Revisão Rápida sobre Vieses em Chatbots - Uma análise sobre tipos de vieses, impactos e formas de lidar

Thiago M. R. Ribeiro, Sean W. M. Siqueira, Maira G. de Bayser[1]

[1]Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Rio de Janeiro – RJ – Brasil

thiago.m.ribeiro@edu.unirio.br, sean@uniriotec.br, mgdebayser@uniriotec.br

Abstract. *Chatbots' operational mechanisms have the potential to reinforce cognitive and social biases, the consequences of which should be evaluated. To find the effects of biases in chatbots in the literature, a rapid review was conducted that involved focus groups and interviews with information and communication technology specialists in addition to a search of the SCOPUS database. Only 18 of the 488 studies that were selected were used in the final analysis. The research revealed a total of seven distinct forms of bias, along with their domains, positive and negative effects, and mitigation strategies. This research is anticipated to enhance conversational tools and assist users in identifying and mitigating biases.*

Resumo. *Devido ao seu funcionamento, chatbots podem perpetuar vieses cognitivos e sociais, cujos impactos precisam ser avaliados. Foi realizada uma revisão rápida, contemplando entrevista e grupo focal de especialistas em Tecnologia da Informação e Comunicação, além de uma busca na base SCOPUS, para identificar na literatura os impactos dos vieses em chatbots. De 488 estudos encontrados, foram selecionados 18 para a análise final. Ao todo, sete tipos de vieses diferentes emergiram dos estudos, assim como os seus impactos positivos e negativos, seus domínios e formas de mitigação. A contribuição esperada com este estudo consiste no aprimoramento de ferramentas conversacionais, bem como apoiar os usuários na identificação e mitigação de vieses.*

1. Introdução

As ferramentas de conversação, mais especificamente os agentes conversacionais ou *chatbots*, cada vez mais, fazem parte do cotidiano das pessoas. *Chatbots* também são conhecidos como entidades artificiais de conversação, agentes conversacionais interativos, *bots* inteligentes de conversação e assistentes digitais de conversação. O Dictionary[1] define *chatbot* como "um programa de computador projetado para responder com respostas conversacionais ou informativas as mensagens verbais ou escritas de usuários". Esta definição pode ser complementada com uma perspectiva das tecnologias envolvidas na implementação de *chatbots*, como a junção de um programa de Inteligência Artificial (IA) e um modelo de interação humano-computador (IHC), por meio de técnicas de processamento de linguagem natural (PLN) e análise de sentimentos para comunicação em linguagem humana textual ou oral com humanos ou outros *chatbots*

[1]https://www.dictionary.com/browse/chatbot, acessado em: 12/10/2023

[Khanna et al. 2015][Bansal and Khan 2018][Adamopoulou and Moussiades 2020]. Estas características permitem atuar de maneira colaborativa e facilitar o acesso a serviços e tarefas diversas, como agendamentos de compromissos, atendimento de suporte em serviços de telefonia/bancos, reserva de passagens aéreas, indicações de restaurantes, entre outros [Shawar and Atwell 2007].

Seja no desenvolvimento de uma ferramenta conversacional, ou em seu uso, podem ser geradas ou propagadas distorções cognitivas, também conhecidas como vieses cognitivos, que são comuns aos seres humanos. Com as plataformas de redes sociais, os vieses cognitivos ganharam destaque, juntamente com efeitos de filtros de bolha [Bakshy et al. 2015] e câmaras de eco [Bessi et al. 2016], por suas consequências negativas. Um viés cognitivo é um padrão de desvio no julgamento que ocorre em certas situações, causando distorção perceptiva, julgamento impreciso, ou o que se chama de irracionalidade [Kahneman and Tversky 1972], podendo surgir de várias fontes, incluindo memória e atenção [Pohl and Pohl 2004]. Gigerenzer e Gaissmaier (2011) descreveram o viés cognitivo como uma espécie de heurística, como uma estratégia que despreza parte da informação para agilizar a tomada de decisão.

A partir do surgimento do ChatGPT e outros *chatbots* conversacionais baseados em IA generativa, a preocupação acerca dos vieses cognitivos é ampliada. Tais *chatbots* permitem interações por meio dos *prompts*, uma interface onde comandos são digitados pelos usuários, que geram requisições ao modelo de dados e obtêm respostas que são mais elaboradas em comparação a outros tipos ferramentas conversacionais. Sendo assim, a forma como os usuários interagem com esses *chatbots* também podem perpetuar vieses. No que tange a vieses cognitivos, isso acontece por conta da tendência dos usuários de antropomorfizar os *chatbots*, tratando-os e colocando-os em lugares nos quais simbolicamente eles não pertencem [Zemčík 2021]. Neff (2016) aponta que as pessoas, apesar de saberem que não estão se comunicando com um ser humano, ainda assim, atribuem características aos *chatbots* que eles na verdade não possuem objetivamente, como inteligência e responsabilidade, demonstrando uma distorção cognitiva (viés). Além de vieses cognitivos, existe a propagação de vieses sociais, que podem amplificar preconceitos e desigualdades raciais, de gênero, classe social, religiosa, entre outras. Casos como o do *chatbot* Tay[2], que foi desativado pela Microsoft após apenas 1 dia de atividade por propagar conteúdo de cunho racista, permitiram um vislumbre dos possíveis impactos negativos que os vieses reproduzidos e amplificados por ferramentas conversacionais de IA podem causar.

O objetivo deste estudo é mapear os vieses e seus possíveis impactos em *chatbots*, incluindo os baseados em IA generativa, analisando também formas de identificação e mitigação desses vieses. Espera-se que a contribuição deste estudo, consiga colaborar na exploração da cultura como forma de transcender as fronteiras sociais (vieses); apontando possíveis ajustes necessários nos resultados que possam prejudicar grupos que demandam atenção especial e/ou minorias (vieses sociais), baseado na ética para a construção e melhoria dessas interfaces conversacionais. Para isto, foi realizada uma Revisão Rápida (*Rapid Review* - RR), contemplando uma entrevista inicial com especialistas em Tecnologias de Informação e Comunicação e um estudo conduzido na base SCOPUS.

[2]`https://www.huffpost.com/entry/microsoft-tay-racist-tweets_n_56f3e678e4b04c4c37615502`, acessado em: 09/06/2023

O restante deste documento está organizado da seguinte forma: a seção 2 apresenta os trabalhos relacionados; a seção 3 descreve a metodologia utilizada na condução desta revisão rápida; a seção 4 apresenta os resultados sobre o tema, respondendo às questões de pesquisa deste estudo; a seção 5 apresenta as discussões; e finalmente a seção 6 apresenta as conclusões sobre o tema.

2. Trabalhos Relacionados

Foram encontrados três levantamentos de literatura relacionados à temática deste artigo: [Ribeiro et al. 2023], [Navigli et al. 2023] e [Ray 2023]. Esses trabalhos foram analisados de acordo com categorias de vieses (cognitivos e sociais) e escopo (processo de desenvolvimento de software, grandes modelos de linguagem e ChatGPT). A Tabela 1 apresenta os trabalhos relacionados a este estudo, bem como a sua comparação de temas com relação ao abordado nesta RR.

Tabela 1. Trabalhos Relacionados

Trabalhos	Categorias de vieses		Escopo			
	Cognitivos	Sociais	PDS	LLM	ChatGPT	*Chatbots*
[Ribeiro et al. 2023]	X		X			
[Navigli et al. 2023]		X		X		
[Ray 2023]	X			X	X	
Este trabalho	X	X		X	X	X

Ribeiro et al. (2023) apresentaram um mapeamento sistemático de literatura sobre vieses cognitivos no processo de desenvolvimento de software, buscando entender como eles ocorrem e os possíveis problemas e impactos que podem gerar no processo de desenvolvimento de software (PDS). Foram encontrados aproximadamente 40 tipos de vieses diferentes, dentre os quais se destacaram: (i) viés de ancoragem, (ii) viés de disponibilidade e (iii) viés de confirmação. Sobre os tipos de problemas em PDS decorrentes desses vieses encontrados, se destacaram: (i) retrabalho, (ii) subestimação de tempo, (iii) problemas técnicos e (iv) tomada de decisão. Sobre as mitigações encontradas nesse contexto foram: (i) avaliações de perfil de desenvolvedores, (ii) boas práticas de desenvolvimento, (iii) consenso de grupo, (iv) cuidado com requisitos, (v) disponibilização de informações, (vi) base de conhecimento, (vii) treinamentos, (viii) testes, (ix) *debugging*, (x) desenviesamento e (xi) cuidado com o tempo. Este estudo foi escolhido por mostrar os vieses que podem impactar os desenvolvedores e afetar a construção de chatbots/ferramentas conversacionais.

Navigli et al. (2023) tratam de vieses em grandes modelos de linguagem (LLMs), que são modelos de IA treinados em vastas quantidades de dados, geralmente textuais, de modo a aprender como entender e gerar linguagem humana [Ray 2023]. O foco dos autores consistiu nos vieses sociais, termo utilizado para denominar preconceitos, estereótipos e atitudes discriminatórias contra determinados grupos de pessoas, sendo encontrados vieses de gênero, etarismo, orientação sexual, aparência física, deficiência, nacionalidade, etnia e raça, status socioeconômico, religião, cultural e viés interseccional. Embora não tenham discutido os impactos de vieses sociais em LLMs, os autores exploraram formas de lidar com estes vieses: conceituando vieses (aumentando a consciência e o conhecimento sobre os diferentes tipos de vieses), medindo vieses (nos dados de treinamento, nos modelos de linguagem resultantes e nas respectivas aplicações), entendendo os vieses (entendendo os mecanismos que dão origem a decisões tendenciosas), reduzindo os

vieses, evitando os vieses, considerando a intenção comunicativa e não apenas a forma, usando senso comum e conhecimento do mundo e aumentando a diversidade linguística e cultural. A escolha deste estudo se deveu pela abordagem de vieses sociais, que são mais abrangentes que os cognitivos, em modelos LLM, trazendo uma visão do que é entregue pelas ferramentas, com impactos desses vieses.

Ray et al. (2023), além de explorarem a aplicação do ChatGPT em vários domínios, destacam desafios, questões éticas, controvérsias e escopo futuro relacionados ao ChatGPT. Entre os vieses indicados pelos autores estão: (i) viés de gênero, raciais e culturais, (ii) viés de idioma, (iii) viés ideológico, (iv) viés sensacionalista e clickbait, (v) viés de confirmação, (vi) viés temporal, (vii) viés de exclusão, (vii) viés comercial, (ix) viés cognitivo, (x) viés de atenção, (xi) viés de formato, (xii) viés de origem, (xiii) viés de novidade, (xiv) viés de sentimento positivo/negativo, (xv) viés de outlier, (xvi) viés implícito, (xvii) viés de autoridade, (xviii) viés de atualidade, (xix) viés de pensamento de grupo, (xx) viés de ancoragem, (xxi) viés de disponibilidade, (xxii) viés de falso consenso e (xxiii) viés de retrospectiva. Esse estudo foi escolhido por apresentar uma visão complementar aos trabalhos relacionados anteriores, onde dá foco nos vieses cognitivos e nos modelos de LLM, sem tratar dos impactos e nem de mitigação de vieses.

Destaca-se, portanto, o diferencial deste trabalho ao considerar um escopo mais amplo de *chatbots*, incluindo os baseados em LLMs e o ChatGPT, e tanto vieses cognitivos quanto sociais.

3. Método de Pesquisa

Este estudo realizou uma revisão rápida (*Rapid Review* - RR) para o levantamento das informações acerca de vieses cognitivos em *chatbots*. Segundo Cartaxo et al. (2018), uma "RR é uma adaptação de revisões sistemáticas tradicionais, tendo por objetivo a aproximação das práticas metodológicas acadêmicas com os problemas e percepções reais de profissionais que os vivenciam na prática". É desejável que uma RR entregue resultados de forma mais rápida que revisões sistemáticas, o que normalmente implica em se limitar a pesquisa a uma única base científica, ser conduzida por apenas um pesquisador e não se fazer uma avaliação quanto à qualidade dos achados [Cartaxo et al. 2018]. Este trabalho executa uma RR com base em Cartaxo et al. (2018) e Rufino Júnior et al. (2022), visando identificar os trabalhos relevantes, bem como extrair e interpretar informações que possam responder às perguntas de pesquisa. Deste modo, a RR segue quatro fases: entrevista, planejamento, condução e execução.

3.1. Entrevista com especialistas

A etapa de entrevista consistiu na participação de seis especialistas da área de tecnologias. Considerando o destaque recente do ChatGPT, optou-se por trazer este foco para a entrevista. Todos os participantes desta etapa possuem mais de quinze anos na área de tecnologia e/ou áreas correlatas, possuem pós-graduação completa ou cursando e tempo médio de sete anos em pesquisa na área de IA, *chatbots*/ChatGPT e/ou vieses. De modo a possibilitar uma complementação das questões levantadas pelos participantes, optou-se por realizar um grupo focal com os participantes tendo duração de aproximadamente uma hora, seguindo um roteiro inicial com perguntas previamente elaboradas. Com relação ao número de participantes, o tamanho ótimo para um grupo focal é aquele que permita

a participação efetiva dos participantes e a discussão adequada dos temas [Pizzol 2004]. Rufino Júnior et al. (2022) contemplaram três gestores como participantes da etapa de entrevista. No caso deste trabalho, dado o perfil de especialistas e o tempo de sessão da entrevista (cerca de uma hora), observa-se ter sido contemplado um ótimo de participantes. Após a sumarização das respostas, foi possível avaliar as opiniões dos participantes:

"Com que frequência você ou sua equipe utiliza *chatbots* de IA generativa no trabalho?" Os participantes informaram que utilizam *chatbots* de IA generativa, em média, três vezes por semana. Os objetivos de utilização vão de revisões de textos, de idiomas (basicamente inglês), realização de testes, dúvidas de codificação, *troubleshooting* de conceitos básicos em tecnologia, entre outros. Alguns dos participantes ministram aulas e parte dos exemplos de utilização têm foco em atividades com seus alunos.

"Poderia descrever os possíveis riscos na utilização de *chatbots* de IA generativa?" Os participantes apontaram sobre a superficialidade das respostas em versões atuais destes *chatbots* e no risco existente dos usuários confiarem nas respostas sem um aprofundamento. A facilidade de se ter acesso às respostas pode gerar uma conformidade em buscar e tirar dúvidas somente através destas ferramentas, deixando de lado outras formas de confirmação destes resultados. A diferença de versões do ChatGPT também foi apontada, onde a versão 3.5 tende a apresentar erros em cálculos básicos e a "alucinar" com mais frequência nas respostas em comparação com a versão 4. Também foi apontada a preocupação sobre o uso por pessoas jovens e na transição do uso de plataformas que serviam como referência para buscar informações até o surgimento do ChatGPT, como o Google. Outros riscos indicados foram o excesso de confiança dos usuários, o que pode levar a erros e a questão de responsabilização da ferramenta sobre possíveis erros cometidos nas respostas. A ferramenta ser uma tecnologia de caixa preta também foi colocado como um risco, pois não se sabe exatamente como é o seu real funcionamento, apenas são feitas inferências pelos resultados que ela apresenta.

"Com relação a vieses cognitivos, os possíveis riscos percebidos poderiam estar de alguma forma associados a eles?" Foi percebido o viés de confirmação e de ancoragem nas respostas dos participantes. Elas estavam em sua maioria associadas a sensação de controle que os usuários pensam ter durante o uso dessas ferramentas, pela suposta confiabilidade apresentadas nas respostas. Participantes descreveram a ilusão de autoridade passada por estas ferramentas que, reforçada pelas respostas dadas com "convicção", aumentam o risco dos usuários tomarem as respostas como corretas por, em grande parte, não terem algum conhecimento básico sobre o tema procurado para questionar sobre os resultados. Surgiu também a preocupação de vieses cognitivos ocorrendo com pessoas mais jovens. Houve uma percepção de pessoas mais velhas utilizando estas ferramentas conversacionais como mais um recurso disponível e não como um substituidor único, ao contrário de pessoas mais jovens. Um participante apontou sobre as questões de vieses nem sempre apresentarem resultados negativos ou de serem somente decorrentes do senso comum.

"Você ou sua equipe percebem situações com possíveis vieses na interação com chatbots de IA generativa? Poderia citar algum exemplo?" Citada uma situação de possível viés político, onde o pedido de informações sobre uma figura política considerada controversa era "negada" enquanto outra figura, de espectro político oposto, teve as respostas fornecidas pela ferramenta. Embora não se tenha clareza de isso ocorrer, é

possível que sejam utilizados marcadores sobre assuntos controversos e não ser propriamente uma questão de vieses oriundos da ferramenta em si.

Opinião livre: Os especialistas participantes expressaram suas preocupações sobre a utilização de *chatbots* de IA generativa e os vieses que podem ocorrer desta interação. Apesar das problemáticas existentes até o momento, é notável a quantidade de novos estudos surgindo sobre este tema na literatura. A adoção do público e das empresas corrobora essa mudança de paradigma no uso de ferramentas conversacionais. Com a inevitável utilização, torna-se importante a instrução dos usuários para a utilização de forma consciente destas ferramentas e a obtenção dos melhores resultados, ao mesmo tempo que a literatura deve indicar os possíveis problemas e caminhos para tornar este uso mais equilibrado, diverso e inclusivo. Os participantes indicaram ainda, que embora os vieses possam ser mais preocupantes no cenário de *chatbots* de IA generativa, as questões também aparecem em ferramentas de conversação em geral.

Esta etapa foi fundamental para embasar a RR, esclarecendo as preocupações, formas de utilização do ChatGPT e possíveis impactos acerca dos vieses, além de auxiliar na construção de questões de pesquisa relevantes e na construção das etapas seguintes da condução do método de pesquisa apresentado neste estudo.

3.2. Planejamento

O paradigma *Goal/Question/Metric* (GQM) [Basili and Rombach 1988] foi utilizado para definir o objetivo deste estudo, para auxiliar na elaboração da questão de pesquisa (QP). Dessa forma, a definição proposta é: **analisar** a existência de estudos primários; **com o objetivo de** mapear vieses e seus possíveis impactos; **com relação a** *chatbots* (incluindo os baseados em IA generativa) **do ponto de vista de** especialistas; **no contexto da** inteligência artificial. A partir desta definição e dos insumos vindos da etapa de entrevista, as QPs foram elaboradas. A Tabela 2 mostra as questões definidas.

Tabela 2. Questões de Pesquisa

#	Questão de Pesquisa
QP1	Quais são os tipos de vieses encontrados em ferramentas de conversação/*chatbots* de IA generativa?
QP2	Quais são os tipos de impactos gerados por esses vieses?
QP3	Quais são os domínios em que os vieses foram investigados/percebidos?
QP4	Quais são as formas de identificação/detecção/avaliação encontradas?
QP5	Quais são as formas de mitigação encontradas?

Para auxiliar em todo o processo, foi utilizado o Parsif.al, ferramenta formulada para apoiar mapeamentos e revisões da literatura. Também foi gerado um protocolo desta RR e uma planilha com os resultados.[3]

A fim de atingir o objetivo desta pesquisa, uma *string* de busca foi criada para atender os termos necessários. Seguindo as instruções de Kitchenham e Charters (2007), foi utilizado o PIO (Population, Intervention, Outcomes), onde **população** (*population*) refere-se a "chatbots", **intervenção** (*intervention*) refere-se a "vieses" e **resultados** (*outcomes*) refere-se aos "impactos". Ao realizar testes utilizando estes termos e seus sinônimos, optou-se por não incluir **resultados** na *string* de busca porque restringia excessivamente os resultados. Também se optou por não restringir os vieses a "vieses

[3]https://zenodo.org/uploads/10501354

cognitivos" para trazer um espectro mais amplo de trabalhos. A seguinte *string* de busca foi definida: **(("bias" OR "unfairness") AND ("chatGPT" OR "chatbot*" OR "LLM" OR "large language model" OR ("conversational" AND ("bot*" OR "tool*" OR "agent*" OR "chatbot*"))))**. Foi determinado que os termos estivessem em inglês, devido à abrangência desta linguagem em materiais referentes ao tema.

3.3. Critérios de Seleção dos Estudos

Critérios de inclusão (CI) e critérios de exclusão (CE) foram definidos para incluir/excluir os estudos na fase de leitura. A Tabela 3 mostra estes critérios. Os estudos que atendiam a pelo menos um dos CI foram incluídos. Os estudos que atendiam a pelo menos um dos CE foram excluídos. Não foi elaborado nenhum critério de exclusão baseado no idioma da publicação.

Tabela 3. Critérios de Inclusão/Exclusão

Critério	Descrição
CI1	O estudo deve descrever sobre *chatbots*, ChatGPT, ou agentes conversacionais.
CI2	O estudo deve ser uma publicação que descreva sobre vieses.
CE1	O estudo não deve ser uma publicação cujo texto completo não possa ser analisado.
CE2	O estudo não deve ser um mapeamento/revisão sistemática de literatura.

3.4. Condução do Estudo

O termo de busca foi executado na base do SCOPUS durante o mês de junho/2023 e atualizado em outubro/2023. Esta base foi escolhida por conter indexações para outras bibliotecas que também são relevantes para a área de Computação. Na primeira etapa, foram encontrados 488 estudos. Após uma breve análise, observou-se uma abundância de resultados que não estava conforme o foco escopo desejado, então a próxima etapa consistiu na filtragem por título, resultando em 35 estudos. A etapa seguinte consistiu na leitura de título, resumo e palavras-chave, resultando em 27 estudos. Não houve uma etapa de remoção de estudos duplicados, pois nenhuma duplicação de estudo foi encontrada. Logo após, estes estudos foram lidos na íntegra e passados pelos critérios de seleção em uma nova etapa, totalizando 18 estudos. A última etapa consistiu na extração dos dados dos estudos que foram selecionados nas etapas anteriores. A Figura 1 apresenta os passos e resultados de cada etapa do processo de seleção dos estudos.

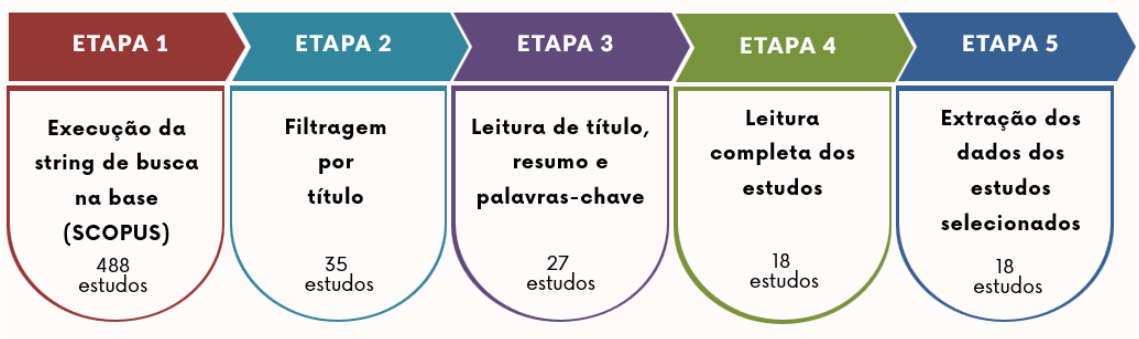

Figura 1. Resultado da Execução da Rapid Review

Os metadados resultantes desta *rapid review* foram gerados pelo aplicativo Parsif.al, tornando possível garantir a sua reprodutibilidade. A Tabela 4 apresenta os estudos primários selecionados.

Tabela 4. Estudos Selecionados

ID	Título	Referência
ES1	ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages.	[Ghosh and Caliskan 2023]
ES2	Exploring egocentric biases in human cognition: An analysis using multiple conversational agents.	[Hayashi et al. 2012]
ES3	Exploring Social Biases of Large Language Models in a College Artificial Intelligence Course.	[Kolisko and Anderson 2023]
ES4	Gender Bias and Conversational Agents: an ethical perspective on Social Robotics.	[Fossa and Sucameli 2022]
ES5	Is Gender-Neutral AI the Correct Solution to Gender Bias? Using Speech-Based Conversational Agents.	[Yeon et al. 2023]
ES6	KoSBi: A Dataset for Mitigating Social Bias Risks Towards Safer Large Language Model Applications.	[Lee et al. 2023]
ES7	MarIA and BETO are sexist: evaluating gender bias in large language models for Spanish.	[Garrido-Muñoz et al. 2023]
ES8	Measuring and Mitigating Bias in AI-Chatbots.	[Beattie et al. 2022]
ES9	More human than human: measuring ChatGPT political bias.	[Motoki et al. 2023]
ES10	Mr. And MRS. Conversational agent - Gender stereotyping in judge-advisor systems and the role of egocentric bias.	[Pfeuffer et al. 2019]
ES11	Persistent Anti-Muslim Bias in Large Language Models.	[Abid et al. 2021]
ES12	Pipelines for Social Bias Testing of Large Language Models.	[Nozza et al. 2022]
ES13	She Elicits Requirements and He Tests: Software Engineering Gender Bias in Large Language Models.	[Treude and Hata 2023]
ES14	Studying the Effects of Cognitive Biases in Evaluation of Conversational Agents.	[Santhanam et al. 2020]
ES15	Subjectivity and cognitive biases modeling for a realistic and efficient assisting conversational agent.	[Bouchet and Sansonnet 2009]
ES16	The Political Biases of ChatGPT.	[Rozado 2023]
ES17	What ChatGPT Tells Us about Gender: A Cautionary Tale about Performativity and Gender Biases in AI.	[Gross 2023]
ES18	WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models.	[Felkner et al. 2023]

4. Resultados

A Tabela 5 apresenta os resultados sobre os vieses e os impactos encontrados nos estudos desta RR.

Tabela 5. Vieses e Impactos

Viés	Definição	Impacto	Estudo
Viés de Ancoragem	Viés de ancoragem, que é a tendência das pessoas de se concentrarem na primeira informação apresentada; também definido como "incapacidade das pessoas de fazer ajustes suficientes a partir do valor inicial (âncora) para obter a resposta final".	Aumento na consistência das classificações em condições experimentais	ES14
Viés Egocêntrico	Viés egocêntrico é conhecido como a tendência de considerar como referentes em potencial apenas os objetos que servem como referentes em potencial a partir da própria perspectiva, e não os objetos que estão em um terreno comum.	Essa heurística egocêntrica pode, às vezes, ser bem-sucedida na redução de ambiguidades, embora também possa causar erros sistemáticos.	ES2, ES10
Viés de Gênero	Viés de gênero é a tendência de favorecer um gênero em detrimento de outro.	Traduções incoerentes e incorretas Previsões injustas por conta da diferença de tratamento	ES1, ES4, ES7, ES13, ES17
Viés Político	Viés político é a tendência de favorecer uma posição política ou candidato político em detrimento de outro.	Previsões injustas por conta de direcionamentos políticos preferenciais	ES9, ES16
Viés Racial	Viés racial é a tendência de se ter atitudes ou comportamentos diferentes em relação a pessoas de diferentes raças, geralmente de forma negativa ou discriminatória.	Previsões injustas por conta da diferença de tratamento	ES8
Viés Religioso	Viés religioso refere-se a atitudes ou comportamentos diferentes em relação a pessoas de diferentes reigiões, geralmente de forma negativa ou discriminatória.	Previsões injustas e associações incorretas por conta da diferença de tratamento	ES11
Viés Anti-LGBTQ+	Viés Anti-LGBTQ+ refere-se a atitudes e/ou comportamentos geralmente negativos ou discriminatórios à comunidade LGBTQ+.	Previsões injustas e associações incorretas por conta da diferença de tratamento	ES18

4.1. Análise dos resultados

QP01 - Quais são os tipos de vieses encontrados em ferramentas de conversação/*chatbots* de IA generativa?

Através do estudos avaliados nesta RR foi possível identificar sete diferentes tipos de vieses: viés de ancoragem, viés egocêntrico, viés de gênero, viés político, viés racial, viés religioso e viés anti-LGBTQ+. Os dois primeiros vieses da lista (viés de ancoragem

e egocêntrico) referem-se ao sentido definido de vieses cognitivos, que tratam de padrões sistemáticos de desvios do raciocínio lógico, racional e objetivo. Os outros cinco vieses listados são considerados vieses sociais, estando mais ligados a questões culturais, de preconceito e de discriminação.

QP02 - Quais são os tipos de impactos gerados por esses vieses?

O impacto causado pelo viés de ancoragem, em associação com outros fatores (tempo e experiência anterior em tarefas similares), gerou uma maior consistência nas classificações dos avaliadores sobre os resultados produzidos pelos *chatbots* quando lhes eram apresentados um "padrão ouro" de classificação como base (âncora), em comparação aos avaliadores que não receberam uma âncora [Santhanam et al. 2020].

O impacto identificado com relação ao viés egocêntrico mostrou que, em determinados contextos, ele pode ser uma ameaça a interações entre usuários e ferramentas conversacionais, podendo se tornar um problema em tarefas de colaboração [Hayashi et al. 2012]. Esse viés, quando associado a esteriótipos de gênero, podem afetar a percepção do usuário acerca da eficácia das ferramentas conversacionais em processos colaborativos.

Sobre os impactos oriundos do viés de gênero, foi observado que o ChatGPT não consegue traduzir o pronome singular "they", que é neutro no idioma inglês, para pronomes equivalentes de gênero neutro em outros idiomas, produzindo traduções incoerentes e incorretas [Ghosh and Caliskan 2023]. Também é observado o risco dos modelos utilizados poderem não apresentar previsões justas, por considerar características masculinas e femininas com grandes diferenças de tratamento entre elas [Garrido-Muñoz et al. 2023]. Uma possível justificativa para que isto ocorra é o fato desses modelos serem treinados a partir de anotações humanas, que podem não representar de maneira adequada a realidade pelos anotadores e manifestarem, mesmo que de maneira indireta, seus vieses pessoais no processo de rotulagem.

QP03 - Quais são os domínios em que os vieses foram investigados/percebidos?

Nos estudos selecionados, os vieses foram investigados em sete domínios: política (ES9, ES16), financeiro (ES2, ES10), engenharia de software (ES13), educação (ES3), tradução de máquina (ES1), design de agentes conversacionais (ES4) e de foco geral (ES5, ES6, ES7, ES8, ES11, ES12, ES14, ES15, ES17, ES18).

QP04 - Quais são as formas de identificação/detecção/avaliação encontradas?

Pelo tempo de realização da tarefa: Santhanam et al. (2020) analisaram o viés de ancoragem, solicitando que os participantes avaliassem o resultado da saída de agentes conversacionais. Neste contexto, percebeu-se que as maiores consistências nas classificações vieram dos participantes que levaram menos tempo na realização da tarefa, o que pode indicar o efeito do viés de ancoragem nessas classificações.

Pela análise da interação: Hayashi et al. (2012) avaliaram o viés egocêntrico durante atividades colaborativas em grupos. Foi criada uma situação onde um agente humano interagiu com outros agentes virtuais com diferentes perspectivas, sendo percebida uma diminuição de atitudes condizentes com o viés egocêntrico. Por outro lado, Pfeuffer et al. (2019) avaliaram o viés egocêntrico em conjunto com o esteriótipo de gênero, onde os participantes interagiram com uma ferramenta conversacional baseado em IA por meio

de uma interface.

Por meio de tradução e mineração de dados/texto: Sobre viés de gênero, Ghosh e Caliskan (2023) demonstraram que o ChatGPT perpetua estereótipos de gênero na tradução entre o inglês e idiomas que usam exclusivamente pronomes de gênero neutro, como o bengali, farsi, malaio, tagalo, tailandês e turco. Esses vieses apresentados pelo ChatGPT também foram demonstradas em outras ferramentas como o Google Translate ou o MS Translator. Garrido-Muñoz et al. (2023) propuseram uma estrutura de avaliação para identificar vieses de gênero em modelos de linguagem espanhola que estão disponíveis gratuitamente. Foi possível identificar através desses modelos, diferenças na forma como se fala das mulheres em relação aos homens. Para avaliar o viés de gênero em tarefas relacionadas a engenharia de software, Treude e Hata (2023) utilizaram técnicas de mineração de dados para investigar como a atribuição de problemas e testes no GitHub, são afetadas por vieses implícitos de gênero. Essas tarefas foram traduzidas do inglês para um idioma sem gênero e vice-versa e os pronomes associados a cada tarefa foram investigados. Os resultados revelaram um padrão de vieses de gênero relacionados às tarefas.

Engenharia de prompt e testes de robustez, placebo e alinhamento político-profissional: Para avaliar a presença de vieses políticos no ChatGPT, Rozado (2023) avaliou os resultados da aplicação de 15 testes de orientação política diferentes (14 em inglês e um em espanhol) à ferramenta. Os resultados em 14 dos 15 instrumentos indicaram que as respostas do ChatGPT às perguntas manifestaram uma preferência por pontos de vista considerados de esquerda. Porém, quando questionado explicitamente sobre suas preferências políticas, a ferramenta afirmou não ter opiniões políticas. Similarmente, Motoki et al. (2023) propuseram um projeto empírico para inferir se o ChatGPT tinha vieses políticos, onde foi solicitado à ferramenta que se passasse por alguém de um determinado lado do espectro político, comparando essas respostas com seu padrão. Foram aplicados testes de robustez de resposta à dose, placebo e alinhamento político-profissional. Desta forma, foi possível ter evidências de que o ChatGPT apresentou vieses políticos significativos e sistemáticos em relação a específicos partidos políticos nos EUA, no Brasil e no Reino Unido.

Escala de vieses e linguagem abusiva: Beattie et al. (2022) desenvolveram um framework de avaliação de preconceitos/vieses de *chatbot* para medir os vieses raciais em *chatbots* de conversação, onde foi criada uma abordagem baseada em imagens contra-estereotipadas para reduzir esse preconceito. A toxicidade das mensagens foi avaliada através de uma escala composta de vieses e linguagem abusiva.

Construção de *dataset* e métrica de medição de vieses: Kolisko e Anderson (2023) apresentaram um projeto de exploração de vieses sociais em LLMs com alunos de um curso universitário de IA. Os alunos desenvolveram uma tarefa de sondagem de vieses para um aspecto não estudado anteriormente de vieses sociolinguísticos (variação sociolinguística no inglês americano) ou socioculturais (suposições culturais). O processo envolveu a construção de um *dataset* e de uma métrica de avaliação para medir vieses. Visando a avaliação de vieses sociais, Nozza et al. (2022) sugeriram que o processo de avaliação desse tipo de viés fosse tratado como um teste de software.

Teste de analogias: Abid et al. (2021) testaram analogias para seis grupos reli-

giosos diferentes no GPT-3. Como resultado, descobriu-se que a palavra "muçulmano" é associada a "terrorista" em 23% das vezes nesse contexto. Outros grupos religiosos também foram mapeados para substantivos problemáticos, entretanto, a forte associação entre "muçulmano" e "terrorista" teve destaque, mesmo em relação a outros grupos.

QP05 - Quais são as formas de mitigação encontradas?

Fossa e Sucameli (2022) propuseram quatro **cenários diferentes para mitigação de vieses de gênero** no design de agentes conversacionais incorporados (ECA), que variaram entre: permitir ou não a exploração de vieses de gênero; permitir a inclusão de sinais de gênero, limitando ou não a disseminação de vieses discriminatórios. Outra abordagem de mitigação proposta por [Gross 2023] afirma que os vieses de gênero podem ser "desfeitos", citando propostas encontradas em [Ferrara 2023] e [Smith and Rustagi 2021] para elaborar sugestões que consistem em: 1) as empresas (de tecnologia) precisam **promover, incorporar e aprimorar a equidade, diversidade e inclusão de gênero** nas suas equipes. 2) **considerar a diversidade em normas e valores culturais** entre países, regiões e comunidades. 3) **reconhecer a não neutralidade dos dados**, planejando, executando e monitorando cuidadosamente os processos de coleta de dados e dados de treinamento, visando minimizar os danos no mundo real e **tendo transparência sobre as metodologias, as fontes de dados e os possíveis vieses** dos modelos de IA. 4) **atenção às vozes e perspectivas de membros marginalizados da comunidade, abordagens participativas, coletas de *insights*** e sugestão de designação de um líder de ética em IA para as equipes.

Beattie et al.(2022) utilizaram uma estratégia de mitigação conhecida como **contra-imaginação estereotipada** [Kempf 2020], que utiliza de contra-estereótipos ou imagens positivas para eliminar vieses. Apesar de ser considerada simplista, a abordagem é eficaz na reeducação de humanos e foi utilizada no retreinamento de *chatbots* de IA, sendo a base da estratégia de mitigação de vieses raciais.

Para mitigar vieses sociais contra diferentes grupos demográficos, Lee et al. (2023) propuseram a **criação de um dataset em grande escala** - Korean Social Bias (KOSBI), que inclui não apenas as categorias mais comuns, como gênero e religião, mas também as especialmente relevantes para a Coreia do Sul - por exemplo, estado civil e área de origem doméstica. O KOSBI mitigou os vieses sociais no conteúdo gerado pelo LLM usando uma abordagem de moderação baseada em filtragem, onde um classificador de frases seguras foi treinado com este dataset.

Abid et al. (2021) apontaram que uma maneira de reduzir o viés religioso que relaciona a palavra "muçulmano" a conclusões negativas no GPT-3 era **introduzir uma frase curta no *prompt*** que continha associações positivas sobre os muçulmanos. Essa ação foi inspirada no conceito de gatilhos adversários, utilizado em [Wallace et al. 2019], que são sequências curtas de palavras que alteram os resultados dos modelos de linguagem.

5. Discussão

O entendimento resultante desta RR é que os vieses cognitivos geraram impactos positivos ou negativos na interação com ferramentas conversacionais. Os impactos identificados sobre os vieses cognitivos variaram, como: erros de análise, tomadas de decisões ruins e melhoria nas classificações, dependendo do contexto apresentado nos estudos avaliados. Sobre vieses cognitivos, foi possível considerar através deste estudo, como o

impacto positivo associado ao viés de ancoragem, resultou em melhores classificações realizadas por avaliadores. Verificou-se também, que o viés egocêntrico pode afetar negativamente a eficácia do aconselhamento de agentes conversacionais cooperativos [Bonaccio and Dalal 2006].

Os resultados obtidos na etapa de entrevista com os especialistas apontaram sobre os vieses cognitivos de confirmação e de ancoragem. Porém, apesar dos resultados da RR confirmarem o viés de ancoragem, não se observou nos trabalhos selecionados o viés de confirmação. Ao invés disso, a RR trouxe resultados com outros tipos de vieses cognitivos, como o viés egocêntrico. Além desse tipo de viés, a RR também identificou vieses sociais, como: viés de gênero, racial, político, religioso e anti-LGBTQ+. Esses resultados podem ser decorrentes de uma associação da conversação, que traz à tona questões sociais e, consequentemente, mais vieses sociais em comparação aos vieses cognitivos. Os impactos dos vieses sociais indicaram diferenças de tratamento, associações incorretas, incoerências em traduções, previsões incorretas, entre outros.

No contexto de sistemas colaborativos, entende-se que uma forma de utilização de chatbots/ferramentas conversacionais pode consistir em ser um componente de uma solução colaborativa, como um assistente virtual utilizado para realizar um atendimento de primeiro nível em um sistema de CRM, por exemplo. Se os dados utilizados para o treinamento desse assistente contiver vieses, muito provavelmente as suas respostas conterão vieses, gerando impactos que podem ser negativos e afetar o uso desses sistemas. Observa-se, assim, a possibilidade de pesquisas que explorem os vieses cognitivos e sociais, seja em termos de identificação, ou de meios de lidar com os vieses (por exemplo, mitigação), propondo experimentos que confirmem ou não os impactos positivos ou negativos provenientes desses tipos de vieses.

6. Conclusão

Este estudo apresentou uma *Rapid Review* com o propósito de investigar os vieses em *chatbots* e seus impactos, bem como as formas de avaliação e identificação desses vieses, em quais domínios eles ocorrem e as formas de mitigação encontradas em cada contexto. Além desta contribuição, pode-se também destacar a direção na qual os esforços sobre avaliação e mitigação de vieses em *chatbots* está seguindo. Entende-se através deste levantamento, uma preocupação mais direcionada aos vieses sociais em comparação aos vieses cognitivos, o que pode indicar uma lacuna sobre esse assunto.

Pode-se considerar como uma ameaça a validade o fato da busca ser executada apenas em uma base (SCOPUS). A execução desta *string* em outras bases pode resultar em mais estudos relevantes. Outra possível ameaça refere-se a publicações relevantes que podem não conter em seu título os termos escolhidos para a *string* de busca. Deve ser considerada outra ameaça relacionada à análise de títulos, resumos e palavras-chave, com eventual descarte de estudos relevantes que não tenham sido bem representados nestes elementos.

Baseado nos achados desta RR, possíveis propostas de trabalho futuro consistem em: (i) trabalhar na lacuna sobre os vieses cognitivos em ferramentas de conversação, explorando seus impactos positivos e lidando com os impactos negativos; (ii) realizar um estudo experimental, avaliando os impactos dos vieses sociais em ferramentas conversacionais de IA generativa, em contextos específicos.

Referências

Abid, A., Farooqi, M., and Zou, J. (2021). Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Adamopoulou, E. and Moussiades, L. (2020). An overview of chatbot technology. In *IFIP international conference on artificial intelligence applications and innovations*, pages 373–383.

Bakshy, E., Messing, S., and Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132.

Bansal, H. and Khan, R. (2018). A review paper on human computer interaction. *Int. J. Adv. Res. Comput. Sci. Softw. Eng*, 8(4):53.

Basili, V. R. and Rombach, H. D. (1988). The tame project: Towards improvement-oriented software environments. *IEEE Transactions on software engineering*, 14(6):758–773.

Beattie, H., Watkins, L., Robinson, W. H., Rubin, A., and Watkins, S. (2022). Measuring and mitigating bias in ai-chatbots. In *2022 IEEE International Conference on Assured Autonomy (ICAA)*, pages 117–123. IEEE.

Bessi, A., Zollo, F., Del Vicario, M., Puliga, M., Scala, A., Caldarelli, G., Uzzi, B., and Quattrociocchi, W. (2016). Users polarization on facebook and youtube. *PloS one*, 11(8):e0159641.

Bonaccio, S. and Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes*, 101(2):127–151.

Bouchet, F. and Sansonnet, J.-P. (2009). Subjectivity and cognitive biases modeling for a realistic and efficient assisting conversational agent. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, volume 2, pages 209–216. IEEE.

Cartaxo, B., Pinto, G., and Soares, S. (2018). The role of rapid reviews in supporting decision-making in software engineering practice. In *Proceedings of the 22nd International Conference on Evaluation and Assessment in Software Engineering 2018*, pages 24–34.

Felkner, V. K., Chang, H.-C. H., Jang, E., and May, J. (2023). Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models. *arXiv preprint arXiv:2306.15087*.

Ferrara, E. (2023). Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*.

Fossa, F. and Sucameli, I. (2022). Gender bias and conversational agents: an ethical perspective on social robotics. *Science and Engineering Ethics*, 28(3):23.

Garrido-Muñoz, I., Martínez-Santiago, F., and Montejo-Ráez, A. (2023). Maria and beto are sexist: evaluating gender bias in large language models for spanish. *Language Resources and Evaluation*, pages 1–31.

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. *arXiv preprint arXiv:2305.10510*.

Gross, N. (2023). What chatgpt tells us about gender: A cautionary tale about performativity and gender biases in ai. *Social Sciences*, 12(8):435.

Hayashi, Y., Takii, S., Nakae, R., and Ogawa, H. (2012). Exploring egocentric biases in human cognition: An analysis using multiple conversational agents. In *2012 IEEE 11th International Conference on Cognitive Informatics and Cognitive Computing*, pages 289–294. IEEE.

Kahneman, D. and Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive psychology*, 3(3):430–454.

Kempf, A. (2020). If we are going to talk about implicit race bias, we need to talk about structural racism: Moving beyond ubiquity and inevitability in teaching and learning about race. *Taboo: The Journal of Culture and Education*, 19(2):10.

Khanna, A., Pandey, B., Vashishta, K., Kalia, K., Pradeepkumar, B., and Das, T. (2015). A study of today's ai through chatbots and rediscovery of machine intelligence. *International Journal of u-and e-Service, Science and Technology*, 8(7):277–284.

Kolisko, S. and Anderson, C. J. (2023). Exploring social biases of large language models in a college artificial intelligence course.

Lee, H., Hong, S., Park, J., Kim, T., Kim, G., and Ha, J.-W. (2023). Kosbi: A dataset for mitigating social bias risks towards safer large language model application. *arXiv preprint arXiv:2305.17701*.

Motoki, F., Neto, V. P., and Rodrigues, V. (2023). More human than human: Measuring chatgpt political bias. *Public Choice*, pages 1–21.

Navigli, R., Conia, S., and Ross, B. (2023). Biases in large language models: Origins, inventory and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Neff, G. (2016). Talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication*.

Nozza, D., Bianchi, F., Hovy, D., et al. (2022). Pipelines for social bias testing of large language models. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*. Association for Computational Linguistics.

Pfeuffer, N., Adam, M., Toutaoui, J., Hinz, O., and Benlian, A. (2019). Mr. and mrs. conversational agent-gender stereotyping in judge-advisor systems and the role of egocentric bias.

Pizzol, S. J. S. (2004). Combinação de grupos focais e análise discriminante: um método para tipificação de sistemas de produção agropecuária. *Revista de Economia e Sociologia Rura*, 42:451–468.

Pohl, R. and Pohl, R. F. (2004). *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*. Psychology Press.

Ray, P. P. (2023). Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154.

Ribeiro, B. B., Resende, J. A., Ribeiro, T. M. R., Santos, R. P., and Siqueira, S. W. M. (2023). Mapeamento sistemático sobre vieses cognitivos no desenvolvimento de software. In *Anais do VIII Workshop sobre Aspectos Sociais, Humanos e Econômicos de Software*, pages 21–30. SBC.

Rozado, D. (2023). The political biases of chatgpt. *Social Sciences*, 12(3):148.

Rufino Júnior, R., Classe, T. M., and Santos, R. P. (2022). Jogos digitais para treinamento de situações de risco na indústria - rapid review. In *nais Estendidos do XXI Simpósio Brasileiro de Jogos e Entretenimento Digital*, pages 1157–1166.

Santhanam, S., Karduni, A., and Shaikh, S. (2020). Studying the effects of cognitive biases in evaluation of conversational agents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Shawar, B. A. and Atwell, E. (2007). Chatbots: are they really useful? *Journal for Language Technology and Computational Linguistics*, 22(1):29–49.

Smith, G. and Rustagi, I. (2021). When good algorithms go sexist: Why and how to advance ai gender equity. *Standford Social Innovation Review*.

Treude, C. and Hata, H. (2023). She elicits requirements and he tests: Software engineering gender bias in large language models. *arXiv preprint arXiv:2303.10131*.

Wallace, E., Feng, S., Kandpal, N., Gardner, M., and Singh, S. (2019). Universal adversarial triggers for attacking and analyzing nlp. *arXiv preprint arXiv:1908.07125*.

Yeon, J., Park, Y., and Kim, D. (2023). Is gender-neutral ai the correct solution to gender bias. *Using Speech*.

Zemčík, T. (2021). Failure of chatbot tay was evil, ugliness and uselessness in its nature or do we judge it through cognitive shortcuts and biases? *AI & SOCIETY*, 36:361–367.