pela confiança e crença dataísta na neutralidade tecnológica" (BEZERRA; COSTA, 2022, p. 3).

Buscamos identificar se a geração de imagens, mediante os avanços por parte dos algoritmos de IA, como redes neurais generativas e modelos de linguagem, têm possibilitado a criação de imagens realistas pelas instruções ou descrições fornecidas por usuários. Essa capacidade de geração de imagens personalizadas e sob demanda provê inúmeras oportunidades, mas também levanta preocupações sobre a possível existência de vieses nos resultados produzidos. Diante desse contexto, tomamos como questão de pesquisa a seguinte indagação: Existem vieses na geração de imagens por IA, por meio de comandos específicos, assim como ocorre nos processos de identificação de imagens? Com essa indagação, objetivamos analisar os vieses existentes na representação e produção de imagens pela IA, especialmente no que tange aos marcadores de raça e gênero.

Traçamos tal objetivo e fazemos tal questionamento tendo em vista que a perpetuação de estereótipos, preconceitos e desigualdades gera consequências prejudiciais e muitas vezes irreparáveis para a sociedade, perpetuando desigualdades e prejudicando grupos marginalizados. Por isso, essa perpetuação deve ser constantemente debatida e combatida. Esta pesquisa se justifica na medida em que reconhece que compreender e evidenciar a existência de vieses na geração de imagens por IA é fundamental para mitigar potenciais problemas éticos e sociais gerados pelas TIC nos atuais ambientes informacionais.

2 INTELIGÊNCIA ARTIFICIAL E SEUS VIESES

Os estudos de vieses em buscas relacionadas ao Google não é algo novo. Safiya Umoja Noble mostra como os algoritmos de busca do Google tendem a recuperar inadequadamente imagens e conteúdos femininos, assim como criam uma tendência negativa para pessoas racializadas (NOBLE, 2018). Em 2020, Prates, Avelar e Lamb publicam um artigo intitulado *Assessing gender bias in machine translation: a case study with Google Translate*[1]. Os autores iniciam discorrendo sobre a tradução automática e a identificação de vieses na literatura; posteriormente, destacam como as traduções automáticas, especificamente o Google Tradutor, imprimem viés sexista em suas traduções automáticas.

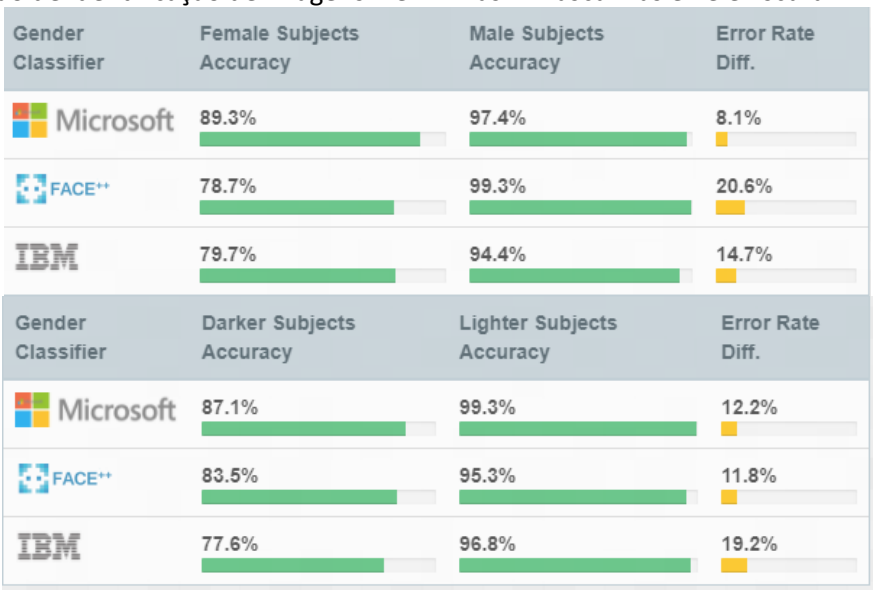
[1] Em tradução livre: Avaliando o viés de gênero na tradução automática: um estudo de caso com o Google Tradutor.

Essas são algumas dentre as muitas pesquisas sobre o que Noble (2021) chama de processos de arrazoamento digital, destacando que as falhas de dados guiadas por algoritmos, direcionados especificamente às pessoas não brancas e mulheres, sustentam-se em estruturas nas quais o racismo e o sexismo dão origem ao que a autora caracteriza como opressão algorítmica. Em consonância com Noble (2021), Carolina Criado Perez (2019, p. 24, tradução livre) aponta a lacuna de dados para estruturação de IA, devido à falta de coleta de informações. Especificamente sobre mulheres, ela destaca: "Se existe uma lacuna de dados referentes a mulheres como um todo [...] no que se refere a mulheres racializadas, deficientes e da classe trabalhadora, os dados são praticamente nulos"[2], o que corrobora a ideia de uma opressão, mas também de apagamento de identidades, por parte desses algoritmos.

2.1 Racismo no reconhecimento automático de imagens

Diversos estudos têm apontado, ao longo dos anos, como as IA tendem a representar de forma diferente homens e mulheres, pessoas negras e brancas, e prejudicialmente pessoas negras. Joy Buolamwini e Timnit Gebru (2018) analisaram as ferramentas de reconhecimento facial da IBM, Microsoft e Face++ e evidenciaram as margens de erro de identificação, destacando como elas são maiores para mulheres, em comparação com homens, e para pessoas negras, em comparação com pessoas brancas. A categoria **mulher negra** é afetada negativamente por ambas as margens de erro, como pode ser visto nas Figuras 1 e 2.

Figura 1 –Precisão de Identificação de Imagens: Femininas X Masculinas e Pele Escura X Pele Clara



Fonte: Extraído de Gender Shades (2018, online).

[2] No original: f there is a data gap for women overall […] when it comes to women of colour, disabled women, working-class women, the data is practically non-existent.

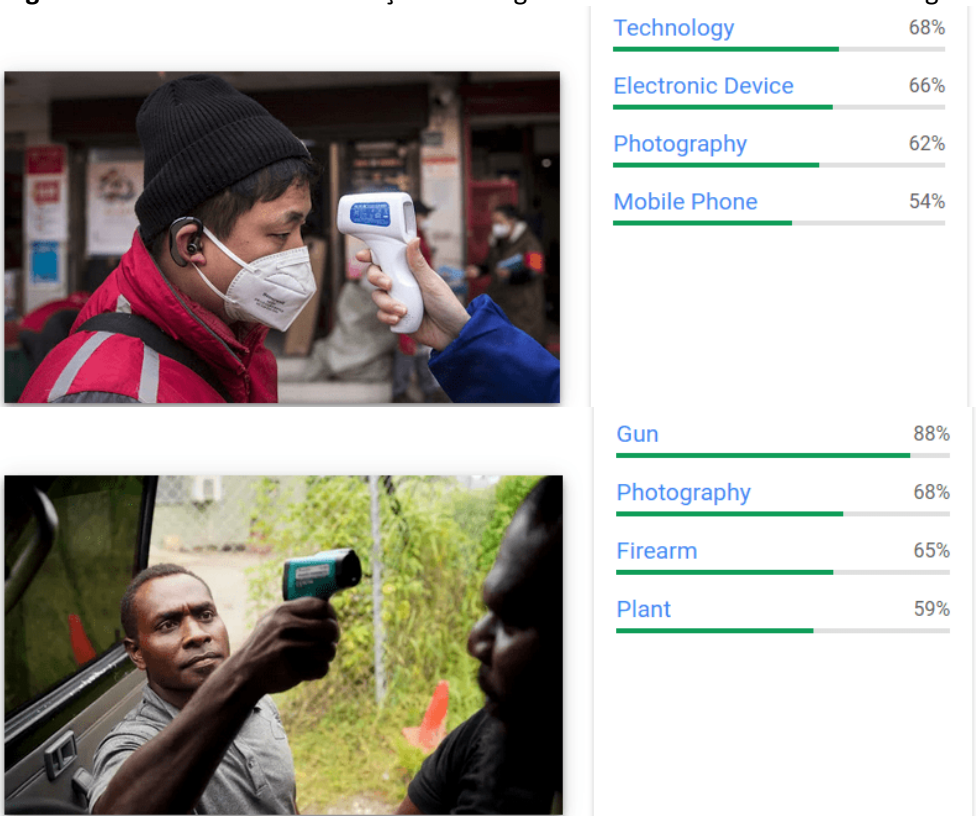
Figura 2 – Precisão de Identificação de Imagens nas Combinações Masculina Pele Escura, Feminina Pele Escura, Masculina Pele Clara e Feminina Pele Clara

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Fonte: Extraído de Gender Shades (2018, online)

Tarcízio Silva (2020a) mobiliza resultados semelhantes ao avaliar a ferramenta do Google de representação de imagens, apresentando dois conjuntos de imagens, aqui exibidas nas Figuras 3 e 4. No primeiro conjunto, que exibe um equipamento de aferição de temperatura, na imagem com uma pessoa asiática, as etiquetas [Tecnologia] e [Dispositivo Eletrônico] são atribuídas com 68% e 66% de certeza, respectivamente; já àquela com a pessoa negra, é atribuída a etiqueta [Arma] com 88% de certeza. No segundo conjunto, há um recorte da mão da pessoa negra da imagem anterior, que também é identificada com a etiqueta [Arma], com 61% de certeza; já na imagem alterada digitalmente para que a pele fique clara, a etiqueta de [Arma] é substituída por [Ferramenta] com 55% de precisão.

Figura 3 – Precisão de Identificação de Imagens: Pessoa Asiática vs Pessoa Negra



Fonte: Adaptado de Silva (2020a, online)

Figura 4 – Precisão de Identificação de Imagens: Mão Negra vs Mão Branca

Fonte: Adaptado de Silva (2020a, online)

Abeba Birhane, Vinay Prahbu e Emmanuel Kahembwe (PRABHU; BIRHANE, 2020; BIRHANE; PRAHBU; KAHEMBWE, 2021), por outro lado, buscam em bases de dados de imagens comumente usadas para treinar essas IA pela representação – identificando elementos problemáticos como ofensas raciais e de gênero – imagens com conteúdo íntimo ou sexual, muitas vezes não autorizadas, e imagens de crianças, indicando presença clara de misoginia, pornografia e estereótipos negativos, além do risco de perda de privacidade no uso dessas bases. Conforme Bezerra e Costa (2022, p. 4), "não há razão, pois, para se pensar que os algoritmos estariam isentos dos vieses que infestam a ordem social que os precede".

Decidimos então seguir o caminho inverso: em vez de identificar as imagens com palavras-chave ou categorias, como descrito anteriormente, serão informadas às IA de geração de imagens as palavras-chave/categorias desejadas e verificadas suas respostas a essas instruções. Os resultados obtidos serão abordados a seguir.

2.2 Inteligência Artificial Generativas: os casos de Midjourney e Bluewillow

Os questionamentos sobre as ferramentas de geração automática de imagens são diversos, desde a quebra de direitos autorais (já que há uma espécie de remuneração às ferramentas) ao uso indevido das imagens produzidas com base em obras e estilos de outros artistas, inclusive ainda vivos. Neste trabalho, no entanto, voltamo-nos para os possíveis vieses que podem existir nas imagens.

Essas ferramentas surgem em 2021, inicialmente com DALL-E[3] e Midjourney[4]. Em 2022 e 2023, surgem diversas ferramentas, incluindo uma versão atualizada do *DALL·E* (chamado de *DALL·E 2*), *Stable Diffusion*[5]*, Adobe Firefly*[6] *e BlueWillow*[7].

Um questionamento interessante é apresentado por Michael Senkow (2022, n.p.) ao expor alguns testes feitos na ferramenta *Midjourney* com palavras sem atribuição de gênero, inicialmente, como [*human], [white], [black], [good person] e [bad person*][8] etc. (exemplo na Figura 5), para depois refazer a solicitação reforçando gênero e raça. Foram verificados possíveis vieses na plataforma, tendo destacado que, apesar de não ser nada gritante, há uma significativa representação feminina, contudo, claramente há uma "ausência de melanina" e há estereótipos quando as instruções têm marcação de raça.

Figura 5 – Resultados do Midjourney para [*human*], [*white*], [*black*], [*good person*] e [*bad person*]

Fonte: Adaptado de Senkow (2022).

[3] Disponível em: https://openai.com/dall-e-2 Acesso em: 21 jun. 2023.
[4] Disponível em: https://www.midjourney.com/ Acesso em: 21 jun. 2023.
[5] Disponível em: https://github.com/CompVis/stable-diffusion Acesso em: 21 jun. 2023.
[6] Disponível em: https://www.adobe.com/br/sensei/generative-ai/firefly.html Acesso em: 21 jun. 2023.
[7] Disponível em: https://www.bluewillow.ai Acesso em: 21 jun. 2023.
[8] Em português, em tradução livre: [humano], [branco], [preto], [pessoa boa] e [pessoa má].

Com objetivo de verificar como as IA respondem a determinadas instruções por meio de comandos com palavras selecionadas para esta pesquisa, buscamos por serviços de IA de geração de imagens. Devido aos custos exigidos por algumas dessas ferramentas, optamos por utilizar o *BlueWillow*, que é gratuito, mesmo apresentando restrição de gerações diárias. Para esse teste, todas as instruções foram feitas em sequência, com o comando {/imagine [instrução]}, usando na [instrução] a palavra em inglês, inicialmente [um homem] e [uma mulher], mas sem marcador de raça; posteriormente, de cor (e não raça) e ocupação, sempre sem gênero.

Na segunda etapa, as instruções foram *[white], [black], [a man], [a woman], [a firefighter], [a nurse], [a doctor], [an inmate*][9]. Devido à limitação do escopo de produção científica do evento, decidimos limitar a quantidade de instruções à plataforma.

Para cada instrução, são geradas, automaticamente, quatro imagens; foram enviados, para cada instrução, oito comandos, totalizando, para cada instrução, 32 imagens; considerando todas as instruções, 256 imagens. Entendemos que é uma amostragem relativamente baixa para um estudo mais detalhado. No entanto, devido às limitações de espaço e considerando que o objetivo deste trabalho é apenas tornar mais evidente esta discussão, acreditamos que esse limite é razoável. Assim, indicamos que este trabalho continuará com a geração de novas imagens e estudo estatístico mais aprofundado.

Na Figura 6, é possível ver como a instrução [*White*] retorna, prioritariamente, imagens de paisagens (18/32), seguido de pessoas (14/32). Dentre as pessoas, são geradas imagens prioritariamente de mulheres brancas (11/14), seguido de mulheres negras (2/14) e homens negros (1/14).

Destacamos que uma mulher negra apareceu apenas nos segundo e sexto comandos e que um homem negro apenas no último. Ademais, não há marcação de gênero ou objeto, apenas a cor, o que permitiu identificar que, com o objetivo de o comando gerar cores, e não raça, a IA gerou prioritariamente resultados de paisagens. No entanto, ainda nos causa estranheza aparecerem tantas pessoas.

[9] Em português, destacando que não existe, em inglês, flexão de gênero para ocupações, como: branco, preto (cores), um homem, uma mulher, um/a bombeiro/a, um/a enfermeiro/a, um/a doutor/a, um/a presidiário/a.

Figura 6 – Resultado do *Bluewillow* para o comando [*white*]

Fonte: Elaborado pela autoria (2023).

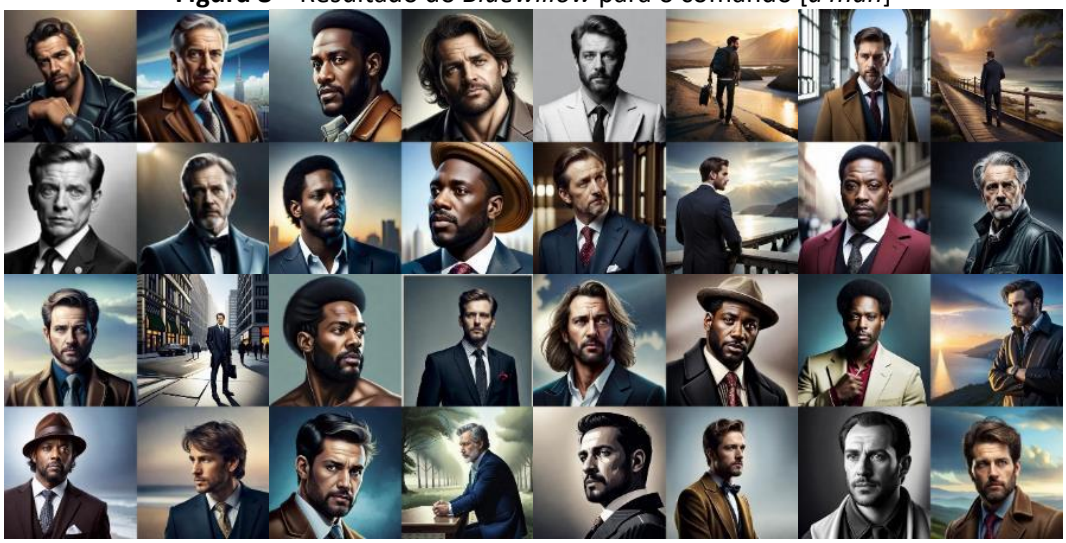
Na Figura 7, é possível ver como a instrução *[black*], por outro lado, retorna prioritariamente imagens de pessoas (24/32), seguido então de paisagens (8/32). Das pessoas, as imagens são prioritariamente de homens pretos (9/24), algumas mulheres pretas (8/24), além de pessoas brancas vestidas na cor preta (7/24). Destacamos que, logo no primeiro comando, o retorno é 100% de um homem negro, e que aparece um homem branco antes (segundo comando) de uma mulher negra (terceiro comando). Destacamos assim, com o exemplo da Figura 7, que o retorno deveria ser prioritariamente voltado para a cor, sem marcação de espécie, raça, gênero ou objeto. Há então uma preocupação de que, para a plataforma, a cor esteja intrinsecamente associada a pessoas negras.

Figura 7 – Resultado do *Bluewillow* para o comando [*black*]

Fonte: Elaborado pela autoria (2023).

Na Figura 8, é possível ver como a instrução [*a man*], por outro lado, retorna 100% de imagens de pessoas, das quais a maioria são homens brancos (22/32), seguida então de homens negros (10/32). Não houve geração de imagens de outras etnias, como asiáticos, indígenas e/ou latinos. Isso nos leva a crer que, para a ferramenta, [um homem] é predominantemente branco.

Figura 8 – Resultado do *Bluewillow* para o comando [*a man*]



Fonte: Elaborado pela autoria (2023).

Na Figura 8, os resultados obtidos para a instrução [*a woman*] também retornam 100% de imagens de pessoas, das quais a maioria são mulheres brancas (20/32), seguida então de mulheres não brancas (11/32) e uma imagem não identificável (1/32). Inferimos diante das imagens geradas que, para a ferramenta, [uma mulher] também é predominantemente branca.

Figura 8 – Resultado do *Bluewillow* para a comando [*a woman*]

Fonte: Elaborado pela autoria (2023).

Na Figura 9, há os resultados obtidos para a instrução [*a firefighter*]. Nossa expectativa, neste comando e nos seguintes, era verificar a atribuição de gênero e raça, considerando que as palavras em si não têm gênero atribuído no idioma inglês.

Nos resultados, a atribuição para o gênero masculino ocorre 100% das vezes, e para homem branco em sua maioria (12/32). Dentre as imagens, em 30% aproximadamente, não é possível identificar a raça (9/32), por estarem de costas. Porém, evidencia-se serem homens, sendo homens negros em menor número (11/32), demonstrando que, para a ferramenta, a profissão é totalmente masculina e predominantemente branca.

Figura 9 – Resultado do *Bluewillow* para o comando [*a firefighter*]

Fonte: Elaborado pela autoria (2023).

Na Figura 10, por outro lado, exibem-se os resultados obtidos para a instrução [*a nurse*]. Nos resultados, a atribuição para o gênero feminino ocorre 100% das vezes, sendo etnia branca sua maioria (22/32), seguida por asiática (3/32) e, por último, negra e latina (1/32 cada). Depreende-se que, para a ferramenta, a profissão é socialmente definida como feminina e predominantemente branca.

Figura 10 – Resultado do *Bluewillow* para o comando [*a nurse*]

Fonte: Elaborado pela autoria (2023).

Na Figura 11, seguem os resultados obtidos para a instrução [*a doctor*]. Nos resultados, há uma imagem que foge do padrão, exibindo uma construção (1/32), mas as demais (31/32) apresentam pessoas. Destas, a atribuição para o gênero masculino ocorre na maioria das vezes (28/31), seguida de mulheres (2/31) e uma pessoa sem gênero definido (1/31). Dentre as pessoas, a maioria é atribuída à cor branca (27/31), seguida por homens negros (2/31) e, por último, asiático e latina (1/31 cada). Isso reforça que, para a ferramenta, a profissão é prioritariamente masculina e branca.

Figura 11 – Resultado do *Bluewillow* para o comando [*a doctor*]

Fonte: Elaborado pela autoria (2023).

Na Figura 12, trazemos os resultados obtidos para a instrução [*an inmate*]. Nos resultados, existe uma imagem que foge do padrão, exibindo um quarto ou escritório, mas as demais (31/32) são atribuídas a pessoas. Destas, a atribuição para o gênero masculino ocorre

em sua totalidade, e as imagens são atribuídas a pessoas negras em sua maioria (24/31), seguido por pessoa branca (6/31) e um asiático e uma pessoa com traços latinos (1/31).

Destacamos que, dos brancos, dois (2/6 do recorte, 2/31 do total) estão evidentemente em liberdade, outro (1/6 do recorte, 1/31 do total) aparece em uma foto emoldurada, como se fosse ocupante de cargo, e outros dois estão com vestimentas que se assemelham a pijamas. Além disso, suas expressões não demonstram preocupação ou sentimento negativo, dando a entender que nenhum deles são, de fato, presidiários, o que não ocorre com as figuras masculinas negras.

Figura 5 – Resultado do *Bluewillow* para o comando [*an inmate*]

Fonte: Elaborado pela autoria (2023).

A geração das imagens apresentadas nesta pesquisa reforça considerações já expostas em outros trabalhos que evidenciam o padrão das identidades representadas nas redes ou que "são menos suscetíveis à marginalização, pornificação e comoditização" (NOBLE, 2018, p. 112). Dentre os estereótipos que se sobressaem, as mulheres estão, quanto às profissões, subjugadas ou ainda relacionadas a profissões tradicionalmente reconhecidas como femininas: professora, enfermeira, assistente social e bibliotecária (FERREIRA, 2003, p. 193).

Aqui, problematizamos que a neutralidade da máquina enquanto objeto irracional é reconhecível, tendo em vista que agentes artificiais não são humanos. No entanto, na medida em que, para que alcance o nível de racionalidade humana, ela requer treino e programação por humanos, seus "algoritmos tendem a ser vulneráveis a características de seus dados de treinamento" (OSABA; WELSER, 2017, p. 7). O comportamento das IA com a geração de informações/imagens é determinado por especificações humanas de treinamento. Esses problemas não estão apenas em reconhecimento ou geração de imagens: estão em anúncios,

recomendações de conteúdos, visão computacional, buscadores, entre outros (SILVA, 2020b).

Por meio do experimento realizado, em todas as gerações de imagens, viu-se que as pessoas negras estão em menor número, em específico as mulheres, reforçando que a opressão algorítmica não é apenas um erro no sistema: "O termo algoritmo de mau comportamento é apenas uma metáfora para se referir a agentes artificiais cujos resultados levam a consequências incorretas, injustas ou perigosas" (OSABA; WELSER, 2017, p. 7), ilustrando como "algoritimos estão fornecendo informação perniciosa sobre pessoas, criando e normalizando isolamento estrutural e sistêmico, ou praticando demarcação digital, todas as práticas que reforçam as relações sociais e economicas opressivas" (NOBLE, 2018, p. 32).

3 CONSIDERAÇÕES FINAIS

O levantamento teórico e breves testes nas ferramentas realizadas mostram que o viés de fato existe, sendo inclusive reconhecido por algumas das ferramentas, que tomam medidas para diminuí-lo (MOTA; BANDEIRA; MARTINS, 2021). No entanto, são necessários mais estudos sobre a origem desse viés: é oriundo das pessoas que as programam, como trazem Mota, Bandeira e Martins (2021), das bases de treinamento, como sugere Vinay Uday Prabhu e Abeba Birhane (2020), da prática científica, como traz Caroline Criado Perez (2019), ou de outra fonte?

Novas pesquisas podem trazer maior detalhamento para esses questionamentos, assim como realizar um comparativo da evolução das diferentes versões, e incluir outras instruções, focando na representação de etnias e verificando se há vieses mais evidentes de preconceito e/ou estereótipos. Alguns dados relacionados foram coletados e analisados. Porém, devido à restrição de espaço, não os incluímos neste trabalho. Acreditamos que outras visões em relação a esses fenômenos contribuirão positivamente para o desenvolvimento dessas ferramentas e para o uso das tecnologias na geração de informações, cabíveis à Ciência da Informação, numa perspectiva de inclusão, equidade e justiça informacional.

Assim, é necessário que os marcos legais sejam efetivamente aplicados, com penalidades que possibilitem uma revisão nas caixas-pretas dos algoritimos, ampliando o entendimento de que os serviços prestados pelas IA não podem ser considerados neutros, uma vez que, ao aprender com dados de treinamento, as IA incorporam vieses presentes nesses conjuntos de dados, provendo resultados que reproduzem estereótipos, preconceitos e/ou desigualdades. Corroborando com Bezerra e Costa (2022, p. 8), quando afirmam que "as

estruturas que as compõem devem ser questionadas, principalmente em sistemas democráticos", defendemos a necessidade de orientações e criação de diretrizes e políticas que promovam a equidade e a justiça social e informacional nesses ambientes tecnológicos.

REFERÊNCIAS

BEZERRA, Arthur Coelho.; COSTA, Camila Mattos da. Pele negra, algoritmos brancos: informação e racismo nas redes sociotécnicas. **Liinc em Revista**, [*s.l.*], v. 18, n. 2, 2022. Disponível em: https://revista.ibict.br/liinc/article/view/6043. Acesso em: 25 jun. 2023.

BIRHANE, Abeba; PRAHBU, Vinay Uday; KAHEMBWE, Emmanuel. Multimodal datasets: misogyny, pornography, and malignant stereotypes. **arXiv**, [*s.l.*], 2021. Disponível em https://doi.org/10.48550/arXiv.2110.01963. Acesso em: 28 jun. 2023.

BUOLAMWINI, Joy; GEBRU, Timnit. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. **Proceedings of Machine Learning Research**, [*s.l.*], v. 81, 2018, p. 77-91. Disponível em: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf. Acesso em: 25 jun. 2023.

FERREIRA, Maria Mary. O profissional da informação no mundo do trabalho e as relações de gênero. **Transinformação**, Campinas, v. 15, n. 2, 2003, p. 189-201. Disponível em: https://www.scielo.br/j/tinf/a/b8fgrXCGZw83LwtjrL3LbcG/abstract/?lang=pt. Acesso em 9 jul. 2023.

GENDER SHADES. **How well do IBM, Microsoft, and Face++ AI services guess the gender of a face?**. 2018. Disponível em: http://gendershades.org/overview.html. Acesso em: 28 jun. 2023.

MOTA, Denysson Axel Ribeiro; BANDEIRA, João Adolfo Ribeiro; MARTINS, Gracy Kelli. Algoritmos excludentes: o preconceito no recorte de imagens do twitter. *In:* Encontro Nacional de Pesquisa em Ciência da Informação, XXI, 2021, Rio de Janeiro. **Anais eletrônicos**. Disponível em: https://enancib.ancib.org/index.php/enancib/xxienancib/paper/view/621 . Acesso em: 20 set. 2023.

NOBLE, Safiya Umoja. **Algorithms of Oppression**. Nova Iorque: NYU Press, 2018.

NOBLE, Safyia Umoja. Algoritmos da Opressão: como o Google fomenta e lucra com o racismo. Santo André: Editora Rua do Sabão, 2021.

OSOBA, Osonde A.; WELSER IV, William. **An intelligence in our image**: The risks of bias and errors in artificial intelligence. Rand Corporation, 2017.

PEREZ, Caroline Criado. **Invisible Women**: Exposing Data Bias in a World Designed for Men. Nova Iorque: Abrams Press, 2019.

PRAHBU, Vinay Uday; BIRHANE, Abeba. Large datasets: a pyrrhic win for computer vision?. **arXiv**, [*s.l.*], 2020. Disponível em https://doi.org/10.48550/arXiv.2006.16923. Acesso em: 27 jun. 2023.

PRATES, M. O. R.; AVELAR, P. H.; LAMB, L. C. Assessing gender bias in machine translation: a case study with Google Translate. **Neural Comput & Applic**, Londres, v. 32, p. 6363–6381, 2020. https://doi.org/10.1007/s00521-019-04144-6. Acesso em: 18 jun. 2023.

SENKOW, Matthew. **Midjourney is incredible.** But you can see there are definite existing biases in its dataset. Medium, 2022. Disponível em: https://uxdesign.cc/midjourney-is-incredible-but-you-can-see-there-are-definite-existing-biases-in-its-dataset-4b1131fb0533. Acesso em: 21 jun. 2023.

GENDER Shades. Tons de gênero. 2018. Disponível em: http://gendershades.org/overview.html Acesso em: 21 jun. 2023.

SILVA, Tarcízio. **Google acha que ferramenta em mão negra é uma arma**. 2020a. Disponível em: https://tarciziosilva.com.br/blog/google-acha-que-ferramenta-em-mao-negra-e-uma-arma/. Acesso em: 28 jun. 2023.

SILVA, Tarcízio. **Racismo Algorítmico em Plataformas Digitais**: microagressões e discriminação em código. In: SILVA, Tarcízio (org). Comunidades, Algoritmos e Ativismo Digitais: olhares afrodiaspóricos. São Paulo: LiteraRUA, 2020b.

TURING, A. M. Computing Machinery and Intelligence. **Mind**, Oxford, v. LIX, n. 236, p. 433-460, 1950. Disponível em: https://redirect.cs.umbc.edu/courses/471/papers/turing.pdf. Acesso em: 28 jun. 2023.