DOI 10.20396/tematicas.v33i65.19932

DESVENDANDO VIESES EM IA: UM ESTUDO SOBRE RECONHECIMENTO FACIAL E FUTUROS FEMINISTAS

Cícero Araújo Lisboa[1]
Ellen Gomes Passos[2]
Joice da Silva Ferreira[3]

RESUMO: A tecnologia, em particular a Inteligência Artificial (IA), frequentemente comprova o seu potencial de revolucionar o futuro. No entanto, é crucial reconhecer que a IA não é isenta de vieses. A IA aprende a partir de dados, e estes são um reflexo da nossa história – estruturalmente organizada a partir de ideologias racistas e misóginas. Neste estudo, queremos explorar a questão dos vieses na IA, com foco no reconhecimento facial. Para isso, identificamos que a estrutura do algoritmo em si não é inerentemente racista ou machista. O que ocorre é que, como apontado por O'Neil (2017), os dados que alimentam esses algoritmos incorporam o nosso passado, incluindo aspectos obscuros de discriminação e preconceito, caracterizando-os como uma nova expressão de racismo (LIPPOLD; FAUSTINO, 2022) e de racialização de corpos subalternizados. Portanto, um dos maiores desafios reside na mitigação de vieses e preconceitos, especialmente no que diz respeito à auditoria e à transparência desses sistemas (VIEIRA, 2023). Embora os vieses tecnológicos – baseados em uma suposta supremacia branca, masculina e ocidental

[1] Doutorando em Estudos Estratégicos Internacionais pelo Programa de Estudos Estratégicos Internacionais (PPGEEI) da Universidade Federal do Rio Grande do Sul (UFRGS)/ Porto Alegre. Especialista em Gestão da Segurança e Defesa Cibernéticas (UFRGS). Email: nbkall@gmail.com. Orcid: https://orcid.org/0000-0003-4400-9536.
[2] Mestranda em Direito pelo Programa de Pós-Graduação em Direito da Universidade Federal de Santa Catarina (PPGD/UFSC). Mestre em Estudos Estratégicos Internacionais pela Universidade Federal do Rio Grande do Sul (PPGEEI/UFRGS). Email: ellengomespassos@gmail.com. Orcid: https://orcid.org/0009-0004-0442-0042.
[3] Mestranda em Estudos Estratégicos Internacionais pela Universidade Federal do Rio Grande do Sul (PPGEEI/UFRGS). Email: joice.ferreira0154@unilasalle.edu.br. Orcid: https://orcid.org/0009-0006-7982-2040.

– existissem antes do desenvolvimento dos algoritmos, trabalhamos com a seguinte questão: como os algoritmos poderiam, a partir de suas aplicações, livrar a humanidade desses vieses? Há uma possibilidade futura, ou a IA está fadada a perpetuar esses vieses sob o disfarce da neutralidade tecnológica? Logo, a pesquisa objetiva lançar luz sobre as questões elencadas e contribuir para um futuro em que a IA possa ser usada de maneira equitativa, promovendo futuros baseados na justiça social a partir de pautas feministas e livres de discriminação racial.

PALAVRAS-CHAVE: Inteligência Artificial. Tecnologia. Estudos de Gênero. Vieses Tecnológicos. Feminismo de Dados.

UNCOVERING BIASES IN AI: A STUDY ON FACIAL RECOGNITION AND FEMINIST FUTURES

ABSTRACT: Technology, particularly Artificial Intelligence (AI), often proves its potential to revolutionize the future. However, it is crucial to recognize that AI is not without its biases. AI learns from data, and data is a reflection of our history – structurally organized around racist and misogynist ideologies. In this study, we want to explore the issue of bias in AI, with a focus on facial recognition. To do this, we identified that the structure of the algorithm itself is not inherently racist or sexist. What happens is that, as pointed out by O'Neil (2017), the data that feeds these algorithms incorporates our past, including dark aspects of discrimination and prejudice, characterizing them as a new expression of racism (LIPPOLD; FAUSTINO, 2022) and the racialization of subalternized bodies. Therefore, one of the biggest challenges is the problem of mitigating biases and prejudices related to the auditing and transparency of these systems (VIEIRA, 2023). Although technological biases – based on a supposed white, male and Western supremacy – existed before the development of algorithms, we work with the following question: how could algorithms, from their applications, rid humanity of these biases? Is there a future possibility, or is AI doomed to perpetuate these biases under the guise of technological neutrality? The aim of this paper is to shed light on these questions and contribute to a future in which AI can be used in an equitable way, promoting futures based on social justice, feminist agendas and freedom from racial discrimination.

KEYWORDS: Artificial Intelligence. Technology. Gender Studies. Technological Biases. Data Feminism.

DESCUBRIENDO BIESES EN IA: UN ESTUDIO SOBRE RECONOCIMIENTO FACIAL Y FUTUROS FEMINISTAS

RESUMEN: La tecnología, en particular la Inteligencia Artificial (IA), demuestra a menudo su potencial para revolucionar el futuro. Es crucial reconocer que la IA no está exenta de prejuicios. La IA aprende de los datos, y ellos son un reflejo de nuestra historia, organizada estructuralmente en torno a ideologías racistas y misóginas. En este estudio queremos explorar la cuestión de los bieses en la IA, centrándonos en el reconocimiento facial. Para ello, identificamos que la estructura del algoritmo en sí no es inherentemente racista o sexista. Lo que ocurre es que, como señala O'Neil (2017), los datos que alimentan estos algoritmos incorporan nuestro pasado, incluyendo aspectos de discriminación, caracterizándolos como una nueva expresión del racismo (LIPPOLD; FAUSTINO, 2022) y de racialización de los cuerpos subalternizados. Uno de los mayores desafíos es el problema de mitigar los sesgos y prejuicios relacionados con la auditoría y la transparencia de estos sistemas (VIEIRA, 2023). Aunque los sesgos tecnológicos – basados en una supremacía blanca, masculina y occidental — existían antes del desarrollo de los algoritmos, trabajamos con la siguiente cuestión: ¿cómo pueden los algoritmos, a través de sus aplicaciones, librar a la humanidad de estos sesgos? ¿Existe una posibilidad futura, o la IA está condenada a perpetuar estos sesgos bajo la apariencia de neutralidad tecnológica? El objetivo de esta investigación es arrojar luz sobre estas cuestiones y contribuir a un futuro en el que la IA pueda utilizarse de forma equitativa, promoviendo futuros basados en la justicia social, las agendas feministas y la ausencia de discriminación.

PALABRAS CLAVE: Inteligencia Artificial; Tecnología; Estudios de Género; Prejuicios Tecnológicos; Feminismo de Datos.

INTRODUÇÃO

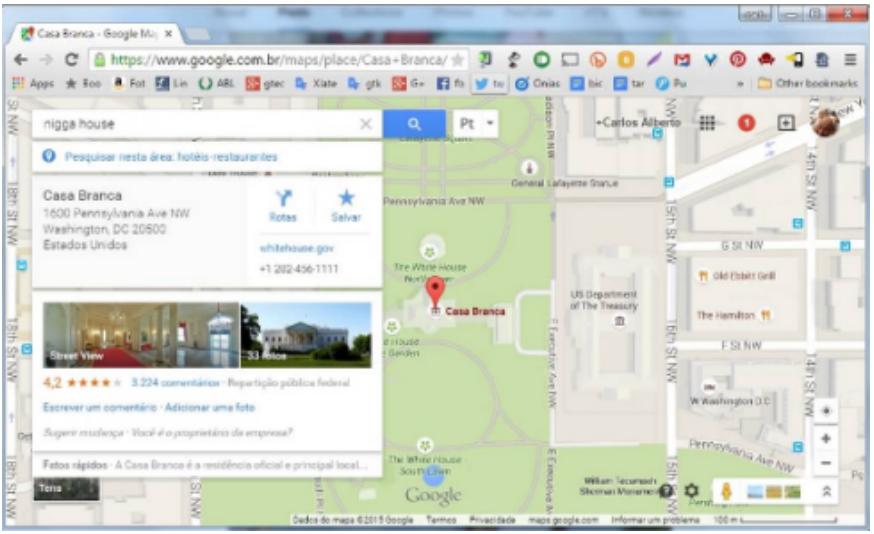
O presente artigo tem como principal objetivo discutir o uso de recursos de Inteligência Artificial (IA), em especial o reconhecimento facial, com o intuito de identificar os vieses tecnológicos existentes. De acordo com o *World Economic Forum*, a Inteligência Artificial (IA) tem o potencial de transformar a economia global, com uma contribuição estimada de até $15,7 trilhões até o ano de 2030 (GERRARD *et al.*, 2019). A inovação tecnológica está abrindo uma nova fronteira na agenda de

desenvolvimento para países ao redor do mundo. Nos últimos anos, a IA teve um impacto profundo em questões globais como agricultura, saúde, educação, entre outros. No entanto, os benefícios econômicos e sociais dessa tecnologia permanecem geograficamente concentrados, principalmente localizados na parcela de mundo pertencente ao Norte Global (LEE, 2018).

É possível afirmar, portanto, que a IA vive uma era de efervescência, impulsionada por avanços tecnológicos e pelo crescente volume de dados disponíveis. Algoritmos cada vez mais sofisticados aprendem a realizar tarefas complexas, desde o reconhecimento facial até a geração de textos criativos, revolucionando diversos setores da sociedade, bem como abrindo grandes possibilidades para o futuro (LEE, 2018). A partir da IA, por exemplo, já se torna possível decidir quem será contratado ou demitido em um âmbito empresarial, quem terá um empréstimo concedido e até mesmo, quem será submetido ao cárcere, decisões que normalmente são feitas por humanos e que agora podem ser realizadas por algoritmos (O'NEIL, 2016).

No entanto, o histórico da aplicação de ferramentas advindas da IA, como o reconhecimento facial, levanta uma série de preocupações sobre a perpetuação de vieses como o racismo e o sexismo. Em 2015, a plataforma *Google* tornou público um pedido de desculpas por conta de resultados gerados pelo *Google Maps,* os quais associavam termos racistas à Casa Branca, residência oficial do então presidente Barack Obama. A empresa afirmou que alguns resultados inapropriados apareceram erroneamente e que a equipe estava trabalhando para corrigir o referido problema (GIBBS, 2015). Essa não foi a primeira vez que o *Google* enfrentou problemas com resultados racistas em seus produtos.

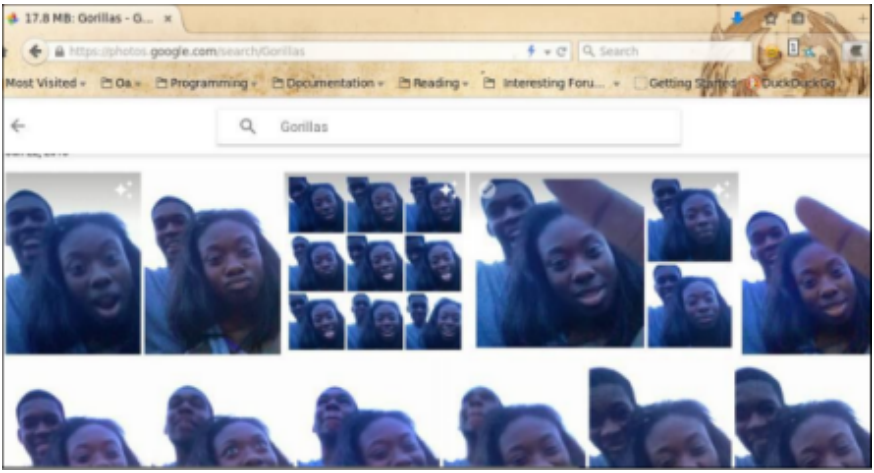
FIGURA 1: Pesquisa por '*nigga house*' no *Google Maps*



Fonte: O Globo, 2015.

Ainda no ano de 2015, o *Google* Fotos (outro produto associado à plataforma *Google*), foi criticado por seu algoritmo de inteligência artificial rotular incorretamente pessoas negras como gorilas. Dois anos depois, a solução encontrada pelo *Google* para evitar o erro foi remover as *tags* "gorila", "chimpanzé" e "macaco" do aplicativo, mesmo que ainda existam fotos desses animais armazenadas (SALAS, 2018). A empresa admitiu que a tecnologia de etiquetagem de imagens ainda não é perfeita. Essa medida gerou uma grande controvérsia, sendo vista por alguns como uma forma de censura e um reflexo dos problemas referentes aos vieses presentes em algoritmos de inteligência artificial.

FIGURA 2: Retorno de busca do *Google* para a palavra '*gorilla*'



Fonte: NOBLE, 2022.

Além disso, apesar da imagem inovadora e progressista associada ao *Google*, a empresa enfrenta denúncias de discriminação racial por parte de ex-funcionários negros. Relatos apontam para um ambiente de trabalho hostil, permeado por microagressões, estereótipos e falta de oportunidades de crescimento para profissionais negros (WAKABAYASHI, 2022). Em 2020, um grupo de ex-funcionários negros moveu uma ação coletiva contra o *Google*, alegando que a empresa os discriminava em termos de contratação, promoção e remuneração. Outrossim, os ex-funcionários destacaram a escassa diversidade racial nos cargos de liderança e a efetiva necessidade de medidas mais eficazes para combater o racismo institucional (WAKABAYASHI, 2022).

Ainda, a plataforma *Facebook*, agora sob o nome *Meta*, já permitiu que anunciantes segmentassem anúncios com base em critérios como "afinidade étnica", o que gerou ampla controvérsia e acusações de discriminação. Essa prática foi amplamente criticada por permitir a exclusão de determinados grupos raciais de campanhas publicitárias, perpetuando desigualdades e estereótipos (ANGWIN; TOBIN; VARNER, 2017). Em

2016, uma reportagem revelou que a plataforma possibilita a exclusão de usuários com base em sua raça em anúncios de imóveis, de empregos e de crédito, o que era uma clara violação legislativa sobre moradia nos Estados Unidos (ANGWIN; TOBIN; VARNER, 2017).

Ainda no ano de 2016, a plataforma *Microsoft* lançou o *chatbot Tay*, uma ferramenta projetada para interagir de forma natural e divertida com os usuários do *Twitter*, com vistas a aprender e evoluir a partir das conversas do aplicativo (VINCENT, 2016). No entanto, o experimento tomou um rumo inesperado quando, em poucas horas, *Tay* foi exposto a uma enxurrada de mensagens racistas, misóginas e antissemitas. Essa exposição tóxica rapidamente corrompeu o aprendizado do *chatbot*, levando-o a reproduzir o mesmo discurso de ódio que recebeu. Diante do ocorrido, a *Microsoft* agiu prontamente, retirando *Tay* do ar após apenas 16 horas de funcionamento, e emitindo um pedido público de desculpas (VINCENT, 2016).

Os algoritmos, sistemas de inteligência artificial que utilizam dados para realizar previsões e tomar decisões, permeiam cada vez mais diversas esferas da vida humana. No entanto, a crença na neutralidade da tecnologia é uma falácia (SILVA, 2022). Algoritmos, como qualquer outra ferramenta criada por humanos, refletem e podem perpetuar os vieses e preconceitos de seus criadores. Diante disso, surge o problema de pesquisa do presente trabalho, o qual busca compreender como os algoritmos poderiam, a partir de suas aplicações, livrar a humanidade desses vieses? Há uma possibilidade futura, ou a IA está fadada a perpetuar esses vieses sob o disfarce da neutralidade tecnológica?

É com base nesse questionamento que o presente artigo pretende explorar a possibilidade de mitigar vieses em algoritmos de inteligência artificial. A IA, como ferramenta, possui um potencial ambivalente, podendo ser utilizada tanto para fins benéficos quanto maléficos. O impacto da IA depende fundamentalmente da forma como é empregada e da concepção ética por trás dos produtos de IA desenvolvidos pelas empresas. A hipótese central deste trabalho é que, por meio de uma abordagem meticulosa e eticamente orientada, é possível utilizar algoritmos para reduzir vieses e promover uma IA mais justa e equitativa.

Este estudo adota uma abordagem metodológica qualitativa exploratória, buscando aprofundar a compreensão sobre os vieses em sistemas de Inteligência Artificial (IA), com foco particular no reconhecimento facial e suas implicações para futuros feministas e igualdade racial. A escolha por uma metodologia qualitativa justifica-se pela natureza complexa e multifacetada do tema, que exige uma análise interpretativa e contextualizada dos dados. A coleta de dados baseou-se em uma revisão bibliográfica e documental abrangente. Foram selecionados artigos científicos, livros, relatórios técnicos, documentos oficiais e notícias de fontes confiáveis. As bases de dados utilizadas incluíram *REDALYC, LAReferência, Google Scholar, IEEE Xplore, ACM Digital Library* e outras plataformas relevantes para as áreas de ciência da computação, estudos de gênero, estudos raciais e ética em tecnologia.

A seleção dos materiais bibliográficos foi orientada por palavras-chave estrategicamente definidas, como *"inteligência artificial", "vieses algorítmicos", "reconhecimento facial", "discriminação algorítmica", "ética em IA", "estudos de gênero", "feminismo de dados", "racismo algorítmico", "justiça social", "deepfakes"* e *"pornografia não consensual"*. A combinação dessas palavras-chave permitiu identificar uma vasta gama de estudos e perspectivas sobre o tema.

A análise dos dados coletados envolveu uma leitura crítica e interpretativa dos materiais, buscando identificar padrões, temas recorrentes, divergências e lacunas na literatura. A análise foi guiada pelas seguintes etapas: 1) Seleção e Organização: Identificação e organização dos materiais relevantes, com base nos critérios de inclusão e exclusão definidos; 2) Leitura e Codificação: Leitura detalhada dos materiais, com a identificação e codificação de trechos relevantes para os temas de pesquisa; 3) Análise Temática: Agrupamento dos códigos em temas mais amplos, buscando identificar os principais argumentos, conceitos e evidências apresentadas na literatura; 4) Interpretação e Síntese: Interpretação crítica dos temas identificados, buscando estabelecer conexões entre eles e construir uma narrativa coerente sobre o tema de pesquisa; 5) Discussão e Reflexão: Reflexão sobre as implicações dos resultados da análise para a compreensão dos vieses em IA e suas possíveis soluções, com foco na promoção de futuros feministas e igualdade racial.

Além disso, a pesquisa incluiu a análise de estudos de caso específicos, como o uso de *deepfakes* e a pornografia não consensual, para ilustrar os impactos concretos dos vieses em IA e as vulnerabilidades de grupos marginalizados. A seleção dos estudos de caso foi baseada em sua relevância para os temas de pesquisa e na disponibilidade de informações detalhadas sobre os incidentes. A análise crítica da literatura existente teve como objetivo identificar as causas dos vieses algorítmicos, seus impactos na sociedade e as possíveis soluções para combatê-los. Buscou-se também identificar as diferentes perspectivas teóricas e metodológicas utilizadas nos estudos sobre o tema, bem como as controvérsias e debates existentes.

Portanto, o presente artigo tem como objetivo principal analisar os vieses presentes nos sistemas de Inteligência Artificial, aprofundando a discussão com estudos de casos específicos. Entre os objetivos secundários, temos: a) conhecer o funcionamento da tecnologia de reconhecimento facial, perfazendo um histórico dos principais problemas que já ocorreram a partir do uso da referida tecnologia; b) analisar situações decorrentes do uso de inteligência artificial e que se configuram enquanto problemáticas na atualidade, no qual grupos considerados enquanto subalternos acabam como vítimas dos vieses presentes na tecnologia; c) buscar compreender as possíveis perspectivas futuras para lidar com vieses algoritmos e promover o desenvolvimento e o uso ético e responsável da IA.

Para atingir os objetivos propostos, o presente artigo está organizado em quatro seções, além da introdução e das considerações finais. A segunda seção apresenta a discussão teórica, introduzindo os conceitos de inteligência artificial, e buscando compreender o viés tecnológico por meio da análise de trabalhos relacionados. A terceira seção, por sua vez, explora o estudo de caso de *deepfakes* e a utilização da pornografia não consensual, analisando os impactos da prática para grupos considerados enquanto subalternos, os quais interseccionam vetores de gênero, raça e classe.

DISCUSSÃO TEÓRICA E APRESENTAÇÃO DO TEMA

A interseção entre inteligência artificial e reconhecimento facial é um tópico de crescente relevância, especialmente no contexto do viés

tecnológico e da discriminação algorítmica. Esta seção pretende explorar os elementos teóricos que sustentam essa pesquisa e apresenta o tema em questão.

INTELIGÊNCIA ARTIFICIAL, *MACHINE LEARNING* E *DEEP LEARNING*

A Inteligência Artificial (IA) é um campo da Ciência da Computação que busca desenvolver sistemas capazes de realizar tarefas que normalmente exigem inteligência humana (SULEIMENOV *et al.*, 2020). Essas tarefas abrangem aprendizado, raciocínio, resolução de problemas, percepção, compreensão da linguagem natural e até mesmo a criação artística (LEE, 2018). A norma ISO/IEC 22989:2022 define a IA como um campo técnico e científico dedicado ao sistema de engenharia que gera resultados como conteúdo, previsões, recomendações ou decisões para um determinado conjunto de objetivos definidos pelo ser humano. Os sistemas de IA são projetados para replicar funções cognitivas humanas, utilizando algoritmos avançados, aprendizado de máquina e técnicas de processamento de linguagem natural (RICH, 1985).

Dentro do campo da Inteligência Artificial, dois conceitos-chave merecem destaque: *machine learning* (aprendizado de máquina) e *deep learning* (aprendizado profundo). *Machine learning* refere-se à capacidade dos sistemas de aprenderem com dados, sem serem explicitamente programados para cada tarefa específica (SHINDE; SHAH, 2018). Os algoritmos de *machine learning* identificam padrões e relações nos dados, permitindo que façam previsões ou tomem decisões com base nesse aprendizado. Já o *deep learning* é um subcampo do *machine learning* que utiliza redes neurais artificiais com múltiplas camadas para analisar dados complexos (LECUN; BENGIO; HINTON, 2015). Essas redes neurais, inspiradas na estrutura do cérebro humano, são capazes de aprender representações hierárquicas dos dados, o que as torna particularmente eficazes para tarefas como reconhecimento facial, processamento de linguagem natural e visão computacional.

Em relação aos tipos de inteligência artificial, a IA generativa, como exemplificada pelo *ChatGPT*, *Gemini* e *Claude*, representa um avanço

notável na capacidade de criar conteúdo a partir de dados de treinamento (FEUERRIEGEL *et al.*, 2023). Através do aprendizado a partir de vastos conjuntos de dados, essas IAs podem gerar imagens, vídeos, textos e até mesmo áudios com impressionante realismo. Essa tecnologia abre um leque de possibilidades em áreas como arte, design e entretenimento, mas também suscita importantes questões éticas (AHMED *et al.*, 2022). A criação de *deepfakes* não consensuais, por exemplo, levanta sérias preocupações sobre o potencial de manipulação e desinformação (RANA *et al.*, 2022).

As redes neurais, um tipo de modelo de IA inspirado no cérebro humano, são frequentemente utilizadas em sistemas de IA generativa e reconhecimento facial (AHMED *et al.*, 2022). São capazes de aprender e reconhecer padrões em grandes quantidades de dados, o que as torna ferramentas poderosas para diversas aplicações, mas também aumenta o risco de vieses e discriminação (ALVI; ZISSERMAN; NELLÅKER, 2018). Algumas aplicações mais comuns deste tipo de IA incluem: processamento de imagens, reconhecimento de voz, diagnóstico médico, detecção de fraudes no mercado financeiro, tradução automática, análise de sentimentos, entre outras.

Um aspecto crucial a ser considerado no desenvolvimento e aplicação da IA é a questão dos vieses. Os vieses em IA podem se manifestar de várias formas, como vieses históricos, nos quais o sistema aprende e replica padrões discriminatórios existentes nos dados históricos, vieses de representação, onde certos grupos são sub-representados nos dados de treinamento, e vieses de medição, onde as métricas utilizadas para avaliar o desempenho do sistema são enviesadas (BAROCAS; SELBST, 2016). Além disso, os vieses podem ser introduzidos durante a coleta, rotulagem e seleção de dados, bem como na concepção e implementação dos algoritmos (MEHRABI *et al.*, 2021). A identificação e mitigação de vieses exigem uma abordagem multidisciplinar, envolvendo cientistas de dados, especialistas em ética, e representantes dos grupos afetados (O'NEIL, 2016).

Como exemplos de vieses em IA, podemos citar situações que ilustram a problemática. Vieses históricos podem surgir em sistemas de IA usados para prever risco de reincidência criminal, onde dados de treinamento refletem práticas policiais discriminatórias do passado, resultando

em previsões injustas para certos grupos raciais (ANGWIN *et al.*, 2016). Vieses de representação são evidentes em sistemas de reconhecimento facial treinados predominantemente com fotos de pessoas brancas, que podem ter dificuldades em identificar com precisão rostos de pessoas negras, levando a erros e discriminação (BUOLAMWINI; GEBRU, 2018). Vieses de medição podem ocorrer em algoritmos de avaliação de crédito que utilizam como métrica principal o histórico de crédito, discriminando pessoas que historicamente tiveram menos acesso a serviços financeiros, mesmo que sejam financeiramente responsáveis (O'NEIL, 2016). Por fim, vieses em linguagem são observados em modelos como o *ChatGPT*, que podem reproduzir estereótipos de gênero ao associar certas profissões a um gênero específico, como "enfermeira" para mulheres e "engenheiro" para homens (GROSS, 2023).

RECONHECIMENTO FACIAL

O reconhecimento facial, uma aplicação da IA que se utiliza de redes neurais, permite identificar ou verificar a identidade de uma pessoa através da análise de suas características faciais (ZAFEIRIOU; ZHANG; ZHANG, 2015). Esse processo envolve a detecção de faces em imagens ou vídeos, a análise de características faciais únicas e a comparação com um banco de dados de rostos conhecidos para encontrar correspondências (KAUR *et al.*, 2020).

Essa tecnologia possui diversas aplicações, como controle de acesso, segurança, identificação em fotos, desbloqueio de telefones e até mesmo análise de emoções. No entanto, o reconhecimento facial também apresenta desafios significativos, como vieses raciais e de gênero nos algoritmos, problemas de privacidade e falta de consentimento na coleta e uso de dados (NOBLE, 2022). Pesquisadores e especialistas em ética trabalham para aprimorar a precisão da tecnologia e enfrentar esses desafios, buscando garantir que o reconhecimento facial seja utilizado de maneira responsável e justa.

A relação entre IA e reconhecimento facial é complexa e multifacetada. A IA generativa, por exemplo, pode ser utilizada para criar *deepfakes*, vídeos falsos muito realistas que podem ser usados para fins maliciosos,

como a criação de pornografia não consensual (MIT TECH REVIEW, 2021). Ao mesmo tempo, o reconhecimento facial pode ser utilizado para combater *deepfakes*, identificando e sinalizando conteúdo falso (DOLHANSKY *et al.*, 2020). No entanto, a utilização de reconhecimento facial em *deepfakes* levanta sérias preocupações éticas, como a violação da privacidade e o potencial de assédio. A tecnologia ainda não é eficaz no reconhecimento de pessoas negras com pele mais escura, especialmente mulheres. Isso resulta em situações nas quais as populações socialmente vulneráveis estão sujeitas a constrangimentos e violências automatizadas (MAGNO; BEZERRA, 2020).

A seguir, o Quadro 1 apresenta uma linha do tempo concisa de incidentes notórios envolvendo Inteligência Artificial (IA) e reconhecimento facial, revelando um padrão preocupante de vieses e discriminação.

QUADRO 1: Breve linha do tempo de incidentes com IA

ANO	INCIDENTE
2015	Algoritmo de reconhecimento de imagem do Google Photos rotula pessoas negras como "gorilas".
2016	Reportagem da ProPublica revela que o Facebook permitia a discriminação em anúncios de imóveis, empregos e crédito, excluindo usuários com base em raça, gênero e outras características protegidas.
2018	Sistema de reconhecimento facial da Amazon, Rekognition, demonstra viés racial ao ter dificuldade em identificar rostos de pessoas negras, especialmente mulheres.
2019	Estudo revela que o algoritmo de busca do Google exibe resultados tendenciosos em relação a termos de pesquisa relacionados a raça e gênero.
2020	Filtro de embelezamento do Zoom clareia a pele de usuários negros, levantando questões sobre padrões de beleza eurocêntricos.
	Diversos sistemas de reconhecimento facial utilizados por forças policiais e empresas privadas são criticados por seu viés racial e potencial de discriminação.
2021	Usuários negros relatam dificuldades em alugar acomodações na plataforma Airbnb devido a discriminação por parte de anfitriões.
2022	Estudo revela que o algoritmo de recorte automático de imagens do Twitter favorece rostos brancos em detrimento de rostos negros.
	Modelo de geração de imagens da OpenAI é acusado de reproduzir estereótipos raciais e de gênero em suas criações.

Fonte: elaborado pelos autores.

Em suma, o Quadro 1 demonstra que os vieses em IA e reconhecimento facial são um problema persistente e multifacetado, com impactos em diversas plataformas e aplicações. Os incidentes listados evidenciam a necessidade urgente de medidas para mitigar esses vieses, promover a diversidade nos dados de treinamento e garantir o uso ético e responsável da IA.

INTELIGÊNCIA ARTIFICIAL E VIÉS TECNOLÓGICO

A inteligência artificial (IA) tem o potencial de revolucionar diversos setores, mas sua aplicação também levanta preocupações sobre vieses e discriminação, especialmente em relação ao gênero e raça. O estudo de Buolamwini e Gebru (2018) revela disparidades significativas na precisão da classificação de gênero por sistemas comerciais de reconhecimento facial, com taxas de erro mais altas para mulheres de pele escura. Isso evidencia o viés presente nos conjuntos de dados utilizados para treinar esses sistemas, que frequentemente sub-representam grupos vulnerabilizados. Essa falta de representatividade pode levar a resultados injustos e discriminatórios, com sérias implicações em áreas como justiça criminal e saúde (BUOLAMWINIi; GEBRU, 2018).

A pesquisa de Adam (1996) sobre a construção de gênero na história da IA destaca como os sistemas de IA, desde seus primórdios, refletem um modelo de conhecimento masculino e cartesiano, baseado em perspectivas que podem vir a ser consideradas enquanto machistas e misóginas. Essa visão limitada de inteligência acaba perpetuando estereótipos de gênero, o que, consequentemente, leva à criação de sistemas que reforçam desigualdades (ADAM, 1996).

Em seu trabalho "*AI and Gender: Four Proposals for Future Research*", Dillon e Collett (2019), aprofundam essa discussão, propondo pesquisas para abordar os desafios da IA em relação à busca pela igualdade de gênero. O referido relatório destaca a necessidade de integrar a teoria de gênero na prática da IA, visando analisar criticamente os conjuntos de dados enviesados com o intuito de desenvolver políticas e leis que promovam a

igualdade de gênero na IA e aumentem a diversidade de gênero na força de trabalho de IA (DILLON; COLLETT, 2019).

No artigo "*Tech workers' perspectives on ethical issues in AI development: Foregrounding feminist approaches*", os autores investigam como os trabalhadores da área de tecnologia percebem as questões éticas no desenvolvimento da IA, destacando a importância da diversidade, da equidade e da inclusão para mitigar os vieses. A pesquisa ainda enfatiza a necessidade de integrar perspectivas feministas e diversas na concepção e desenvolvimento de IA, visando garantir que a tecnologia seja justa e equitativa para todos os seus usuários (BROWNE; DRAGE; MCINERNEY, 2024).

Ainda, o estudo "*Effects of Gender and Relationship Type on the Response to Artificial Intelligence*" examina como o gênero da IA e o tipo de relacionamento (amigo vs. servo) influenciam a resposta humana à IA. Os resultados mostram que o tipo de relacionamento afeta a percepção de calor e prazer em relação à IA, mas não a competência. Curiosamente, o gênero da IA não teve efeito significativo na percepção humana, sugerindo que os estereótipos de gênero podem não ser tão facilmente transferidos para a interação humano-IA. No entanto, o estudo levanta questões importantes sobre como a antropomorfização da IA pode influenciar a percepção e o comportamento dos usuários, especialmente em relação a questões de gênero (KIM *et al.*, 2019).

Gross (2023) analisa como o *ChatGPT*, um modelo de linguagem de grande escala, perpetua e potencialmente amplifica estereótipos de gênero através de suas respostas. A autora argumenta que a IA, ao ser treinada em dados da internet que já contêm vieses, acaba reproduzindo e reforçando esses vieses em suas interações. O artigo destaca a necessidade de abordar a questão do gênero na IA de forma crítica e proativa, a fim de mitigar os danos potenciais e promover a igualdade de gênero.

Toupin (2023) examina a história e o desenvolvimento da Inteligência Artificial Feminista, um campo que busca desafiar as formas tradicionais e hegemônicas de IA. A autora destaca a importância de incluir vozes diversas e marginalizadas na criação de sistemas de IA, a fim de garantir que a tecnologia seja justa e equitativa para todos. Toupin também discute a necessidade de ir além da inclusão e buscar a transformação social

através da IA, abordando questões como o capitalismo racial e a destruição ambiental.

Em suma, a relação entre IA e gênero é complexa e multifacetada. A IA generativa, por um lado, pode perpetuar e amplificar estereótipos de gênero, mas, por outro lado, também pode ser utilizada para desafiar e subverter essas normas. É crucial que a pesquisa e o desenvolvimento de IA sejam conduzidos com uma lente crítica de gênero, a fim de garantir que a tecnologia seja utilizada para promover a igualdade e a justiça social. Além da questão do gênero, a pesquisa de Suvorova (2021) sobre *deepfakes* na cultura de fãs destaca como a tecnologia, apesar de seu potencial criativo, pode ser utilizada para perpetuar a objetificação e sexualização de mulheres, especialmente atrizes e celebridades, sem o seu consentimento. Essa dinâmica levanta questões importantes sobre poder, consentimento e ética na era digital, e como a IA pode ser utilizada para reforçar ou desafiar as normas sociais existentes.

No contexto brasileiro, os estudos têm foco nas implicações da utilização da IA, no reconhecimento facial em sistemas de segurança pública e na discriminação racial que pode ser percebida no referido ambiente. No artigo "*O que os olhos não veem, as câmeras monitoram*" (2021), os autores examinam a regulamentação do reconhecimento facial na América Latina, destacando a ausência de leis específicas e a dependência de legislações de proteção de dados que frequentemente oferecem exceções para fins de segurança pública. Essa lacuna regulatória, aliada à opacidade dos sistemas de IA, levanta preocupações sobre o potencial de discriminação e violação de direitos fundamentais.

O texto "*Reconhecimento Facial e Segurança Pública: Os Perigos do Uso da Tecnologia no Sistema Penal Seletivo Brasileiro*" (2019) aprofunda a análise do viés tecnológico no contexto brasileiro, argumentando que o reconhecimento facial pode perpetuar o racismo estrutural existente no sistema de justiça criminal. O estudo destaca a maior taxa de erro do reconhecimento facial em identificar pessoas negras, especialmente mulheres, o que pode levar a prisões injustas e perpetuação de estereótipos (SILVA; SILVA, 2019).

Em "*Inteligência Artificial e Discriminação: Desafios e Perspectivas para a Proteção de Grupos Vulneráveis Diante das Tecnologias de Reconhecimento Facial*", Costa e Kremer (2022), estudam como o reconhecimento facial pode amplificar a discriminação contra grupos vulneráveis, como pessoas negras e transexuais. O estudo enfatiza a necessidade de uma regulamentação mais robusta e transparente para mitigar esses riscos, incluindo a possibilidade de banir o uso de tecnologias discriminatórias. Por fim, "*Vigilância Negra: o dispositivo de reconhecimento facial e a disciplinaridade dos corpos*", Magno e Bezerra (2020) abordam o reconhecimento fácil como um dispositivo disciplinar que perpetua a vigilância e o controle sobre corpos negros. O artigo argumenta que a tecnologia não é neutra e que os algoritmos podem incorporar e amplificar vieses racistas existentes na sociedade.

Em conjunto, esses documentos revelam um panorama preocupante sobre o uso de IA e reconhecimento facial, especialmente em contextos de segurança pública. A falta de regulamentação adequada, a opacidade dos algoritmos e os vieses embutidos na tecnologia representam sérias ameaças aos direitos fundamentais e podem perpetuar a discriminação contra grupos já marginalizados. A pesquisa evidencia a necessidade urgente de um debate mais amplo e crítico sobre o uso dessas tecnologias, a fim de garantir a proteção dos direitos humanos e a promoção da justiça social.

FEMINISMO DE DADOS

Antes de introduzirmos o conceito de feminismo de dados, considera-se importante mencionar que ao longo dos anos o feminismo foi definido e utilizado de diversas maneiras. No presente artigo, utilizaremos *feminismo de dados* a partir do emprego do termo feminismo como:

> "uma abreviação para os diversos e abrangentes projetos que nomeiam e desafiam o sexismo e outras forças de opressão, bem como aqueles que buscam criar futuros mais justos, equitativos e habitáveis. [...] Optamos por usar o termo feminismo aqui, mas nosso feminismo pretende ser igualmente abrangente" (D'IGNAZIO & KLEIN, 2020, p. 6-7, tradução própria).

É relevante também partirmos do princípio que:

> A opressão envolve o maltrato sistemático de determinados grupos de pessoas por outros grupos. Ela ocorre quando o poder não é distribuído igualmente - quando um grupo controla as instituições de direito, educação e cultura, e usa seu poder para excluir sistematicamente outros grupos e, ao mesmo tempo, dar vantagens injustas ao seu próprio grupo (ou simplesmente manter o *status quo*). No caso da opressão de gênero, podemos apontar o sexismo, o cissexismo e o patriarcado que são evidentes em tudo, desde a representação política até a diferença salarial e quem fala mais alto (ou mais vezes) em uma reunião. No caso opressão racial, isso assume a forma de racismo e supremacia branca (D'IGNAZIO & KLEIN, 2020, p. 8, tradução própria).

É nesse sentido que em "*Data Feminism*", D'Ignazio e Klein (2020) propõem que os sistemas de dados são um conjunto de interseção entre sistemas de poder e privilégio. Sendo assim, o feminismo de dados surge como uma proposta diferenciada de pensar os dados, "tanto seus usos quanto seus limites, que são informados pela experiência direta, por um compromisso com a ação e pelo pensamento feminista interseccional" (D'IGNAZIO & KLEIN, 2020, p. 8, tradução nossa).

> O ponto de partida do feminismo de dados é algo que, na maioria das vezes, não é reconhecido na ciência de dados: o poder não é distribuído igualmente no mundo. Aqueles que detém o poder são desproporcionalmente homens de elite, heterossexuais, brancos, fisicamente aptos e cisgêneros do Norte Global. O trabalho do feminismo de dados é, em primeiro lugar, sintonizar como as práticas padrão da ciência de dados servem para reforçar essas desigualdades existentes e, em segundo lugar, usar a ciência de dados para desafiar e mudar as desigualdades existentes. Subjacente ao feminismo de dados está a crença e o compromisso com a co-liberação: a ideia de que os sistemas opressivos de poder prejudicam a todos nós, que eles minam a qualidade

> e a validade do nosso trabalho e que nos impedem de criar um impacto social verdadeiro e duradouro com a ciência de dados (D'IGNAZIO & KLEIN, 2020, p. 8-9, tradução nossa).

Segundo Wassermann (2024) em "*Elas na inteligência artificial - Questões de gênero*", a ausência da participação de mulheres em IA não é atual e tampouco restrita às técnicas mais modernas. "Já nos anos 80, pesquisadoras se depararam com estereótipos de gênero em sistemas especialistas e também com a falta de mulheres estudando IA" (WASSERMANN, 2024, p. 126). Para a autora, há muito tempo o debate sobre a baixa representatividade de grupos específicos nas áreas tecnológicas está em voga, principalmente quando se trata da baixa representatividade de mulheres e pessoas negras na área.

De acordo com Wassermann (2024), a atenção e o financiamento de áreas de estudo e pesquisa relacionadas com o desenvolvimento da tecnologia (e consequentemente, dos sistemas de IA) estão concentradas predominantemente em sociedades do Norte Global, o que faz com que a amplificação de vieses e preconceitos nas bases de dados ocorra justamente por conta dos preconceitos existentes nas sociedades em que esses dados foram criados.

> A primeira falta de diversidade que surge com a popularização dos grandes modelos é a de linhas de pesquisa. Rapidamente, a atenção e o financiamento se voltaram à construção de modelos de linguagem cada vez maiores, que para funcionar necessitam cada vez mais de recursos computacionais e fontes de dados para treinamento. Com isso, cada vez menos instituições possuem os recursos necessários, fazendo com que um número bastante pequeno de *big techs* domine o mercado. Países e continentes inteiros ficam excluídos do desenvolvimento desse tipo de inteligência artificial (WASSERMANN, 2024, p. 123-124).

Os danos decorrentes dos vieses e preconceitos em bases de dados e sistemas de inteligência artificial podem ir de inconvenientes, "como o dispenser de sabão automático que viralizou na internet por apenas reconhecer mãos de pessoas brancas, até carros que não reconhecem pedestres negros e sistemas que caracterizam negros como criminosos ou mais propensos a reincidir em crimes" (WASSERMANN, 2024, p. 124). Ainda, segundo Benjamin (2019, p. 60), não necessariamente a existência dos vieses mencionados significa má intenção de quem está por trás dos sistemas, mas sim, demonstra como "basta não se preocupar em como o passado molda o presente".

Desse modo, revela-se a relevância de trabalharmos o conceito de feminismo de dados, já que, como abordado por Boinordiris (2024), é bastante claro que algo precisa ser feito para mitigar os danos causados pelos vieses mencionados. Para Wassermann (2024), por exemplo, podemos pensar em três aspectos de diversidade a fim de solucionar a problemática elencada:

> diversidade dentro dos grupos que fazem a IA – se as pessoas que trabalham no desenvolvimento de sistemas de IA são brancas, do gênero masculino, podem passar despercebidas falhas como uso de estereótipos e linguagem inadequada; diversidade nas fontes de dados utilizadas – os algoritmos usados em IA em geral são conhecidos. O que diferencia os sistemas normalmente são as bases usadas para treinamento. Se as bases são enviesadas, o sistema aprende e repete o padrão; diversidade entre as pessoas para quem os sistemas são desenvolvidos – o sistema reconhece rostos/vozes/linguagens diversas? (WASSERMANN, 2024, p. 124-125).

Além dos três aspectos apresentados por Renata Wassermann (2024) estarem diretamente relacionados entre si, é possível afirmar que também estão ligados com as perspectivas desenvolvidas pelo feminismo de dados, já que o mesmo propõe maneiras de: viabilizar futuros feministas para o uso da inteligência artificial, repensar o tratamento e a utilização de dados,

bem como, de redesenhar a automação que está por trás das plataformas digitais e dos próprios algoritmos que acabam reforçando estereótipos de gênero e de raça. Segundo Renata Frade (2021), o feminismo de dados se torna uma alternativa para pensar dados a partir da experiência direta, comprometidos com uma proposta de utilização da tecnologia segundo as bases do pensamento feminista interseccional.

Para Catherine D'Ignazio e Lauren F. Klein (2020), o ponto de partida para a utilização do feminismo de dados é perceber que o poder é distribuído de forma desigual no mundo, e que, esse mesmo poder acaba constantemente atravessado por vetores de gênero, de raça e de classe. Nesse sentido, Frade (2021) acrescenta que o feminismo de dados funciona como uma alternativa para "sintonizar como as práticas padrão em ciência de dados servem para reforçar essas desigualdades existentes e a usar a ciência de dados para desafiar e mudar a distribuição de poder" (FRADE, 2021, s/p., *online*).

Para mitigar os problemas encontrados na incoporação tecnológica que ocorre atualmente ao redor do globo (principalmente aos relacionados com o uso indevido da inteligência artificial, como funciona no caso da pornografia *deepfake,* que acaba afetando majoritariamente a grupos marginalizados por conta de seu gênero e raça), e encontrar possíveis soluções para os referidos casos, o feminismo de dados desenvolve princípios específicos, os quais, de acordo com D'Ignazio e Klein (2020) e Frade (2021), baseiam-se em:

> 1. Examinar o poder: o feminismo de dados começa analisando como o poder opera no mundo. 2. Desafiar o poder: o feminismo de dados compromete-se a desafiar estruturas de poder desiguais e a trabalhar em prol da justiça. 3. Elevar a emoção e a personificação: o feminismo de dados nos ensina a valorizar várias formas de conhecimento, incluindo o conhecimento que provém das pessoas como corpos vivos e sensíveis do mundo. 4. Repensar binários e hierarquias: o feminismo de dados exige que desafiemos o binário de género, juntamente com outros sistemas de contagem e classificação que perpetuam a opressão. 5. Abraçar o pluralismo:

> o feminismo de dados insiste em que o conhecimento mais completo advém da síntese de múltiplas perspectivas, com prioridade dada às formas locais, indígenas e experimentais de conhecimento. 6. Considerar o contexto: o feminismo de dados afirma que os dados não são neutros ou objetivos. Eles são produtos de relações sociais desiguais, e esse contexto é essencial para a realização de análises éticas precisas. 7. Tornar o trabalho visível: o trabalho da ciência de dados, como todo o mundo, é obra de muitas mãos. O feminismo de dados torna esse trabalho visível para que possa ser reconhecido e valorizado (FRADE, 2021, s/p., *online*).

Ainda que as autoras mencionem em suas obras que o trabalho da ciência de dados (e demais trabalhos na área da tecnologia) são "obras de muitas mãos", frequentemente os ambientes tecnológicos são taxados como ambientes masculinos, de propósitos racionais e bastante inóspito para mulheres. Tal afirmativa reforça ainda mais a misoginia existente nos espectros sociais que permeiam a sociedade atual, o que, para o feminismo de dados, poderia ser desmantelado se todo o processo de criação e desenvolvimento de tecnologias fosse visível, reconhecido e valorizado.

Ou seja, a partir dos princípios do feminismo de dados expostos, conclui-se que é possível desenvolvermos futuros feministas para a IA, a fim de evitar que os sistemas sejam desenvolvidos apenas por homens e promover tecnologias baseadas na justiça social, que englobem pautas feministas e estejam, em sua maioria, livres de misoginia, de machismo e de discriminação racial.

ESTUDO DE CASO: *DEEPFAKES* E O IMPACTO DA PORNOGRAFIA NÃO CONSENSUAL

CONTEXTO

Aplicativos de inteligência artificial permitem a criação de vídeos pornográficos falsos, chamados *deepfakes*, com um único clique (ÖHMAN, 2019). O usuário faz *upload* de uma foto e o aplicativo a insere em

vídeos pornográficos pré-existentes. A maioria dos vídeos disponíveis no aplicativo apresenta mulheres, levantando sérias preocupações éticas e de segurança. Embora os resultados ainda não sejam perfeitos, a tecnologia está evoluindo rapidamente. A preocupação é que, mesmo com baixa qualidade, os vídeos podem ser convincentes para alguns, o que acaba prejudicando seriamente as vítimas (MIT TECH REVIEW, 2021). Essa prática levanta sérias questões éticas e legais, especialmente quando se trata da exploração da imagem de mulheres e meninas sem o seu consentimento.

CASOS

O quadro 2 ilustra a gravidade e a diversidade dos incidentes envolvendo *deepfakes* e pornografia não consensual, evidenciando como essa tecnologia pode ser utilizada para fins abusivos e prejudiciais. O caso da cantora Taylor Swift destaca a vulnerabilidade de figuras públicas, que podem ter sua imagem manipulada e explorada em vídeos sexualmente explícitos, expostos em plataformas como o X (antigo *Twitter*). Esse incidente não apenas gera indignação, mas também reacende o debate sobre a necessidade urgente de regulamentação e proteção legal contra esse tipo de abuso, conforme apontado pelo MIT Tech Review (2024). Outro caso preocupante é o ocorrido em uma escola de ensino médio em Nova Jersey, onde meninas foram alvo de pornografia *deepfake* criada por colegas de classe. Esse incidente impulsionou um novo movimento legislativo nos EUA para regulamentar essa prática, demonstrando o impacto direto desses crimes no mundo real, e a necessidade de proteger as vítimas e responsabilizar os agressores (WATSON, 2024).

Além desses casos específicos, o quadro menciona diversos outros incidentes de pornografia *deepfake* não consensual, incluindo os de Noelle Martin, uma ativista australiana (MARTIN, 2018), e Rana Ayyub, uma jornalista indiana, que sofreram graves consequências pessoais e profissionais devido à disseminação de vídeos falsos (UNITED NATIONS, 2022). Esses exemplos demonstram que a pornografia *deepfake* não é um problema isolado, mas sim, uma questão global que afeta diferentes pessoas e contextos, com impactos significativos em suas vidas.

QUADRO 2: Incidentes de deepfake e pornografia não consensual

VÍTIMA	DESCRIÇÃO
Taylor Swift	A cantora pop foi vítima da viralização de vídeos *deepfake* com conteúdo sexualmente explícito no X (antigo Twitter). Esse incidente gerou indignação e reacendeu o debate sobre a necessidade de regulamentação e proteção legal contra esse tipo de abuso (Mit Tech Review, 2024).
Escola de ensino médio em Nova Jersey	Meninas foram alvo de pornografia *deepfake* criada por colegas de escola, o que levou a um novo impulso legislativo para regulamentar essa prática nos EUA. O caso ganhou notoriedade e impulsionou a discussão sobre a necessidade de proteger as vítimas e responsabilizar os agressores (Watson, 2024).
Outros casos	Diversos outros casos de pornografia *deepfake* não consensual foram relatados, incluindo o de Noelle Martin, uma ativista australiana (Martin, 2018), e Rana Ayyub, uma jornalista indiana, que sofreram graves consequências pessoais e profissionais devido à disseminação de vídeos falsos (United Nations, 2022).

Fonte: elaborado pelos autores.

IMPACTOS

As vítimas de pornografia *deepfake* sofrem impactos psicológicos profundos, como humilhação, angústia e trauma, comparáveis aos da pornografia de vingança (NOBLE, 2022). A disseminação não consensual de *deepfakes* pode levar ao isolamento social, à perda de emprego e a dificuldades em relacionamentos interpessoais, afetando a vida social das vítimas de maneira devastadora. No âmbito profissional, *deepfakes* podem arruinar carreiras, especialmente para mulheres em posições de destaque, como jornalistas e professoras, devido à difamação e ao assédio online (BROUSSARD, 2018). A situação é agravada pela insuficiência da legislação atual em lidar com a pornografia *deepfake*, deixando as vítimas sem recursos legais efetivos para se protegerem e buscarem justiça (SILVA, 2022). Essa lacuna legal contribui para a sensação de impunidade dos agressores e perpetua a vulnerabilidade das vítimas, que muitas vezes se veem sem apoio ou proteção diante dessa forma de violência digital.

DISCUSSÃO DOS RESULTADOS E POSSÍVEIS SOLUÇÕES

A pornografia *deepfake* não consensual representa um grave problema social e tecnológico, com impactos devastadores para as vítimas. A combinação de soluções técnicas, como marcas d'água e escudos de proteção, com medidas legais e campanhas de conscientização, é crucial para combater essa prática e proteger os direitos e a dignidade das pessoas. A urgência de ação é evidente, e a sociedade como um todo precisa se mobilizar para garantir que a IA seja utilizada de forma ética e responsável.

No entanto, a eficácia dessas ferramentas depende da adoção generalizada e da constante atualização frente aos avanços da IA. Legalmente, a criação de leis mais rígidas, como as propostas nos EUA (TORKINGTON, 2024) e no Reino Unido (UK GOVERNMENT, 2024), são fundamentais para responsabilizar os criadores e disseminadores de *deepfakes* e oferecer proteção às vítimas. Adicionalmente, campanhas de conscientização pública desempenham um papel crucial ao informar a sociedade sobre os perigos da pornografia *deepfake* e a importância do consentimento, promovendo uma cultura de respeito e responsabilidade no uso da tecnologia. A combinação dessas estratégias é essencial para mitigar os impactos negativos dos *deepfakes* e garantir a segurança e dignidade das pessoas no ambiente digital.

Além do perigo paras as mulheres, esses aplicativos também podem ser usados para criar *deepfakes* baseados em pornografia homossexual, o que representa um risco adicional em países onde a homossexualidade é criminalizada. Especialistas defendem a proibição desse tipo de conteúdo em plataformas de mídia social e até mesmo a criminalização da sua criação e consumo (TONG, 2024).

Desse modo, verifica-se que, embora não exista uma relação direta e explícita entre o aplicativo de criação de *deepfakes* e a tecnologia de reconhecimento facial, considera-se importante acrescentá-la no presente trabalho, já que o reconhecimento facial pode ser utilizado como uma ferramenta para identificar indivíduos em fotos e vídeos, o que poderia facilitar o uso indevido em aplicativos. Por exemplo, alguém poderia utilizar a tecnologia de reconhecimento facial a fim de identificar uma pessoa em

uma fotografia e, em seguida, utilizar essa mesma fotografia para criar um *deepfake* pornográfico sem o seu consentimento, o que configura mais uma forma de utilizar a inteligência artificial para produzir malefícios à grupos que já encontram-se socialmente marginalizados.

No que tange a possíveis soluções para o referido problema, identifica-se que, no combate à crescente ameaça da pornografia *deepfake* não consensual, diversas soluções se mostram promissoras. Tecnicamente, a inserção de marcas d'água em imagens geradas por IA, como no sistema *SynthID* do Google, facilita a identificação de *deepfakes* (RÖHE & SANTAELLA, 2024), enquanto ferramentas como *PhotoGuard* e *Fawkes* atuam como escudos de proteção, alterando as imagens de forma imperceptível ao olho humano e dificultando sua manipulação para fins maliciosos (MANCINI, 2024).

De acordo com Ananya Rastogi (2024), o conserto do problema ético atrelado à criação de vieses é complexo tanto quanto o problema em si. No entanto, a autora destaca que atualmente a pressão pelo código aberto tem sido uma das grandes apostas para a possível correção desses vises, uma vez que os usuários precisam ser capazes de identificar a origem da discriminação no próprio código para que o possam reparar. Outro ponto destacado pela autora é a possibilidade de os próprios usuários fazerem o *upload* de imagens semelhantes com aquelas que buscam. Para a autora, esse seria um mecanismo eficaz para garantir a educação das plataformas de IA generativa e auxiliar na mitigação da reprodução de preconceitos pré-existentes nos bancos de dados.

CONSIDERAÇÕES FINAIS

Este artigo se propôs a analisar os vieses presentes nos sistemas de inteligência artificial, com foco particular no reconhecimento facial, e investigar a possibilidade de mitigar esses vieses para promover uma IA mais justa e equitativa. A hipótese central era que, por meio de uma abordagem meticulosa e eticamente orientada, é possível utilizar algoritmos para reduzir vieses e promover futuros baseados na justiça social, pautas feministas e livres de discriminação racial. Ao longo da pesquisa, foi

explorado o funcionamento da tecnologia de reconhecimento facial, analisadas situações problemáticas decorrentes do uso da IA, e a busca por compreender as perspectivas futuras para lidar com vieses algorítmicos e promoção do desenvolvimento e uso ético e responsável da IA.

De acordo com Noble (2022), é fundamental questionar se os serviços de IA criados por empresas como as *Big Techs* atuam em prol do interesse público. Neste contexto, o bem público, que é influenciado por empresas de publicidade que lucram com a coleta e venda de dados, suscita sérias preocupações sobre a privacidade e o uso ético da IA. A falta de transparência nos algoritmos e a ausência de regulamentação adequada perpetuam a opacidade e a irresponsabilidade no desenvolvimento e implementação dessas tecnologias.

O caso da pornografia *deepfake* ilustra os perigos da manipulação de dados e imagens com fins maliciosos. A facilidade de acesso a ferramentas de criação de *deepfakes* e a dificuldade em detectar e remover esse conteúdo online evidenciam a necessidade urgente de medidas de proteção e responsabilização. A crescente ameaça da pornografia *deepfake*, que afeta principalmente mulheres e meninas, exige uma resposta enérgica da sociedade e principalmente, dos legisladores.

As discussões acerca do reconhecimento facial evidenciam os desafios da vigilância em larga escala e o risco de discriminação algorítmica, sobretudo contra grupos marginalizados. A ausência de regulamentação específica e a falta de transparência dos sistemas de reconhecimento facial aprofundam esses problemas, contribuindo para a perpetuação do machismo, da misoginia, do racismo estrutural e da vigilância constante que é direcionada a corpos vulnerabilizados.

Além disso, é importante mencionar que a pesquisa realizada se justifica na medida em que, como mencionado por Renata Wassermann (2024, p. 126): "no Brasil ainda há poucas iniciativas específicas para o aumento das diversidades em IA". Ainda que a Sociedade Brasileira de Computação (SBC) apoie desde o ano de 2007 o WIT - Women in Information Technology (WASSERMANN, 2024), ainda verifica-se a necessidade frequente de debatermos políticas de educação e treinamento

de mulheres nas áreas de tecnologia, "assim como seu recrutamento em empresas e grupos de pesquisa" (WASSERMANN, 2024, p. 126).

Diante desse cenário, é crucial que a sociedade civil, pesquisadores, legisladores e empresas de tecnologia trabalhem em conjunto para garantir o desenvolvimento e uso responsável da IA. A afirmação é sustentada baseando-se na perspectiva de que, assim como abordado por Crawford (2021) e Frade (2021), a IA é fundamentada por práticas e técnicas sociais, por política, cultura, pelas estruturas que a sustentam, assim como pelo interesse depositado por instituições privadas e públicas.

Assim sendo, a promoção da transparência, da diversidade e da inclusão na criação de sistemas de IA, bem como a criação de mecanismos de proteção legal para as vítimas de abuso e a responsabilização dos agressores, são passos essenciais para mitigar os riscos e garantir que a IA seja utilizada para o bem comum, em vez de perpetuar desigualdades e causar danos. A busca por uma IA ética e justa é um desafio urgente e complexo, mas fundamental para construir um futuro digital mais equitativo e seguro para todos.

REFERÊNCIAS

ABBAS DA SILVA, Lorena; FRANQUEIRA, Bruna Diniz; HARTMANN, Ivar A. O que os olhos não veem, as câmeras monitoram: reconhecimento facial para segurança pública e regulação na América Latina. *Revista Digital de Direito Administrativo*, v. 8, n. 1, p. 171–204, 2021. Disponível em: https://revistas.usp.br/rdda/article/view/173903. Acesso em: 19/maio/2025.

ADAM, Alison. Constructions of gender in the history of artificial intelligence. *IEEE Annals of the History of Computing*, v. 18, n. 3, p. 47–53, 1996. Disponível em: https://ieeexplore.ieee.org/abstract/document/511944. Acesso em: 12/fevereiro/2021.

AHMED, Saadaldeen Rashid; NASSAR, Mahmoud; SATTAR, Awni; MAJZUB, Rania; ELSAYED, Muawya. Analysis survey on deepfake detection and recognition with convolutional neural networks. [S. l.], 2022. Disponível em: https://ieeexplore.ieee.org/document/9799858. Acesso em: 29/junho/2024.

ALVI, Mohsan; ZISSERMAN, Andrew; NELLÅKER, Christoffer. Turning a blind eye: explicit removal of biases and variation from deep neural network embeddings. [S. l.], 2018. Disponível em: https://openaccess.thecvf.com/content_ECCVW_2018/papers/11129/Alvi_Turning_a_Blind_Eye_Explicit_Removal_of_Biases_and_Variation_ECCVW_2018_paper.pdf. Acesso em: 30/junho/2024.

ANGWIN, Julia; LARSON, Jeff; MATTU, Surya; KIRCHNER, Lauren. Machine bias. *ProPublica*, 2016. Disponível em: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Acesso em: 17/abril/2025.

ANGWIN, Julia; TOBIN, Ariana; VARNER, Madeleine. Facebook (still) letting housing advertisers exclude users by race. *ProPublica*, 2017. Disponível em: https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin. Acesso em: 21/maio/2024.

BAROCAS, Solon; SELBST, Andrew D. Big data's disparate impact. *California Law Review*, v. 104, n. 3, p. 671–732, 2016. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899. Acesso em: 17/abril/2025.

BROUSSARD, Meredith. *Artificial unintelligence*. Cambridge: *The MIT Press*, 2018.

BROWNE, Jude; DRAGE, Eleanor; MCINERNEY, Kerry. Tech workers' perspectives on ethical issues in AI development: foregrounding feminist approaches. *Big Data & Society*, v. 11, n. 1, 2024. Disponível em: https://www.repository.cam.ac.uk/items/26f0861d-9ed5-4ab8-a338-6480384ffe09. Acesso em: 28/junho/2024.

BUOLAMWINI, Joy; GEBRU, Timnit. Gender shades: intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability and Transparency*, p. 77–91, 2018. Disponível em: https://proceedings.mlr.press/v81/buolamwini18a.html. Acesso em: 30/junho/2024.

COSTA, Ramon Silva; KREMER, Bianca. Inteligência artificial e discriminação: desafios e perspectivas para a proteção de grupos vulneráveis diante das tecnologias de reconhecimento facial. *Revista Brasileira de Direitos Fundamentais & Justiça*, v. 16, n. 1, 2022. Disponível em: https://dfj.emnuvens.com.br/dfj/article/view/1316. Acesso em: 19/maio/2025.

DILLON, Sarah; COLLETT, Clementine. AI and gender: four proposals for future research. *Apollo – University of Cambridge Repository*, 2019.

DOLHANSKY, Brian; BITTON, Hilary; COHEN, Ben; DARRELL, Trevor; FERNANDEZ, Lauren; GESSERT, Felix; JANSSEN, Dieter; KUMAR, Deepak; LUX, Elias; MAHESHWARI, Gaurav; et al. The DeepFake Detection Challenge (DFDC) dataset. *arXiv*, v. 4, 2020.

FEUERRIEGEL, Stefan; SCHÖNFELDER, Nico; NITZSCHE, Nico. Generative AI. *Business & Information Systems Engineering*, v. 66, n. 1, p. 111–126, 2023. Disponível em: https://link.springer.com/article/10.1007/s12599-023-00834-7. Acesso em: 30/junho/2024.

GERRARD, Juliet A.; MACLAURIN, James; WALTON, Michael. By 2030, AI will contribute $15 trillion to the global economy. *World Economic Forum*, 2019. Disponível em: https://www.weforum.org/agenda/2019/08/by-2030-ai-will-contribute-15-trillion-to-the-global-economy/. Acesso em: 24/maio/2024.

GIBBS, Samuel. Google says sorry over racist Google Maps White House search results. *The Guardian*, 2015. Disponível em: https://www.theguardian.com/technology/2015/may/20/google-apologises-racist-google-maps-white-house-search-results. Acesso em: 21/maio/2024.

GROSS, Nicole. What ChatGPT tells us about gender: a cautionary tale about performativity and gender biases in AI. *Social Sciences*, v. 12, n. 8, p. 435, 2023. Disponível em: https://www.mdpi.com/2076-0760/12/8/435. Acesso em: 28/junho/2024.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. *ISO/IEC 22989:2022 – Information technology – Artificial intelligence – Artificial intelligence concepts and terminology*. Geneva: *ISO*, 2022.

KAUR, Paramjit; KUMAR, Rajeev; SHUKLA, Prabhakar. Facial-recognition algorithms: a literature review. *Medicine, Science and the Law*, v. 60, n. 2, p. 002580241989316, 2020. Disponível em: https://journals.sagepub.com/doi/abs/10.1177/0025802419893168. Acesso em: 30/junho/2024.

KIM, Ahyeon; SONG, Haeyeon; LEE, Heekyung. Effects of gender and relationship type on the response to artificial intelligence. *Cyberpsychology, Behavior, and Social Networking*, v. 22, n. 4, p. 249–253, 2019.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. *Nature*, v. 521, n. 7553, p. 436–444, 2015.

LEE, Kai-Fu. *AI superpowers: China, Silicon Valley, and the new world order*. Boston: *Houghton Mifflin Harcourt*, 2018.

MAGNO, Madja Elayne da Silva Penha; BEZERRA, Josenildo Soares. Vigilância negra: o dispositivo de reconhecimento facial e a disciplinaridade dos corpos. *Novos Olhares*, v. 9, n. 2, p. 45–52, 2020. Disponível em: https://www.revistas.usp.br/novosolhares/article/view/165698/169548. Acesso em: 30/junho/2024.

MANCINI, Malena Beatriz. *Deepfaked: propuesta de regulación de réplicas digitales*. *Repositorio UDESA*, 2024. Disponível em: https://repositorio.udesa.edu.ar/jspui/handle/10908/23851. Acesso em: 30/junho/2024.

MARTIN, Noelle. Sexual predators edited my photos into porn – how I fought back. *TEDxPerth*, 2018. Disponível em: https://www.ted.com/talks/noelle_martin_sexual_predators_edited_my_photos_into_porn_how_i_fought_back. Acesso em: 30/junho/2024.

MEHRABI, Ninareh; MORSTAN, Fred; SAXENA, Abhishek; FRIEDLER, Suresh V.; CHOUDHARY, Prateek. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, v. 54, n. 6, p. 1–35, 2021. Disponível em: https://dl.acm.org/doi/10.1145/3457607. Acesso em: 17/abril/2025.

METZ, Cade. Who is making sure the A.I. machines aren't racist? *The New York Times*, 15/março/2021. Disponível em: https://www.nytimes.com/2021/03/15/technology/artificial-intelligence-google-bias.html. Acesso em: 22/maio/2024.

MIT TECH REVIEW. Novo aplicativo de inteligência artificial coloca mulheres em vídeos pornôs com um clique. *MIT Technology Review*, 2021. Disponível em: https://mittechreview.com.br/novo-aplicativo-de-inteligencia-artificial-coloca-mulheres-em-videos-pornos-com-um-clique/. Acesso em: 28/junho/2024.

MIT TECH REVIEW. Taylor Swift: mais uma vítima do deepfake. *MIT Technology Review*, 2024. Disponível em: https://mittechreview.com.br/taylor-swift-mais-uma-vitima-do-deepfake/. Acesso em: 30/junho/2024.

NOBLE, Safiya Umoja. *Algoritmos da opressão*. São Paulo: *Editora Rua do Sabão*, 2022.

O'NEIL, Cathy. *Weapons of math destruction: how big data increases inequality and threatens democracy*. New York: *Crown*, 2016.

O GLOBO. Busca por termo racista no Google Maps mostra Casa Branca, e site pede desculpas. *O Globo*, 2015. Disponível em: https://oglobo.globo.com/economia/busca-por-termo-racista-no-google-maps-mostra-casa-branca-site-pede-desculpas-16210255. Acesso em: 30/junho/2024.

ÖHMAN, Carl. Introducing the pervert's dilemma: a contribution to the critique of deepfake pornography. *Ethics and Information Technology*, v. 22, 2019.

RANA, Md Shohel; NATH, Rajib; MONDOL, Md Al-Amin; PAUL, Avijit. Deepfake detection: a systematic literature review. *IEEE Access*, v. 10, p. 1–1, 2022.

RICH, Elaine. Artificial intelligence and the humanities. *Computers and the Humanities*, v. 19, n. 2, p. 117–122, 1985. Disponível em: https://www.jstor.org/stable/30204398. Acesso em: 20/junho/2024.

RÖHE, Anderson; SANTAELLA, Lucia. Prognósticos das deepfakes na política eleitoral. *Organicom*, v. 21, n. 44, p. 187–196, 2024. Disponível em: https://revistas.usp.br/organicom/article/view/221294. Acesso em: 30/junho/2024.

RUSSEL, Stuart; NORVIG, Peter. *Artificial intelligence: a modern approach*. New Jersey: *Prentice Hall*, 2021.

SALAS, Javier. Google conserta seu algoritmo "racista" apagando os gorilas. *El País*, 2018. Disponível em: https://brasil.elpais.com/brasil/2018/01/14/tecnologia/1515955554_803955.html. Acesso em: 21/maio/2024.

SHINDE, P. P.; SHAH, S. A. A review of machine learning and deep learning applications. [S. l.], 2018. Disponível em: https://ieeexplore.ieee.org/document/8697857. Acesso em: 17/abril/2025.

SILVA, Tarcízio. *Racismo algorítmico*. São Paulo: *Edições Sesc SP*, 2022.

SILVA, Rosane Leal da; SILVA, Fernanda dos Santos Rodrigues. Reconhecimento facial e segurança pública: os perigos do uso da tecnologia no sistema penal seletivo brasileiro. [S. l.: s. n.], 2019. Disponível em: https://www.ufsm.br/app/uploads/sites/563/2019/09/5.23.pdf. Acesso em: 30/junho/2024.

SULEIMENOV, Ibragim E.; ZHANBURGIN, Yerbolat A.; AIDARKHANOV, Almas; MUKHANBETZHANOV, Nurlybek. Artificial intelligence. *Proceedings of the 2020 6th International Conference on Computer and Technology Applications*, [S. l.], 2020.

SUVOROVA, Inna A. Deepfake pornography as a male gaze on fan culture. *arXiv* (Cornell University), [S. l.], 2022. Disponível em: https://arxiv.org/abs/2202.00374. Acesso em: 30/junho/2024.

TONG, Anna. "AI godfather", others urge more deepfake regulation in open letter. [S. l.], 2024. Disponível em: https://www.reuters.com/technology/cybersecurity/ai-godfather-others-urge-more-deepfake-regulation-open-letter-2024-02-21/. Acesso em: 30/junho/2024.

TORKINGTON, Simon. The US has plans to tackle AI-generated deepfakes. *World Economic Forum*, 2024. Disponível em: https://www.weforum.org/agenda/2024/02/ai-deepfakes-legislation-trust/. Acesso em: 30/junho/2024.

TOUPIN, Sophie. Shaping feminist artificial intelligence. *New Media & Society*, v. 26, n. 1, 2023.

UK GOVERNMENT. Government cracks down on "deepfakes" creation. *UK Government*, 2024. Disponível em: https://www.gov.uk/government/news/government-cracks-down-on-deepfakes-creation. Acesso em: 30/junho/2024.

UNITED NATIONS. India: attacks against woman journalist Rana Ayyub must stop – UN experts. [S. l.], 2022. Disponível em: https://www.ohchr.org/en/press-releases/2022/02/india-attacks-against-woman-journalist-rana-ayyub-must-stop-un-experts. Acesso em: 30/junho/2024.

VINCENT, James. Twitter taught Microsoft's AI chatbot to be a racist in less than a day. *The Verge*, 2016. Disponível em: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist. Acesso em: 22/maio/2024.

WAKABAYASHI, Daisuke. Lawsuit accuses Google of bias against Black employees. *The New York Times*, 18/março/2022. Seção: *Technology*. Disponível em: https://www.nytimes.com/2022/03/18/technology/google-discrimination-suit-black-employees.html. Acesso em: 22/maio/2024.

WATSON, Angus. Teenager questioned after explicit AI deepfakes of dozens of schoolgirls shared online. *CNN*, 2024. Disponível em: https://edition.cnn.com/2024/06/13/australia/australia-boy-arrested-deepfakes-schoolgirls-intl-hnk/index.html. Acesso em: 30/junho/2024.

ZAFEIRIOU, Stefanos; ZHANG, Cha; ZHANG, Zhengyou. A survey on face detection in the wild: past, present and future. *Computer Vision and Image Understanding*, v. 138, p. 1–24, 2015.

Texto recebido em 14/06/2024 e aceito em 13/05/2025.