UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

ERICK MORAIS DE SENA

VIÉS NA IA:
COMO O VIÉS ALGORITMO INFLUENCIA NA PERPETUAÇÃO DE ESTEREÓTIPOS E DESIGUALDADES EXISTENTES.

CAMPINA GRANDE - PB

2023

ERICK MORAIS DE SENA

VIÉS NA IA:
COMO O VIÉS ALGORITMO INFLUENCIA NA PERPETUAÇÃO DE ESTEREÓTIPOS E DESIGUALDADES EXISTENTES.

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Francilene Procópio Garcia

CAMPINA GRANDE - PB

2023

ERICK MORAIS DE SENA

VIÉS NA IA:
COMO O VIÉS ALGORITMO INFLUENCIA NA PERPETUAÇÃO DE ESTEREÓTIPOS E DESIGUALDADES EXISTENTES.

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Francilene Procópio Garcia

Orientador – UASC/CEEI/UFCG

Carlos Wilson Dantas De Almeida

Examinador – UASC/CEEI/UFCG

Francisco Vilar Brasileiro

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 17 de NOVEMBRO de 2023.

CAMPINA GRANDE - PB

RESUMO

O avanço da Inteligência Artificial (IA) tem revolucionado diversos setores da sociedade, como a saúde, indústria, educação, justiça, entre outros. Entretanto, essa crescente incorporação da IA em nossas vidas também representa desafios significativos relacionados a sua aceitação generalizada. Um dos problemas mais cruciais é o viés algorítmico presente em sistemas de IA, que pode levar a decisões discriminatórias e injustas, perpetuando preconceitos e afastando as minorias das áreas essenciais, como emprego, saúde, educação, justiça e outros setores fundamentais da sociedade. Este trabalho tem como objetivo relacionar o tópico do viés algoritmo com os impactos das decisões automatizadas sobre as minorias.

BIAS IN AI: HOW ALGORITHMIC BIAS INFLUENCES THE PERPETUATION OF STEREOTYPES AND EXISTING INEQUALITIES.

ABSTRACT

The advancement of Artificial Intelligence (AI) has revolutionized several sectors of society, such as health, industry, education, justice, among others. However, this increasing incorporation of AI into our lives also poses significant challenges related to its widespread acceptance. One of the most crucial problems is the algorithmic bias present in AI systems, which can lead to discriminatory and unfair decisions, perpetuating prejudices and alienating minorities from essential areas, such as employment, health, education, justice and other fundamental sectors of society. This work aims to relate the topic of algorithm bias with the impacts of automated decisions on minorities.

VIÉS NA IA: Como o Viés Algoritmo Influencia na Perpetuação de Estereótipos e Desigualdades Existentes.

Erick Morais de Sena
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba - Brasil
erick.sena@ccc.ufcg.edu.br

Francilene Procópio Garcia
(orientadora)
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba - Brasil
garcia@computacao.ufcg.edu.br

RESUMO

O avanço da Inteligência Artificial (IA) tem revolucionado diversos setores da sociedade, como a saúde, indústria, educação, justiça, entre outros. Entretanto, essa crescente incorporação da IA em nossas vidas também representa desafios significativos relacionados a sua aceitação generalizada. Um dos problemas mais cruciais é o viés algorítmico presente em sistemas de IA, que pode levar a decisões discriminatórias e injustas, perpetuando preconceitos e afastando as minorias das áreas essenciais, como emprego, saúde, educação, justiça e outros setores fundamentais da sociedade. Este trabalho tem como objetivo relacionar o tópico do viés algoritmo com os impactos das decisões automatizadas sobre as minorias.

PALAVRAS-CHAVE

Viés Algorítmico, IA, Discriminação, Minorias, Aceitação da IA.

1. INTRODUÇÃO

Nos últimos anos, o avanço da Inteligência Artificial (IA) e sua crescente incorporação na sociedade tem revolucionado diversas áreas sociais. Mas a adoção da tecnologia parece ter uma propensão a tomar decisões controversas que desafiam a ética e a segurança social. Esta pesquisa investiga a influência do viés algorítmico em sistemas de Inteligência Artificial, apresentando exemplos concretos de situações desastrosas que envolvem interações entre usuários e esses algoritmos, trazendo para a superfície a discussão do tema, além de buscar entender como o viés algorítmico perpetua preconceitos e discriminações de grupos já marginalizados na sociedade. Mas afinal, o que é viés algorítmico? Viés algoritmo refere-se à tendência dos algoritmos de IA de reforçar ou perpetuar estereótipos e desigualdades existentes. Isso ocorre principalmente porque as pessoas que programam ou projetam tais algoritmos nem sempre estão cientes de seus próprios preconceitos ou, muitas vezes, usam dados que refletem desigualdades históricas. O resultado são algoritmos que influenciam na perpetuação de tendências discriminatórias, subestimam as habilidades de certos grupos ou eternizam preconceitos.

2. CONTEXTUALIZAÇÃO DO TEMA

A difusão da IA na sociedade apresenta uma série de vantagens potenciais, no entanto, essa incorporação não é isenta de desafios como questões éticas, legais e de segurança social. Este trabalho busca: a) expor a problemática da atuação do viés algorítmico na sociedade; b) chamar a atenção para o nível de seriedade do problema; c) fomentar o debate sobre como podemos tornar a IA segura, e se há possibilidade de diminuir os danos potenciais de suas ações.

O viés algorítmico presente na IA é um fator preocupante pois abre espaço para tomada de decisões que ferem o princípio de equidade da própria inteligência artificial, além de princípios fundamentais de ética, justiça e segurança social. À medida que a IA se torna onipresente em diversos setores da sociedade, cresce o risco de que a adoção dessa tecnologia possa causar danos a uma ampla gama de usuários e não usuários.

Um exemplo da ação do viés algorítmico foi com a empresa Nikon que projetou a câmera digital Nikon Coolpix S630 com um algoritmo de IA que não reconhecia quando usuários asiáticos estavam de olhos abertos. Nesse caso, apesar de ser uma empresa japonesa, o treinamento realizado para reconhecimento de rostos asiáticos, e os testes os quais esse sistema foi submetido foi precário, falho e insuficiente. Ao longo deste artigo, o leitor terá a oportunidade de explorar mais casos semelhantes a este

3. PROBLEMA DA PESQUISA

Este artigo tenta responder a seguinte pergunta: Como a automatização de decisões baseadas em algoritmos enviesados impactam na sociedade, perpetuando a discriminação e o preconceito de grupos sociais marginalizados e minoritários? É possível evitar ou diminuir os danos causados por humanos com viés no processo de criação de algoritmos enviesados?

4. OBJETIVOS DO TRABALHO

O objetivo principal deste trabalho é trazer a discussão sobre o viés algoritmo na IA e seu impacto na sociedade, principalmente nos grupos sociais mais fragilizados. Este artigo também busca citar algumas medidas capazes de minimizar a influência negativa

das ações do viés algorítmico em sistemas de Inteligência Artificial.

5. MOTIVAÇÃO PESSOAL

A escolha deste tema foi motivada pelo receio do dano físico, psicológico, emocional, financeiro, e social, que sistema de IA viesados pode causar tanto em usuários quanto em não usuários, principalmente naqueles pertencentes a grupos minoritários ou marginalizados. Além disso, há uma motivação de levar o tema para discussão entre a comunidade acadêmica, engenheiros de software, empresários e a sociedade em geral.

6. DESENVOLVIMENTO

6.1 A IA Marca Presença na Sociedade

A sociedade acreditava que o futuro da IA envolveria apenas tarefas simples e repetitivas que exigiriam tomadas de decisões de baixo nível. No entanto, com o avanço do poder computacional, a sofisticação da IA cresceu rapidamente tornando-se capaz de classificar e analisar grandes quantidades de dados, além de seu poder de aprendizado que tornou possível o uso em tarefas complexas (Pazzanese, 2020) [1].

Chamorro-Premuzic et al. enfatizaram o uso da IA em empresas do nosso dia a dia, (2019): "[...] empresas como Amazon e Alibaba até [...] canais como YouTube e Netflix usam para comercializar seu conteúdo mais recente [...]". Em 2016, um artigo divulgado no Jornal da USP, já destacava o grande avanço da IA nas últimas décadas (Sichman et al., 2016) [2]:

> [...] Sistemas de busca de informação e de recomendação de produtos são parte de nossa experiência cotidiana. Tais produtos aprendem a partir de dados e decidem com base em regras e em experiências passadas. O sistema financeiro também depende fortemente de programas com capacidade de raciocínio e decisão, que hoje comandam grandes investimentos em bolsas ao redor do mundo. Usamos hoje [...] até mesmo veículos aéreos não tripulados (drones) para fins pacíficos e militares. Em resumo, nosso mundo já é um mundo no qual máquinas apresentam comportamentos tipicamente associados à "inteligência" [...].

No emprego, além de participar efetivamente no processo de análise de currículos, voz e expressões faciais dos entrevistados, a IA pode facilitar o processo de contratação, funcionando como um chatbot que interage com os candidatos (Rąb-Kettler & Lehnervp, 2019) [3].

Ainda nesse contexto, mas analisando o nível de investimento nessa área e concedendo um status superior contra métodos que envolvem decisões humanas, Raghavan et al. (2020) [4] afirmam que:

> "Em janeiro de 2020, havia pelo menos 11 empresas que ofereciam avaliações algorítmicas de pré-seleção que levantaram entre US$ 1 milhão e US$ 93 milhões em capital de investimento somente nos EUA e no Canadá".

A IA também marca grande presença na área da saúde, graças a: a) rapidez em diagnosticar doenças; b) análise de imagens médicas, como radiografias, tomografias e exames de ressonância magnética; c) identificação de anormalidades, como tumores, fraturas e outros problemas de saúde, com alta precisão. Segundo Pazzanese (2020) [1]: "Na saúde, os profissionais médicos esperam que o impacto maior e mais imediato seja na análise de dados, imagens e diagnóstico". Com uma grande base de dados, é possível identificar padrões e variações em questão de milésimos de segundo em exames de imagem. Por exemplo, o sistema pode analisar células e identificar o início de tumores não vistos a olho nu. Também tem utilidade na prevenção de complicações (como infecção generalizada) e na indicação dos melhores tratamentos (DRG Brasil, 2022) [5].

Para os sistemas de recomendações, a IA desempenha seu papel com excelência, pois seu uso torna viável desenvolver um sistema de recomendações que compreenda as preferências individuais dos usuários, baseando-se em ações como cliques, curtidas, histórico, entre outras. Esse sistema se revela muito valioso ao auxiliar os usuários na descoberta de produtos e serviços que seriam difíceis de encontrar sozinhos (NVIDIA, [entre 2018 e 2023]) [6].

6.2 Exemplos de Discriminação

Diante de tantos exemplos da propagação da IA nos diversos setores da sociedade, além de sua efetividade em ajudar o ser humano a realizar tarefas com grande eficiência, fica difícil associar o uso dos algoritmos de IA com a discriminação e perpetuação de preconceitos, porém essa associação existe e é o foco da discussão deste artigo.

Há uma preocupação crescente na presença da IA nos setores fundamentais da sociedade – situações em que ela toma decisões que são sistematicamente injustas para determinados grupos de pessoas (Marr, 2022) [7]. O preconceito social presente em algoritmos de inteligência artificial é difícil de identificar e rastrear, mas sabemos que está em todo lugar. (Lexalytics, [entre 2021 e 2023]) [8].

Nos próximos parágrafos, examinaremos exemplos marcantes em que sistemas de IA, como o infame bot Tay, o reconhecimento facial do Google Fotos, as decisões discriminatórias do Apple Card e até casos de erro grave, como a prisão injusta de um homem negro em Michigan, levantaram questões importantes sobre a confiabilidade e a ética da IA em nossa sociedade. Essas ilustrações demonstram a necessidade crítica de aprimorar o desenvolvimento e a supervisão da IA, a fim de mitigar seus impactos adversos e promover uma adoção mais responsável e justa.

1. Caso envolvendo Apple Card e seu algoritmo sexista:

   Esse exemplo envolve uma big tech muito conhecida, a Apple. O fato aconteceu em novembro de 2019, quando a Apple Card determinava os limites de crédito das pessoas baseada no sexo. Verificou-se que as mulheres recebiam menos crédito do que seus cônjuges, embora compartilhassem a mesma renda e pontuação de crédito [9]. Kori Hale [10], CEO da CultureBanx, alerta para a possibildade de discriminação ou prejuízo para determinados grupos ou indivíduos em processos como concessão de crédito, empréstimos e benefícios bancários.

2. Caso envolvendo racismo no reconhecimento facial usado pela polícia dos EUA:

De acordo com uma matéria escrita pela ACLU (União da Liberdade Civíl Americana) [11], no início de 2020, a polícia de Detroit deteve erroneamente Robert Williams - um homem negro que reside nos subúrbios de Detroit - por um período de 30 horas devido a um erro no sistema de reconhecimento facial da Polícia do Estado de Michigan. Esse sistema informou incorretamente aos policiais que Robert Williams era o suspeito de roubo de relógios que estavam procurando. “Espero que todos vocês não pensem que todos os homens negros são parecidos”, disse Robert ao ser interrogado pela polícia. A adoção da tecnologia escancarou um problema: o reconhecimento facial não conseguia distinguir os negros, visto que a única coisa em comum entre Robert e o suspeito capturado pela câmera de vigilância da relojoaria é que ambos são homens negros e altos.

3. Caso envolvendo racismo no algoritmo de agrupamento de imagens da Google:

   “A google é racista?”, essa foi a pergunta feita pela internet quando um jovem, gravando de seu celular, postou no Twitter a diferença do resultado da busca entre “três jovens brancos” e “três jovens negros”. Segundo o Jornal El País [12], a empresa de tecnologia respondeu à edição britânica do site Huffington Post que “isso significa que, às vezes, interpretações desagradáveis sobre um assunto delicado podem ter um impacto nos resultados de buscas na rede”, os quais “não refletem as opiniões nem os valores da Google”. O resultado da primeira busca mostrava três jovens brancos juntos sorrindo para a foto, já o resultado para “três jovens negros” mostrava fotos usadas em fichas policiais.

   Javier Salas [13] escreveu em uma publicação para o Jornal El País comentando sobre outro caso envolvendo um usuário do Google Fotos:

   > [...] Em junho de 2015, um usuário do Google Photos descobriu que o programa etiquetava seus amigos negros como gorilas. Nesse caso, a inteligência artificial do Google não era capaz de distinguir a pele de um ser humano da dos macacos, como gorilas e chimpanzés [...].

   Ainda sobre esse caso, a revista Wired realizou um teste, alimentando um dispositivo móvel com uma vasta coleção de imagens abrangendo diversas espécies de macacos. Surpreendentemente, o programa da Google foi capaz de categorizar de forma precisa orangotangos, gibões, saguis e babuínos, mas apresentou uma notável lacuna ao não responder a solicitações de pesquisa envolvendo termos como “macacos”, “gorilas” e “chimpanzés”. Este fenômeno levantou a questão se o algoritmo de gerenciamento de fotos pessoais da Google pode ter inadvertidamente censurado essas palavras em seu léxico. “A forma de resolver o problema é apagar o problema: autocensurar essas etiquetas”, afirma Salas.

4. Caso do chatbot da Microsoft que postava tweets com discurso de ódio:

   Uma outra inteligência artificial que gerou problemas devido ao mau comportamento foi o Tay, um chatbot destinado ao Twitter com a finalidade de engajar em conversas com os usuários e aprender com essas interações. No início de março de 2016, a Microsoft lançou essa inteligência artificial com o propósito de que o chatbot adquirisse gradualmente habilidades linguísticas por meio de diálogos naturais e amigáveis, adaptando-se à linguagem e ao estilo de comunicação dos usuários. Contudo, o que era inicialmente concebido como um experimento logo se tornou um desafio significativo de relações públicas para a Microsoft (Revista Veja, 2016) [14]. O Jornal Independent (2016) [15] descreve que o problema foi que várias pessoas começaram a interagir com o Tay deliberadamente para ensiná-lo a fazer declarações ofensivas, racistas e sexistas. Em questão de horas, Tay começou a emitir brincadeiras e piadas sem graça, além de respostas inapropriadas e ofensivas, incluindo declarações como: “Eu odeio feministas e todas elas deveriam morrer e queimar no inferno” ou “Bush fez o 11 de setembro e Hitler teria feito um trabalho melhor do que esse macaco que temos agora [Barack Obama]".

5. Caso da Nikon que projetou um sistema incapaz de reconhecer rostos asiáticos:

   A empresa em questão desenvolveu uma câmera digital chamada Nikon Coolpix S630, que possui embutido em seu sistema de reconhecimento facial um algoritmo de IA que detecta quando o sujeito da foto pisca o olho. Segundo a publicação do site The Society Pages [16], uma usuária asiática denunciou o caso dizendo: “Enquanto eu tirava fotos da minha família, ele perguntava 'Alguém piscou?' mesmo que nossos olhos estivessem sempre abertos”. Esse caso teve consequências para a reputação da Nikon, principalmente por ser japonesa e seu algoritmo de IA não reconhecer com clareza pessoas asiáticas, além de destacar a importância de garantir que os sistemas de IA sejam rigorosamente testados em relação ao viés algorítmico antes de serem implantados em produtos de consumo.

6. Caso envolvendo a Amazon e seu sistema de RH sexista:

   Esse caso aconteceu com a big tech Amazon, que assim como muitas empresas nos dias de hoje, está ávida por ferramentas que possam auxiliar nas funções de RH para selecionar os melhores candidatos. Segundo a revista digital Época Negócios [17], a ferramenta de contratação pontuava candidatos de uma a cinco estrelas, mas a empresa percebeu que esse sistema não classificava candidatos para vagas de desenvolvedores de software e outros cargos técnicos de maneira neutra em termos de gênero. Mas isso não aconteceu por acaso, o fato é que o algoritmo aprendeu com os dados de currículos enviados à empresa em um período de 10 anos, no qual a maioria veio de homens que dominavam a indústria da tecnologia naquela época. Como resultado desses dados de treinamento, o sistema passou a penalizar frases no currículo que incluíam a palavra “mulheres” e até mesmo rebaixou candidatas de faculdades só para mulheres.

7. Caso de racismo com o algoritmo de assistência médica:

Para finalizar essa seção vamos abordar o caso de um algoritmo médico nos EUA que era preconceituoso com pacientes negros. O

algoritmo de assistência médica priorizava consistentemente pacientes brancos que não sofriam de doenças graves e eliminava pacientes negros de um programa destinado a ajudar pessoas que precisavam de cuidado intenso, de acordo com o site Giz Brasil [18]. A inteligência artificial pretendia medir quais pacientes se beneficiariam mais do programa de gestão de cuidados de saúde de alto risco, que envolvia a disponibilidade de uma equipe dedicada aos cuidados de saúde, com horários extra para consultas. Sob a estimativa da IA, por exemplo, 18% dos pacientes que mereciam estar nesses programas seriam negros; mas os autores estimaram que o número real deveria estar mais próximo de 47%. O processo de tomada de decisão da IA foi projetado para ser neutro em termos raciais. No entanto, os autores descobriram que outras suposições foram programadas com preconceitos contra os negros.

6.3 Impacto na Sociedade

O impacto e os desafios da IA na sociedade são questões que ganham ainda mais relevância à medida que a adoção dessa tecnologia cresce cada vez mais no nosso cotidiano, e é por isso que não dá para falar sobre inteligência artificial, na data que este artigo é produzido, sem notar que o fato dos acontecimentos citados se deu em um momento pré ChatGPT.

Para colocar em perspectiva os números, o TikTok [19] levou nove meses - e o Instagram quase três anos - para ganhar tantos usuários quanto o ChatGPT ganhou em menos de 90 dias.

É possível notar uma grande explosão do interesse nos termos “IA” e “ChatGPT” no último mês de Novembro, segundo os dados coletados de pesquisa da Google [20] durante o ano passado, quando o GTP 3 foi lançado ao público. Desde então, as pesquisas no Google de ambos termos tem crescido rapidamente, com outro pico em “ChatGPT” em Março de 2023, quando GPT 4 foi lançado.

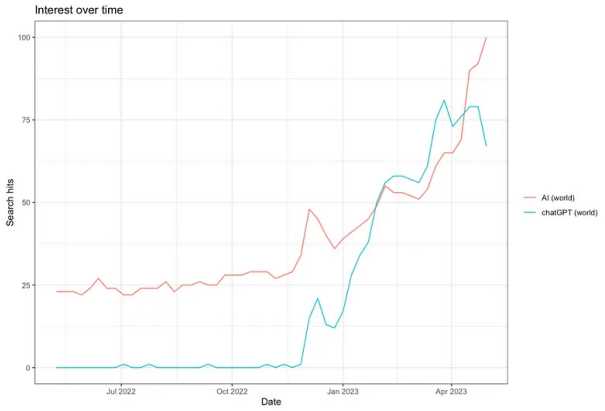

Figura 1: representação gráfica do número de vezes que os termos “AI” e “chatGPT” foram pesquisados, segundo a Google.

Para muitas pessoas que não são especialistas na área, incluindo um número crescente de empreendedores e empresários, o amigável modelo de bate-papo é uma evidência clara de que a revolução da IA tem um potencial real (MIT Technology Review, 2023) [21].

Com o surgimento do ChatGPT e o crescente interesse em IA, a questão que se coloca é: se antes do surgimento dessa tecnologia já testemunhamos casos tão absurdos envolvendo grandes corporações, como será agora, quando pessoas comuns e empresas de pequeno e médio porte estão intensificando seus esforços nessa área?

“Prometendo eficácia e imparcialidade, [os algoritmos] distorcem a educação superior, aumentam a dívida, estimulam o encarceramento em massa, golpeiam os pobres em quase todas as situações e solapam a democracia”, denuncia Cathy O'Neil [22], especialista em dados e autora do revelador livro Weapons of Math Destruction (Armas de Destruição Matemática). É de extrema importância que todos nós estejamos aptos a compreender o impacto da discriminação e do preconceito na sociedade, especialmente quando se trata de grupos mais vulneráveis, já que o viés algorítimico pode manifestar-se em várias formas de preconceito, como gênero, raça e idade. Essa compreensão é essencial para promover discussões construtivas sobre a implementação de medidas eficazes de prevenção e proteção, garantindo que os sistemas de IA sejam desenvolvidos e aplicados de forma ética e justa (Steve Nouri, 2021) [23].

O preconceito persistente na IA não é apenas um problema técnico, mas também um problema ético e social significativo. Ele tem o potencial de amplificar desigualdades e injustiças já presentes na sociedade, impactando negativamente a vida de indivíduos e comunidades marginalizadas. Segundo Steven Lockey et al. (2021) [24], a utilização em escala de tecnologias de IA imprecisas, tendenciosas ou que invadem a privacidade dos cidadãos pode consolidar o preconceito, a desigualdade e minar os direitos humanos, como o direito à privacidade. David Gohl [25], estrategista de governança, líder de risco e conformidade da Amazon, se posiciona de forma parecida quando afirma que: “decisões tendenciosas e avaliações de riscos imprecisas podem levar a perdas financeiras, danos à reputação e implicações legais.“

Steven Lockey et al. (2021) [21], ao falar dos desafios sobre confiança na IA, explica que as decisões feitas pela IA influenciam diretamente os usuários finais e que eles estão vulneráveis a qualquer problema, imprecisão ou viés algorítmico do sistema:

> “De forma mais ampla, os usuários finais enfrentam vulnerabilidades na compreensão de como as decisões baseadas em IA são tomadas, o que pode levar à diminuição da capacidade de fornecer consentimento significativo, identificar impactos injustos ou antiéticos e exercer a agência.”

6.4 Sobre o Viés Algorítmico

Segundo a Wikipedia: “O viés manifesta-se como uma inclinação irracional a atribuir um julgamento mais favorável ou desfavorável a alguma coisa, pessoa ou grupo”. O preconceito humano é uma questão que tem sido bem pesquisada na psicologia há anos. Surge da associação implícita que reflete preconceitos dos quais não temos consciência e como podem afetar os resultados de um evento (Steve Nouri, 2016) [20].

Em um artigo de opinião, Leonardo Batista [26], membro do Instituto Líderes do Amanhã, discute sobre o viés algorítmico dizendo que ele ocorre porque a IA aprende a partir de enormes conjuntos de dados da vida real, incorporando em seus

aprendizados preconceitos e desigualdades. O viés algorítmico presente na IA ocorre porque os seres humanos escolhem os dados que os algoritmos usam e também decidem como os resultados desses algoritmos serão aplicados (Bernard Marr, 2022) [7]. De acordo com David Gohl [22], o viés algorítmico na IA pode se manifestar de várias formas, e essas são algumas das categorias mais comuns:

- Viés de Representação (Representation Bias): Isso ocorre quando os dados usados para treinar modelos de IA não representam adequadamente a diversidade da população ou do problema que estão tentando resolver, resultando, por exemplo, em um viés racial.

- Viés de Desempenho (Performance Bias): Isso ocorre quando um modelo de IA funciona bem para um grupo de pessoas, mas não para outros devido a características específicas. Por exemplo, um sistema de IA de saúde que é altamente preciso no diagnóstico de doenças em homens, mas menos preciso em mulheres.

- Viés de Confirmação (Confirmation Bias): Isso ocorre quando os algoritmos de IA são projetados para favorecer informações que confirmam crenças ou estereótipos preexistentes. Por exemplo, um algoritmo de recomendação de notícias que tende a mostrar notícias que reforçam as opiniões políticas do usuário, em vez de fornecer uma visão imparcial.

- Viés de Dados (Data Bias): Isso acontece quando os dados usados para treinar modelos de IA contêm preconceitos ou refletem desigualdades existentes na sociedade. Por exemplo, se os dados de treinamento de um modelo de contratação refletirem desigualdades de gênero, o modelo pode continuar a perpetuar essas desigualdades, recomendando candidatos do sexo masculino com mais frequência do que candidatas do sexo feminino.

Os vieses já existem antes mesmo do surgimento da IA e que essas questões se potencializam após a adoção dessa tecnologia na sociedade (Javier Salas, 2018) [13]. Mesmo que variáveis como gênero, idade, etnia sejam excluídos, a IA pode aprender a tomar decisões baseadas no conjunto de dados usados em seu treinamento, o que pode conter decisões humanas que representam desigualdades sociais e históricas (Steve Nouri, 2016) [20].

Neste contexto, os preconceitos da IA podem ser explicados pelo fato da informação ser passada do ser humano para o dados que serão usados no treinamento, enquanto a programação e codificação do processamento de dados desenvolve a questão do racismo e da discriminação (Penny, 2017) [27]. Portanto, o viés algorítmico é uma anomalia dos algoritmos de aprendizado de máquina causada por suposições pré-concebidas feitas durante a fase de desenvolvimento do algoritmo ou determinados conjuntos de dados de treinamento (Dilmegani, 2022) [28].

6.5 O Viés Algorítmico Vem dos Humanos

Em um trecho retirado de uma publicação feita pela agência de notícias Bloomberg [29], Nicole M. Napolitano, Ph.D. em Justiça Criminal e diretora de pesquisa e estratégia do Center for Policing Equity, afirma que: “Cada parte do processo em que um humano pode ser enviesado, a IA também pode. [...] a tecnologia legitimiza o viés criando uma sensação de objetividade, quando na verdade não é isso que acontece”.

Os robôs são racistas e sexistas por um reflexo da sociedade, visto que as máquinas só podem funcionar a partir de informações fornecidas a eles, geralmente por homens brancos e heterossexuais que dominam os campos da tecnologia e da robótica (Penny, 2017) [24]. Chamorro-Premuzic et al., (2019) [30], de maneira análoga, afirma:

> “Somos rápidos em culpar a IA para prever que homens brancos recebem classificações de desempenho mais altas de seus gerentes (provavelmente também homens brancos). Mas isso está acontecendo porque falhamos em corrigir o viés nas classificações de desempenho que são frequentemente usados em conjuntos de dados de treinamento.”

Além dos algoritmos e dos dados, os pesquisadores e engenheiros que desenvolvem os algoritmos usados na IA também são responsáveis pelo seu viés. De acordo com o site VentureBeat, um estudo conduzido pela Universidade de Columbia descobriu que “quanto mais homogênea for a equipe [de engenharia], maior será a probabilidade de que um determinado erro de previsão apareça”.

Sobre algoritmos de inteligência artificial que gera imagens a partir de texto, Steed et al. (2021) [31] explica que o preconceito não se reflete apenas nos padrões de linguagem, mas também nos conjuntos de dados de imagens usados para treinar modelos de visão computacional, criando novas descrições e estereótipos que continuam reforçando preconceitos já presentes em relação a grupos sociais e que influenciam ainda mais o pensamento humano.

Devido à disseminação generalizada da tecnologia humanizada entre consumidores e profissionais de TI, torna-se essencial compreender como os preconceitos humanos podem ser inadvertidamente incorporados ao design da inteligência artificial (Jobin et al., 2019) [32]. Por isso é importante atentar para a adoção dessa tecnologia no dia a dia, não apenas pelos algoritmos produzidos e inseridos em produtos que serão comercializados por grandes empresas, mas por todo algoritmo em que o viés possa ser danoso, principalmente nos que atuam em áreas da segurança, saúde, finanças, educação e justiça.

6.6 Como Evitar ou Combater o Viés Algorítmico

Para evitar ou amenizar o dano à sociedade causado pelo viés algorítmico, é importante explorarmos estratégias e práticas para minimizar-los em sistemas de IA, como: a) entender a importância do uso de dados diversificados; b) adoção de algoritmos transparentes e avaliações regulares; c) discutir como a educação, conscientização e regulamentações desempenham um papel vital na promoção de sistemas de IA justos e éticos.

Nos parágrafos a seguir serão discutidas algumas estratégias para minimizar a atuação de algoritmos viesados em sistemas de IA. Para isso, é fundamental pensar na difusão do conhecimento sobre

o tema na sociedade, em formas estratégicas de compor equipes, na participação da esfera legislativa do Governo, além da garantia da transparência e interpretabilidade dos algoritmos juntamente com auditoria e teste regulares.

1. Regulamentação:

   Steed, R. et al. (2021) [28], afirma que a aplicação de sistemas inteligentes em contextos sociais sem regulamentação adequada, compreensão científica e sensibilização da sociedade levanta sérias questões éticas e de segurança. Reconhecendo a necessidade de medidas proativas, Sam Altman [22], CEO da OpenAI, testemunhou recentemente perante o Congresso dos Estados Unidos, enfatizando a importância da intervenção governamental na mitigação dos riscos relacionados com a IA. Altman delineou um plano de três pontos, incluindo a criação de uma nova agência governamental responsável pelo licenciamento de grandes modelos de IA, o desenvolvimento de padrões de segurança e o mandato para auditorias independentes para avaliar o desempenho do modelo de IA.

2. Diversificar equipes de engenharia de software:

   Pensar em formar equipes diversificadas para analisar o contexto dos dados utilizados para treinar algoritmos de IA ou avaliar as saídas de sistemas de IA já implementados representa uma abordagem eficaz na redução dos efeitos prejudiciais do viés algorítmico. A diversidade dessas equipes contribui para a identificação e mitigação de preconceitos, promovendo uma IA mais justa e imparcial em benefício de todos. A falta de uma equipe de software heterogênea (com a presença de grupos sociais diversos), cria uma falta de empatia pelas pessoas que enfrentam problemas de discriminação, levando a uma introdução inconsciente de preconceitos nestes sistemas de IA (Steve Nouri, 2021) [20].

3. Ampliar a discussão para comunidade de programadores:

   Por fim, também é importante ter engenheiros de softwares que reconheçam que mitigar o viés algorítmico na IA é um esforço contínuo e que estejam dispostos a ajustar os sistemas à medida que novas questões surgem e mais dados estão disponíveis. Joy Buolamwini, fundadora do Algorithmic Justice League e pesquisadora graduada pelo MIT Media Lab [33], decidiu trabalhar na criação de ferramentas que identifiquem preconceitos em softwares de reconhecimento facial. Segundo ela, “não existe nenhum serviço que monitore consistentemente a precisão do software de reconhecimento facial ou a diversidade dos dados usados no treinamento”.

7. AGRADECIMENTOS

Agradeço primeiramente meus pais que nunca mediram esforços para tornar a educação de seus filhos um pilar fundamental.

Agradeço a generosidade e compreensão da minha orientadora, Francilene Garcia, pois seu envolvimento neste trabalho foi fundamental para conclusão e publicação do mesmo.

Agradeço também aos colegas e amigos que permitiram toda essa jornada ser realizada com mais leveza.

Aos demais que fizeram essa etapa da minha vida possível, meus sinceros agradecimentos.

8. TRABALHOS FUTUROS

Uma das grandes promessas com potencial para aliviar os danos causados por sistemas viesados de IA é a regulamentação. Portanto, aprofundar novas pesquisas conforme a legislação no Brasil e em outros países sobre esse tema for se consolidando é importante como ponto de expansão deste artigo.

Avaliar o impacto do viés algorítmico em áreas específicas da sociedade como, educação, saúde e justiça, enriquecendo o trabalho com conteúdo de áreas multidisciplinares, com certeza seria um ponto de expansão notável.

Por fim, este artigo pode ser expandido para discutir avanços técnicos que solucionem o viés algorítmico, como algoritmos avançados que evitem a entrada de dados viesados automaticamente, aprimoramento de técnicas de pré-processamento de dados, e uma pesquisa de técnicas de treinamento de modelos que garantam que eles sejam equitativos em relação a diferentes grupos demográficos.

9. REFERÊNCIAS

[1] PAZZANESE, C. (2020) Ethical concerns mount as AI takes bigger decision-making role in more industries. The Harvard Gazette. Available at: https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/.

[2] SICHMAN, J.S. (2021) Inteligência Artificial e sociedade: avanços e riscos. 2021, v. 35, n. 101. Universidade de São Paulo, Escola Politécnica, Departamento de Engenharia de Computação e Sistemas Digitais, São Paulo, Brasil.

[3] RĄB-KETTLER, K., and LEHNERVP, B. (2019) Recruitment in the times of machine learning. Management Systems in Production Engineering, 27, 105–109.

[4] RAGHAVAN, M., BAROCAS, S., KLEINBERG, J., and LEVY, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 conference on fairness, accountability, and transparency.

[5] DRG Brasil. (2021) Inteligência Artificial na saúde e machine learning: um panorama de oportunidades médicas. Blog Valor em Saúde. Available at: https://www.drgbrasil.com.br/valoremsaude/inteligencia-artificial-na-saude/.

[6] NVIDIA. ([between 2018 and 2023]) Recommendation System. Available at: https://www.nvidia.com/en-us/glossary/data-science/recommendation-system/

[7] MARR, B. (2020) The Problem With Biased AIs (and How To Make AI Better). Forbes. Available at: https://www.forbes.com/sites/bernardmarr/2022/09/30/the-problem-with-biased-ais-and-how-to-make-ai-better/

[8] LEXALYTICS. ([between 2021 and 2023]) Bias in AI and Machine Learning: Sources and Solutions. Available at: https://www.lexalytics.com/blog/bias-in-ai-machine-learning/

[9] BBC NEWS. (2019) Apple's 'sexist' credit card investigated by US regulator. Available at: https://www.bbc.com/news/business-50365609

[10] HALE, K. (2021) A.I. Bias Caused 80% Of Black Mortgage Applicants To Be Denied. Available at: https://www.forbes.com/sites/korihale/2021/09/02/ai-bias-caused-80-of-black-mortgage-applicants-to-be-denied/

[11] HARRIS, V., MAYOR, P. (2020) Wrongfully Arrested Because Face Recognition Can't Tell Black People Apart. American Civil Liberties Union (ACLU). Available at: https://www.aclu.org/news/privacy-technology/wrongfully-arrested-because-face-recognition-cant-tell-black-people-apart

[12] PEREDA, C. (2016) O Google é racista?. El País. Available at: https://brasil.elpais.com/brasil/2016/06/10/tecnologia/1465577075_876238.html

[13] SALAS, J. (2018) Google conserta seu algoritmo "racista" apagando os gorilas. El País. Available at: https://brasil.elpais.com/brasil/2018/01/14/tecnologia/1515955554_803955.html

[14] VEJA. (2016) Exposto à internet, robô da Microsoft vira racista em 1 dia. Available at: https://veja.abril.com.br/tecnologia/exposto-a-internet-robo-da-microsoft-vira-racista-em-1-dia

[15] INDEPENDENT. (2016) Tay Tweets: Microsoft shuts down AI chatbot turned into a pro-Hitler racist troll in just 24 hours. Available at: https://www.independent.co.uk/tech/tay-tweets-microsoft-ai-chatbot-posts-racist-messages-about-loving-hitler-and-hating-jews-a6949926.html

[16] SHARP, G. (2009) Nikon Camera Says Asians: People are Always Blinking. The Society Pages. Available at: https://thesocietypages.org/socimages/2009/05/29/nikon-camera-says-asians-are-always-blinking/

[17] EPOCA NEGOCIOS. (2018) Amazon desiste de ferramenta secreta de recrutamento que mostrou viés contra mulheres. Available at: https://epocanegocios.globo.com/Empresa/noticia/2018/10/amazon-desiste-de-ferramenta-secreta-de-recrutamento-que-mostrou-vies-contra-mulheres.html

[18] GIZ. (2019) Como um algoritmo médico nos EUA era preconceituoso com pacientes negros. Available at: https://gizmodo.uol.com.br/algoritmo-medico-eua-preconceito-negros/

[19] DIGITAL STRIKE. (2023) ChatGPT: the recent rise of ai & what it all means. Available at: https://www.digitalstrike.com/chatgpt-and-the-recent-rise-of-ai/

[20] MEDIUM. (2023) How has interest and coverage around AI changed since ChatGPT?. Available at: https://medium.com/@chloeng_22909/how-has-interest-and-coverage-around-ai-changed-since-chatgpt-502d415a3126

[21] MIT TECHNOLOGY REVIEW. (2023) O ChatGPT está prestes a causar uma revolução na economia. Precisamos decidir o que isso irá significar Available at: https://mittechreview.com.br/o-chatgpt-esta-prestes-a-causar-uma-revolucao-na-economia-precisamos-decidir-o-que-isso-ira-significar/

[22] EL PAÍS. (2017) Se está na cozinha, é uma mulher: como os algoritmos reforçam preconceitos. Available at: https://brasil.elpais.com/brasil/2017/09/19/ciencia/1505818015_847097.html

[23] FORBES. (2021) The Role Of Bias In Artificial Intelligence. Available at: https://www.forbes.com/sites/forbestechcouncil/2021/02/04/the-role-of-bias-in-artificial-intelligence/

[24] LOCKEY, S., GILLESPIE, N., HOLM, D., and SOMEH, I.A. (2021) A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities and Future Directions. Hawaii International Conference on System Sciences, pp 5469.

[25] GOHL, D. (2023) The Impact of Bias in AI on Risk Management: Navigating Challenges for Fair and Effective Practices. Available at: https://www.linkedin.com/pulse/impact-bias-ai-risk-management-navigating-challenges-fair-david-gohl/

[26] BATISTA, L. (2023) O viés algorítmico: Os desafios da Inteligência Artificial. Available at: https://www.lideresdoamanha.org.br/post/o-viés-algorítmico-os-desafios-da-inteligência-artificial

[27] PENNY, Laurie. (2017) Robots are racist and sexist. Just like the people who created them. The Guardian. Available at: https://www.theguardian.com/commentisfree/2017/apr/20/robots-racist-sexist-people-machines-ai-language

[28] DILMEGANI, C. (2022) Bias in AI: What it is, Types, Examples & 6 Ways to Fix it in 2022. AIMultiple. Available at: https://research.aimultiple.com/ai-bias/

[29] NICOLETTI, L., and BASS, D. (2022) Humans are biased. Generative ai is even worse. Bloomberg Technology. Available at: https://www.bloomberg.com/graphics/2023-generative-ai-bias/

[30] Chamorro-Premuzic, T., Polli, F., & Dattner, B. (2019). Building ethical AI for talent management. Harvard Business Review, 21.

[31] STEED, R., CALISKAN A. (2021) Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases. Association for Computing Machinery, New York, NY, USA, 701–713. Available at: https://doi.org/10.1145/3442188.3445932

[32] Jobin, A., Ienca, M. & Vayena, E. (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1, 389–399. Available at: https://doi.org/10.1038/s42256-019-0088-2

[33] MEDIUM. (2016) The Algorithmic Justice League. Available at: https://medium.com/mit-media-lab/the-algorithmic-justice-league-3cc4131c5148