As Causas Sistêmicas por trás do Viés de Gênero em IA: Um Mapeamento Sistemático da Literatura.

Rafaela Toledo Dolabella[1], **Thais Regina de Moura Braga Silva**[1], **Gláucia Braga e Silva**[1] **e Estela Miranda Batista**[1]

[1]Instituto de Ciências Exatas e Tecnológicas
Universidade Federal de Viçosa (UFV)
Florestal - MG - Brazil

{rafaela.dolabella,thais.braga,glaucia,estela.batista}@ufv.br

Abstract. *Artificial Intelligence (AI) has been transforming daily life, impacting critical decisions in various fields such as healthcare, security, and the labor market. However, its use raises ethical and social concerns, particularly regarding biases present in AI systems. This study specifically addresses gender bias, investigating the systemic causes (data, algorithmic bias, or user behavior) behind this issue through a Systematic Literature Mapping (SLM). The research analyzed 176 studies from the IEEE Xplore, ACM Digital Library, and Scopus databases, of which 12 were deemed relevant to the study's objective. The selection process was guided by inclusion and exclusion criteria. The analysis identified that the primary cause of gender bias is the use of biased data, amplified by algorithmic bias in half of the studies. User behavior was minimally explored, being the main focus in only one article.*

Resumo. *A Inteligência Artificial (IA) tem transformado o cotidiano das pessoas, impactando decisões críticas em diversas áreas, como saúde, segurança e mercado de trabalho. Contudo, seu uso levanta preocupações éticas e sociais, especialmente no que diz respeito aos vieses presentes em sistemas de IA. Este trabalho aborda especificamente o viés de gênero, investigando as causas sistêmicas (dados, viés algorítmico e comportamento do usuário) por trás desse problema por meio de um Mapeamento Sistemático da Literatura (MSL). A pesquisa analisou 176 estudos provenientes das bases IEEE Xplore, ACM Digital Library e Scopus, dos quais 12 foram considerados relevantes para o objetivo do trabalho. A seleção foi orientada por critérios de inclusão e exclusão. A análise identificou que a maioria dos artigos encontrados aborda o uso de dados enviesados como causa do viés de gênero, seguido do viés algorítmico, encontrado em metade dos estudos. O comportamento do usuário ainda é pouco explorado, aparecendo como foco principal em apenas um trabalho.*

1. Introdução

A Inteligência Artificial (IA) está cada vez mais presente no cotidiano das pessoas, seja por meio de chatbots, como o ChatGPT[1], assistentes virtuais, como a Alexa[2], ou sistemas de tomada de decisão aplicados a processos de contratação, diagnósticos médicos e

[1]https://openai.com/index/chatgpt/
[2]https://www.amazon.com.br

segurança pública. Desse modo, é notável o quanto a IA causa um impacto significativo nas rotinas diárias das pessoas, influenciando opiniões, comportamentos e até mesmo decisões críticas que afetam diretamente a vida daqueles submetidos aos resultados dessas tecnologias [GLOBO 2024]. Como resultado, o uso da IA tem gerado debates éticos e sociais, destacando a necessidade de garantir que esses sistemas sejam justos, inclusivos e livres de preconceitos. Casos como o dos serviços de triagem de currículos baseados em IA da Amazon e do LinkedIn, que priorizavam candidatos do sexo masculino para vagas em tecnologia da informação (TI), ilustram como o viés de gênero presente nos algoritmos pode gerar impactos negativos significativos. Esses exemplos evidenciam a urgência de abordar e mitigar os vieses nos sistemas de aprendizado de máquina, promovendo uma IA mais equitativa e responsável [Njoto et al. 2022a].

Investigar as causas por trás dos preconceitos presentes nas IAs é essencial para desenvolver soluções que mitiguem esses problemas e promovam sistemas mais justos e inclusivos. O objetivo geral deste trabalho é apresentar um Mapeamento Sistemático da Literatura (MSL) que buscou e classificou estudos primários que abordaram o viés de gênero em modelos de IA relacionado a "sistemas", utilizando as categorizações propostas em [Lima et al. 2023] e [da Silva Souza 2023], respondendo a seguinte questão de pesquisa: Quais artigos abordam as causas do viés de gênero nas 3 categorias investigadas, isto é, dados enviesados, viés algorítmico ou comportamento do usuário?

O trabalho de [Lima et al. 2023] identifica seis categorias principais de causas do viés de gênero: sistemas, preconceito, cultura, desigualdade, relacionamento e interação. Especificamente na categoria "sistemas", os autores destacam como decisões técnicas e sociais ao longo do desenvolvimento do modelo de IA podem perpetuar esses vieses. Por sua vez, os autores de [da Silva Souza 2023] argumentam que o viés sistêmico pode se manifestar em diferentes etapas do processo de desenvolvimento, abrangendo desde os dados utilizados no treinamento até o próprio algoritmo (viés algorítmico) e o comportamento dos usuários.

Os resultados deste trabalho poderão contribuir com uma visão geral da literatura atual sobre as causas sistêmicas do viés de gênero, auxiliando no avanço de estratégias de mitigação no desenvolvimento de IAs. O restante deste artigo está organizado da seguinte maneira: a Seção 2 apresenta os trabalhos relacionados encontrados na literatura; a Seção 3 descreve a metodologia adotada neste MSL, detalhando os métodos utilizados para a escolha dos estudos primários; a Seção 4 apresenta os principais achados do estudo; e, por fim, a Seção 5 traz as conclusões do trabalho, sintetizando os resultados, destacando suas implicações e sugerindo possíveis direções para pesquisas futuras.

2. Trabalhos Relacionados

No que diz respeito a trabalhos relacionados a revisões e mapeamentos sistemáticos ou *ad hoc* que abordam o viés de gênero nas IAs, [Hall and Ellis 2023] realizaram uma Revisão Sistemática da Literatura, com uma busca inicial de 3401 artigos retornados, e, ao final, conduziu uma análise de 64 artigos selecionados. Tais estudos foram encontrados nas bases de dados IEEE Xplore, ACM Digital Library, SCOPUS e Web of Science e se concentraram em artigos de 2013 até 2023. O objetivo do estudo foi categorizar as causas sociais do preconceito de gênero nos algoritmos de IA, delinear as consequências sociais deste preconceito e explorar soluções sociais propostas. O artigo destacou que, na maioria

dos estudos analisados, as principais causas apontadas foram o design algorítmico e os dados enviesados, destacando as causas sociais por trás disso. Dentre as consequências mais abordadas pelos estudos primários está a amplificação de desigualdades, criando ciclos de feedback onde resultados enviesados reforçam o viés original.

[Nadeem et al. 2022] apresentaram uma revisão *Ad Hoc* que capturou 6024 artigos, sendo 31 destes selecionados ao final a partir da base de dados Scopus, com filtro de tempo de 2010 a 2020. Seu principal objetivo foi explorar o preconceito de gênero em sistemas de tomada de decisão baseados em IA, identificar e examinar as características desse preconceito, investigar os fatores que contribuem para sua ocorrência e as estratégias relatadas para sua mitigação. Ao considerar os fatores que contribuem para o preconceito de gênero nos sistemas de tomada de decisão baseados em IA, a partir dos estudos primários, foram estabelecidos temas principais, como estereótipos de gênero, conjuntos de dados tendenciosos, falta de diversidade de gênero nas equipes e falta de regulamentação. Por fim, os autores descrevem abordagens para mitigar o viés de gênero nas IAs, como alteração da tecnologia e abordagens justas de gestão.

Por fim, [Nadeem et al. 2020] apresentam uma Revisão *Ad Hoc* que analisou 34 artigos encontrados via Google Scholar e cujo objetivo foi explorar o conceito de preconceito de gênero na IA, identificando os fatores que contribuem para sua ocorrência, os impactos relacionados a esse viés e as abordagens potenciais para sua mitigação. O estudo estabeleceu os principais fatores contribuintes para o preconceito de gênero na IA, como a falta de diversidade nos dados de treinamento e na equipe de desenvolvedores, preconceito na sociedade e nos dados, fatores econômicos e decisões tendenciosas. Embora o viés de gênero em IA possa ser influenciado por limitações técnicas e arquitetônicas dos sistemas, as causas identificadas estão mais associadas a fatores sociais e culturais do que a aspectos puramente sistêmicos. Por fim, o estudo apontou três abordagens principais para lidar com o preconceito de gênero na IA: implementar um desenvolvimento de IA justo e ético, reduzir o preconceito no algortimo e garantir a diversidade na equipe de desenvolvimento de IA.

Neste MSL foram analisadas apenas as causas sistêmicas do viés de gênero, diferentemente dos estudos citados acima, que abordam mais de um tipo de causa, como sociais, culturais e de regulamentação. Além disso, diferentemente dos estudos apresentados, neste trabalho não foram analisadas formas de se diminuir o viés de gênero nas IAs e sim os fatores por trás da sua ocorrência, de forma a proporcionar uma compreensão mais aprofundada dos mecanismos sistêmicos que contribuem para a sua perpetuação. Ademais, neste estudo o MSL foi sistemático, enquanto [Nadeem et al. 2022] e [Nadeem et al. 2020] não apresentaram um protocolo para a realização da revisão realizada. Por fim, não foi utilizado um filtro temporal para o retorno dos estudos primários neste artigo, enquanto em [Hall and Ellis 2023] e [Nadeem et al. 2022] os recortes foram de 2013 a 2023 e 2010 a 2020, respectivamente.

3. Metodologia

Neste estudo, o MSL realizado abrangeu três fases: planejamento, condução e relatório. Na etapa de planejamento, foi elaborado um protocolo que definiu as questões de pesquisa, os critérios de inclusão e exclusão, as bases de dados, a *string* de busca e os procedimentos de revisão. Na fase de condução, realizou-se a seleção dos estudos, a avaliação

Referência	Metodologia	Objetivo	Causas abordadas
[Hall and Ellis 2023]	Revisão Sistemática da Literatura	Categorizar causas sociais e explorar soluções para o viés de gênero em IA.	Design algorítmico e dados enviesados causados por fatores sociais.
[Nadeem et al. 2022]	Revisão *Ad Hoc*	Identificar características e estratégias para mitigar o preconceito de gênero em IA.	Estereótipos de gênero, conjuntos de dados tendenciosos, falta de diversidade de gênero nas equipes e falta de regulamentação.
[Nadeem et al. 2020]	Revisão *Ad Hoc*	Explorar fatores, impactos e estratégias para mitigar o viés de gênero na IA.	Falta de diversidade nos dados de treinamento e na equipe de desenvolvedores, preconceito na sociedade e nos dados, fatores econômicos e decisões tendenciosas.
Este trabalho	Mapeamento Sistemático da Literatura	Classificar estudos sobre causas do viés de gênero em IA, abordando dados, algoritmos e comportamento do usuário.	Dados enviesados, viés algorítmico e comportamento do usuário.

Tabela 1. Trabalhos relacionados.

de sua qualidade, a extração e o monitoramento dos dados, além da síntese dos resultados. Por fim, na fase de relatório, foram apresentados os principais achados e divulgadas as conclusões obtidas.

A *string* de busca definida neste trabalho e apresentada na Tabela 2 possui cinco aspectos principais. O primeiro deles engloba o viés de gênero em si, o segundo aborda termos sinônimos à causas, o terceiro é relacionado à IA, o quarto trata das causas sistêmicas (dados, viés algorítmico e comportamento do usuário) e o quinto traz a tentativa de excluir ao máximo artigos que abordem outros tipos de vieses, como o racial.

Aspectos	**Palavras-Chave**
Viés de gênero	*"Gender bias"*
Sinônimos de causas	*"Cause"*, *"Source"*, *"Origin"*, *"Foundation"*, *"Factor"*, *"Reason"*
Inteligência Artificial	*"AI"*, *"artificial intelligence"*, *"machine learning"*
Causas sistêmicas	*"system"*, *"data"*, *"algorithmic bias"*, *"algorithm"*, *"program"*, *"user interaction"*, *"user behavior"*
Exclusão de outros vieses	*"race"*, *"ethnicity"*
***String* de Busca:**	
("Gender Bias") AND ("Cause"OR "Source"OR "Origin"OR "Foundation"OR "Factor"OR "Reason") AND ("AI"OR "artificial intelligence"OR "machine learning") AND ("system"OR "data"OR "algorithmic bias"OR "algorithm"OR "program"OR "user interaction"OR "user behavior") NOT ("race"OR "ethnicity")	

Tabela 2. *String* de busca usada neste MSL.

Para atingir o objetivo geral desta pesquisa, as seguintes questões de pesquisa foram especificadas:

- QP1:Qual o panorama geral referente aos estudos analisados, em termos de nacionalidade dos autores, ano de publicação, aplicações e modelos de IA e os impactos

abordados?

- QP2:Qual a distribuição de gênero das pessoas autoras dos trabalhos?
- QP3:Quais artigos abordam as causas do viés de gênero nas 3 categorias investigadas, isto é, dados enviesados, viés algorítmico ou comportamento do usuário?

Para a seleção dos trabalhos, foram empregados 1 Critério de Inclusão (CI) e 3 Critérios de Exclusão (CE). O critério de inclusão em questão é: (CI) O artigo aborda causas sistêmicas (dados, viés algorítmico e comportamento do usuário) de viés de gênero em modelos de IA. Os critérios de exclusão são: (CE_1) O estudo ser publicado como resumo; (CE_2) O estudo ser uma versão mais antiga de outro estudo; (CE_3) O artigo completo não estar disponível.

Por fim, as bases de dados selecionadas para encontrar os estudos foram: IEEE Xplore[3], ACM Digital Library[4] e Scopus[5]. As buscas abrangeram os resumos, títulos e palavras-chave dos estudos nas três bases de dados. Com o intuito de encontrar mais trabalhos em português, foi utilizada a mesma *string* de busca na base da SOL/SBC[6]. No entanto, não foram encontrados resultados. Diante disso, a *string* de busca foi simplificada para: ((Gender Bias Cause OR Source OR Reason AI OR machine learning system OR data OR algorithmic bias OR algorithm OR program OR user)) NOT race OR ethnicity). Com essa alteração, quatro trabalhos foram retornados, mas nenhum deles atendeu ao critério de inclusão (CI).

Na base de dados IEEE Xplore, foram analisados 64 artigos, dos quais 60 foram excluídos por não atenderem ao Critério de Inclusão (CI), e 1 foi excluído por não se enquadrar no (CE_1). Ao final, 3 trabalhos foram considerados relevantes. Na ACM Digital Library, dos 65 artigos analisados, 10 foram excluídos por não atenderem ao (CE_3), 27 por não se adequarem ao (CI), 24 por ambos os critérios (CE_3 e CI), e 1 por não atender ao (CE_1). No total, 3 trabalhos foram considerados relevantes. Na Scopus, dos 43 artigos analisados, 1 foi excluído por não atender ao (CE_1), 7 por não atenderem ao (CE_3), 20 por não atenderem ao (CI), 2 por não atenderem aos critérios (CE_3 e CI), e 4 por não atenderem os critérios (CE_1, CE_3 e CI). Ao final, 9 artigos foram considerados relevantes, 3 deles sendo trabalhos relacionados. Em resumo, foram analisados 176 artigos, dos quais 16 foram selecionados, após a aplicação dos critérios de exclusão nas três bases de dados e uma leitura detalhada dos estudos, 15 artigos foram selecionados como relevantes, sendo 3 classificados como trabalhos relacionados, por serem RSL ou MSL, e 12 como estudos primários, conforme ilustra a Figura 1.

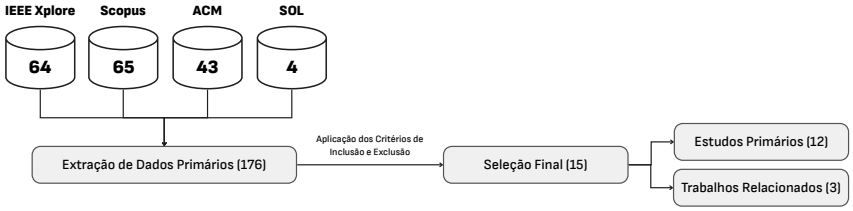


Figura 1. Sumarização de trabalhos por base

[3]https://ieeexplore.ieee.org/Xplore/home.jsp
[4]https://dl.acm.org/
[5]https://www.elsevier.com/products/scopus
[6]https://sol.sbc.org.br/index.php/indice

Após as etapas descritas na metodologia desta pesquisa foram selecionados 12 trabalhos, descritos na Tabela 3.

ID	Ano	Referência	Descrição
T1	2022	[Njoto et al. 2022b]	O viés de gênero é introduzido na automação de contratações ao observar um painel humano e desenvolver um protótipo algorítmico. São identificadas fontes de viés humano e algorítmico, com propostas de mitigação que combinam ciências sociais e de dados.
T2	2023	[Sogancioglu et al. 2023]	O viés de gênero em *embeddings* usados em tarefas de saúde mental varia conforme os dados e o tipo de *embedding*. A troca de palavras de gênero mostrou-se eficaz na redução do viés.
T3	2021	[Dervişoğlu and Fatih Amasyali 2021]	O viés de gênero em análise de sentimentos foi comparado entre dados em turco e inglês, explorando suas causas e testando métodos de mitigação, como *word2vec*, *fasttext*, *T5* do Google e modelos *LSTM*, para análise comparativa dos resultados.
T4	2021	[Cho et al. 2021]	O viés de gênero em traduções entre diferentes idiomas é analisado com foco na generalização para línguas menos exploradas. A precisão de inferência de gênero e a fluência da tradução foram avaliadas em alemão, coreano, português e tagalo.
T5	2023	[Kopeinik et al. 2023]	Os usuários refletem vieses de gênero em interações com sistemas algorítmicos, especialmente ao formular consultas de busca *online*.
T6	2023	[Ghosh and Caliskan 2023]	A precisão do *ChatGPT* na tradução entre inglês e línguas com pronomes neutros, como o bengali, revela vieses de gênero semelhantes aos de outras ferramentas de tradução, destacando a importância de uma abordagem centrada no humano para o desenvolvimento futuro de IA.
T7	2023	[Parreira et al. 2023]	Viés de gênero foi identificado em robôs ouvintes, que ofereceram mais *feedback* a homens. As causas foram analisadas e soluções propostas para evitar comportamentos tendenciosos em robôs sociais.
T8	2023	[Cheong et al. 2023]	O viés de gênero em algoritmos e dados de saúde mental é analisado, destacando a pouca atenção dada à justiça no aprendizado de máquina na área. Foram confirmados vieses em três conjuntos de dados e nos algoritmos, com estratégias de mitigação avaliadas nas etapas de pré-processamento, processamento e pós-processamento.
T9	2022	[Matthews et al. 2022]	Viés de gênero e racial em *embeddings* de palavras em textos jurídicos dos EUA persiste mesmo sem dados históricos. O estudo propõe uma abordagem para detecção e estratégias de mitigação para reduzir o impacto em ferramentas de pesquisa jurídica.
T10	2022	[Rizhinashvili et al. 2022]	O estudo apresenta um método para reduzir o viés de gênero em processamento de fala, eliminando o parâmetro de gênero. A técnica transforma a voz em um tom neutro, tornando o gênero indistinguível para humanos e IA. A rede *Wav2Vec* foi utilizada para validar a neutralização. O sistema pode atuar como pré-processamento para treinamento de modelos, eliminando o viés de gênero.
T11	2019	[Brunet et al. 2019]	Algoritmos de *word embedding* frequentemente exibem vieses estereotipados, como o viés de gênero, que podem ser amplificados em sistemas de aprendizado de máquina. Este estudo apresenta uma técnica para entender como esses vieses surgem durante o treinamento, analisando como alterações no *corpus* de treinamento afetam o viés do *embedding*.
T12	2018	[Thelwall 2018]	É investigado se o aprendizado de máquina induz vieses de gênero, favorecendo autores masculinos ou femininos, e se treinar variantes separadas para cada gênero poderia melhorar a precisão em análises de sentimentos.

Tabela 3. Descrição dos artigos analisados sobre viés de gênero em IA.

4. Resultados

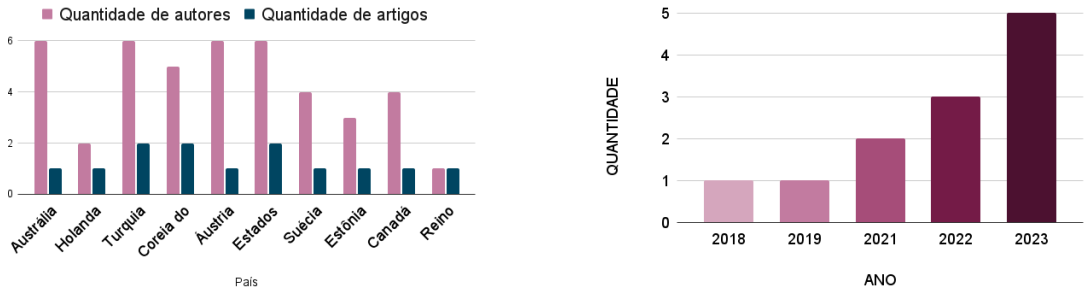
Esta seção apresenta os principais achados da pesquisa, com base nas respostas das 3 questões de pesquisa, juntamente com uma análise dos artigos selecionados.

4.1. QP1:Qual o panorama geral referente aos estudos analisados, em termos de nacionalidade dos autores, ano de publicação, aplicações e modelos de IA e os impactos abordados?

Inicialmente, foi analisada a nacionalidade dos autores, bem como a quantidade de artigos por país. A partir da Figura 2(a), é possível perceber que a maior parte dos autores são australianos, turcos, austríacos e estadunidenses. Entretanto, existe apenas 1 artigo da Austrália e da Áustria, enquanto tem-se 2 artigos da Turquia, Coreia do Sul e Estados

Unidos. Um ponto a ser destacado, é a falta de artigos que contemplem este tema no Brasil e em outros países da América do Sul.

Além disso, foi feita uma análise quanto ao ano de publicação de cada estudo selecionado. Observando a Figura 2(b), é possível perceber que os estudos analisados estão concentrados entre os anos de 2018 e 2023, possuindo um crescimento explícito dos estudos ao longo deste período, com a maioria das publicações ocorrendo em 2023. Esse panorama evidencia que o tema do viés de gênero em Inteligências Artificiais é atual e possui um grande potencial de crescimento, refletindo a relevância e a urgência dessa discussão na sociedade.

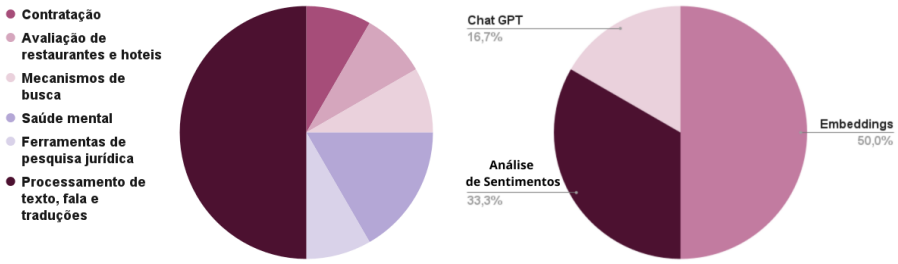


(a) Quantidade de autores X Quantidade de artigos por país

(b) Publicação dos artigos ao longo dos anos

Figura 2. Quantidade de autores por país e publicação ao longo dos anos.

Também foram analisados quais as aplicações de IA abordadas nos estudos. De acordo com a Figura 3(a), é notável que a IA para processamento de texto, fala e traduções foi a aplicação mais abordada nos artigos, abrangendo metade deles. Tal aplicação engloba tanto traduções no geral, quanto uso de robôs ouvintes e algortimos de incorporação de palavras. Na sequência, tem-se 2 artigos aplicando IA no contexto de saúde mental, 1 em contratação, 1 nas avaliação de restaurantes e hotéis, 1 em mecanismo de busca e 1 de ferramenta de pesquisa jurídica.



(a) Aplicações de IA

(b) Modelos de IA

Figura 3. Aplicações e modelos de IA

Ao analisar os modelos de IA discutidos nos artigos considerados, observa-se na Figura 3(b) que apenas 6 dos 12 artigos especificam claramente tal informação. Essa ausência de detalhamento representa uma lacuna que dificulta a replicação e limita a compreensão das discussões e resultados apresentados. Para os 6 trabalhos em que os modelos

usados foram explicitados, conforme mostrado na Figura 3(b), percebe-se que apenas os modelos *Embeddings*, análise de sentimentos e ChatGPT foram utilizados, havendo uma predominância do primeiro.

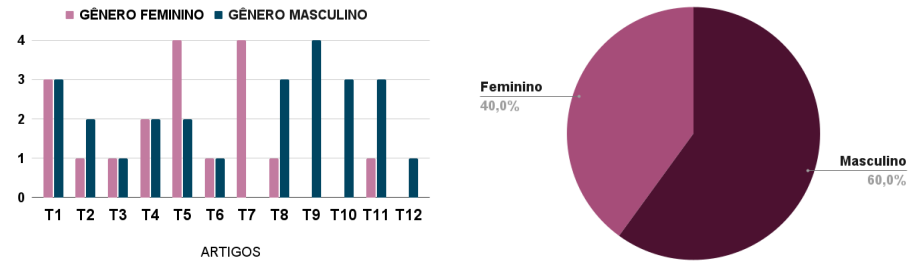
Impactos	Artigos
Decisões injustas que afetam negativamente a vida de diferentes grupos, especialmente mulheres.	T1, T2, T3, T8, T9, T11
Criação de um ambiente tecnológico prejudicial e desinformado.	T5, T7, T10, T12
Perpetuação de desigualdades linguísticas e culturais.	T4, T6

Tabela 4. Impactos abordados nos artigos analisados.

Os dados dispostos na Tabela 4 mostram que os impactos relacionados a decisões injustas feitas pelas IAs foram os mais abordados pelos estudos, destacando a preocupação crescente com as consequências éticas e sociais dessas tecnologias. Isto ocorre visto que atualmente elas estão cada vez mais presentes em tomadas de decisão e com um grande potencial de perpetuar desigualdades e discriminações em recrutamento, justiça e saúde.

4.2. QP2: Qual a ditribuição de gênero das pessoas autoras dos trabalhos?

A análise da diversidade de gêneros entre os autores dos estudos selecionados é um aspecto de grande relevância para compreender esta representatividade no campo de pesquisa sobre viés de gênero em Inteligências Artificiais. Inicialmente, foi realizado um levantamento do número total de autores de cada gênero, sendo que dos autores identificados, 24 são homens e 16 são mulheres, representando 60% e 40% do total, respectivamente. Esses dados revelam que, embora haja uma presença significativa de mulheres, os homens ainda lideram a produção científica nesse tema.



(a) Diversidade de gênero em cada estudo (b) Gênero dos autores principais

Figura 4. Diversidade de gênero dos autores

A Figura 4(a), mostra a diversidade de gênero entre os autores de cada um dos artigos analisados, em que é possível observar que o artigo T7 é o único a contar exclusivamente com mulheres como autoras. Já os artigos T1, T3, T4 e T6 apresentam uma distribuição equilibrada, com uma divisão igualitária entre autores masculinos e femininos. Já o artigo T5 destaca-se por ter mulheres como a maioria entre os autores, enquanto os demais estudos ou possuem apenas autores do gênero masculino ou apresentam uma predominância masculina na autoria.

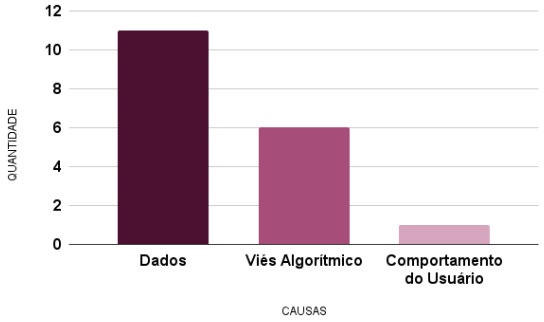


Figura 5. Causas sistêmicas mais abordadas

Na Figura 4(b), apresentam-se o gênero dos autores principais, em que os homens também predominam, sendo 7 homens e 5 mulheres no total.

Portanto, há uma disparidade de gênero dos autores com relação ao tema em questão. Sob esta ótica, a inclusão de mulheres é particularmente importante em estudos que trazem um problema que afeta diretamente a equidade de gênero, uma vez que mulheres podem trazer experiências e sensibilidades únicas que enriquecem a discussão do tema. Assim, esses dados reforçam a necessidade de mitigar a desigualdade na academia e no desenvolvimento de IA, tanto para diversificar as perspectivas nos estudos quanto para garantir que as soluções propostas sejam mais abrangentes e inclusivas.

4.3. QP3: Quais artigos abordam as causas do viés de gênero nas 3 categorias investigadas, isto é, dados enviesados, viés algorítmico ou pelo comportamento do usuário?

A Tabela 5 e a Figura 5 apresentam a distribuição dos trabalhos nas 3 categorias. Nelas observam-se que 11 dos 12 artigos abordaram dados enviesados como a principal causa sistêmica para a ocorrência do viés de gênero nas IAs. Dentro destes 11 artigos, 6 incluíam tanto dados enviesados quanto viés algorítimico. Apesar do crescente número de estudos sobre o viés de gênero em Inteligências Artificiais, existe uma lacuna significativa na literatura no que diz respeito à investigação do comportamento do usuário como uma causa sistêmica para a ocorrência desse viés, já que apenas 1 estudo abordou este aspecto como uma causa principal. Assim, é de extrema importância que o comportamento do usuário seja mais investigado pela literautra, já que é um fator complexo e difícil de se medir, visto que envolve variáveis subjetivas e dinâmicas que não podem ser plenamente controladas. Diferente de aspectos técnicos, como dados e algoritmos, que podem ser ajustados e monitorados, as interações dos usuários com sistemas de IA são imprevisíveis e influenciadas por fatores individuais, sociais e culturais.

Categorias de causas	Artigos
Dados	T1, T2, T3, T4, T6, T7, T8, T9, T10, T11, T12
Viés Algorítmico	T1, T2, T4, T8, T9, T10
Comportamento do Usuário	T5

Tabela 5. Causas do viés de gênero abordadas nos artigos analisados.

Destaca-se ainda que para entender a diversidade de gênero dos autores seguindo

as categorias estudadas, a Figura 6 mostra que a causa de comportamento do usuário foi estudada apenas por autoras femininas. Para o restante dos tipos de causas sistêmicas, os trabalhos que os abordam possuem homens como a maioria dos autores principais.

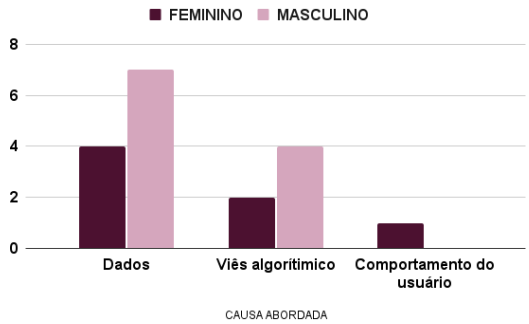


Figura 6. Autor principal versus Causa abordada

5. Conclusão

Este trabalho apresentou um MSL que resultou na seleção de 12 estudos relevantes sobre as causas sistêmicas do viés de gênero em modelos de IA. Foram destacadas as causas mais recorrentes e lacunas que ainda exigem maior atenção da comunidade científica. Os resultados oferecem uma visão geral do estado da arte, contribuindo para o entendimento do preconceito de gênero em IAs.

A maioria dos artigos aponta os dados enviesados como a principal origem do viés (11 de 12 estudos), comprometendo a imparcialidade dos sistemas desde o treinamento. Metade também discute o viés algorítmico, indicando que os algoritmos podem reforçar distorções pré-existentes. Apenas um estudo aborda o comportamento do usuário como causa relevante — justamente o único com autoria exclusivamente feminina — sugerindo que diferentes perspectivas influenciam a análise das causas e reforçando a importância da diversidade na pesquisa. Entre as ameaças à validade deste trabalho, destacam-se o número reduzido de estudos selecionados e um possível viés linguístico e geográfico, devido à formulação da *string* de busca na plataforma SOL/SBC.

Como trabalhos futuros, propõe-se a realização de revisões sistemáticas com foco em recortes mais específicos, bem como a ampliação do escopo para incluir causas sociais, culturais e regulatórias. Além disso, é necessário aprofundar a investigação sobre o impacto do comportamento do usuário, uma área ainda pouco explorada, mas essencial para o desenvolvimento de estratégias de mitigação mais eficazes.

Agradecimentos

Os autores agradecem ao CNPq [Processo 440447/2024-0] pelo apoio concedido.

Referências

Brunet, M.-E., Alkalay-Houlihan, C., Anderson, A., and Zemel, R. (2019). Understanding the origins of bias in word embeddings. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, pages 1275–1294, Long Beach, California, USA. PMLR.

Cheong, J., Kuzucu, S., Kalkan, S., and Gunes, H. (2023). Towards gender fairness for mental health prediction. In Elkind, E., editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 5932–5940. International Joint Conferences on Artificial Intelligence Organization. AI for Good.

Cho, W. I., Kim, J., Yang, J., and Kim, N. S. (2021). Towards cross-lingual generalization of translation gender bias. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 449–457, New York, NY, USA. Association for Computing Machinery.

da Silva Souza, N. C. (2023). Uma abordagem para identificação do viés de gênero em modelos de pln. Trabalho de Conclusão de Curso (MBA) -– Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

Dervişoğlu, H. and Fatih Amasyali, M. (2021). Bias detection and mitigation in sentiment analysis. In *2021 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–6.

Ghosh, S. and Caliskan, A. (2023). Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 901–912, New York, NY, USA. Association for Computing Machinery.

GLOBO (2024). Na educação, na saúde e até dentro de casa: como a inteligência artificial já faz parte da nossa rotina — g1.globo.com. https://g1.globo.com/globo-reporter/noticia/2024/11/23/na-educacao-na-saude-e-ate-dentro-de-casa-como-a-inteligencia-artificial-ja-faz-parte-da-nossa-rotina.ghtml. [Acessado em 03/02/2025].

Hall, P. and Ellis, D. (2023). A systematic review of socio-technical gender bias in ai algorithms. *Online Information Review*, 47(7):1264 – 1279.

Kopeinik, S., Mara, M., Ratz, L., Krieg, K., Schedl, M., and Rekabsaz, N. (2023). Show me a "male nurse"! how gender bias is reflected in the query formulation of search engine users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Lima, R. M. d., Pisker, B., and Correa, V. S. (2023). *Journal of Telecommunications and the Digital Economy*, 11(2):8–30.

Matthews, S., Hudzina, J., and Sepehr, D. (2022). Gender and racial stereotype detection in legal opinion word embeddings. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI 2022)*, volume 36, pages 12026–12033, Vancouver, Canada. AAAI Press.

Nadeem, A., Abedin, B., and Marjanovic, O. (2020). Gender bias in ai: A review of contributing factors and mitigating strategies. In *ACIS 2020 Proceedings - 31st Australasian Conference on Information Systems*.

Nadeem, A., Marjanovic, O., and Abedin, B. (2022). Gender bias in ai-based decision-making systems: a systematic literature review. *Australasian Journal of Information Systems*, 26. All Open Access, Gold Open Access.

Njoto, S., Cheong, M., Lederman, R., McLoughney, A., Ruppanner, L., and Wirth, A. (2022a). Gender bias in ai recruitment systems: A sociological-and data science-based case study. In *2022 IEEE International Symposium on Technology and Society (ISTAS)*, volume 1, pages 1–7.

Njoto, S., Cheong, M., Lederman, R., McLoughney, A., Ruppanner, L., and Wirth, A. (2022b). Gender bias in ai recruitment systems: A sociological-and data science-based case study. In *2022 IEEE International Symposium on Technology and Society (ISTAS)*, volume 1, pages 1–7.

Parreira, M. T., Gillet, S., Winkle, K., and Leite, I. (2023). How did we miss this? a case study on unintended biases in robot social behavior. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, page 11–20, New York, NY, USA. Association for Computing Machinery.

Rizhinashvili, D., Sham, A. H., and Anbarjafari, G. (2022). Gender neutralisation for unbiased speech synthesising. *Electronics (Switzerland)*, 11(10). All Open Access, Gold Open Access.

Sogancioglu, G., Kaya, H., and Salah, A. A. (2023). The effects of gender bias in word embeddings on patient phenotyping in the mental health domain. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8.

Thelwall, M. (2018). Gender bias in machine learning for sentiment analysis. *Online Information Review*, 42(3):343 – 354. All Open Access, Green Open Access.