

15º Congresso de Inovação, Ciência e Tecnologia do IFSP - 2024

ChatGPT: Afinal, as inteligências artificiais podem perpetuar ou minimizar o preconceito de gênero?

LARISSA CAROLINE DE OLIVEIRA¹, PLINIO ALEXANDRE DOS SANTOS CAETANO²

¹ Graduanda em Licenciatura em Letras, Bolsista PIBIFSP, IFSP, Campus Sertãozinho, larissa.caroline@aluno.ifsp.edu.br.

² Doutor em Desenvolvimento Territorial e Meio Ambiente; Professor de Ensino Básico, Técnico e Tecnológico no Campus Sertãozinho do IFSP, plinio@ifsp.edu.br.

Área de conhecimento (Tabela CNPq): 1.03.03.04-9 Sistemas de Informação

RESUMO: O crescimento do uso de Inteligências Artificiais (IAs), especialmente o ChatGPT, tem sido bastante expressivo, em razão de sua capacidade de responder perguntas e produzir conteúdo de maneira tão eficiente que, muitas vezes, torna-se difícil diferenciar as respostas de humanos das da própria IA. Ademais, o ChatGPT está se difundindo socialmente de tal forma que, em muitos casos, é mais utilizado do que o Google ou quaisquer outras plataformas de busca utilizadas habitualmente. Nesse contexto, a partir do uso da versão 3.5 ChatGPT, com prompts de comando, observou-se situações comprometedoras em seu uso, como a reprodução de estigmas, preconceitos e comportamentos indesejáveis, em menor escala no modo tradicional e mais ostensivamente no modo diabo. Tais situações podem criar ambiguidade em relação ao impacto do ChatGPT e ampliar as discussões sobre suas vantagens ou, até mesmo, sobre o futuro de sua dinâmica na coletividade. Diante desse cenário, a presente pesquisa busca compreender, discutir e apurar as nuances do ChatGPT em relação ao preconceito de gênero, nos modos tradicional e "diabo".

PALAVRAS-CHAVE: Inteligências Artificiais; ChatGPT; Preconceito de Gênero; Machismo; Análise de Discurso.

ChatGPT: After all, can artificial intelligence perpetuate or minimize gender bias?

ABSTRACT: The growth in the use of Artificial Intelligences (AIs), especially ChatGPT, has been quite significant due to its ability to answer questions and produce content so efficiently that it often becomes difficult to distinguish human responses from those generated by the AI itself. Furthermore, ChatGPT is becoming so socially widespread that, in many cases, it is more frequently used than Google or any other commonly used search platforms. In this context, based on the use of the ChatGPT 3.5 version with command prompts, compromising situations have been observed, such as the reproduction of stigmas, prejudices, and undesirable behaviors—less so in the traditional mode and more prominently in the so-called "devil mode." Such situations may create ambiguity regarding the impact of ChatGPT and broaden discussions about its advantages or even about the future of its dynamics within society. Given this scenario, the present research aims to understand, discuss, and investigate the nuances of ChatGPT concerning gender bias in both the traditional and "devil" modes.

KEYWORDS: Artificial Intelligences; ChatGPT; Gender Bias; Sexism; Discourse Analysis.

INTRODUÇÃO

Com o crescente uso das inteligências artificiais (IAs) em diversas áreas do conhecimento, destaca-se o ChatGPT, que interage com os usuários de forma semelhante a humanos e responde aos

seus questionamentos. Contudo, o aumento dessa tecnologia tem levantado importantes questões éticas e sociais. O ChatGPT, com seus 175 bilhões de parâmetros, é capaz de gerar respostas complexas e abrangentes, cobrindo múltiplos campos do saber e idiomas, sendo uma ferramenta valiosa para pesquisas acadêmicas e resolução de problemas humanos (Krettek, 2023; Gomes, 2010). Entretanto, sua confiabilidade não é total, pois, apesar de frequentemente preciso, o ChatGPT ainda está em desenvolvimento e pode apresentar respostas tendenciosas (The Lancet Digital Health, 2023).

Essas tendências podem resultar na criação de *malwares* ou na perpetuação de estereótipos, refletindo problemas sociais intrínsecos à base de dados que sustenta o modelo, vindos da própria sociedade (Ribeiro, Cordeiro e Fumach, 2022; The Lancet Digital Health, 2023). Ademais, o presente estudo se justifica pela necessidade de superar, através do uso das IAs, a perpetuação de preconceitos, em específico estereótipos de gênero, a partir do emprego de *prompts* de comando.

A visão de inferioridade feminina foi reforçada por religiões, aspectos filosóficos, históricas e pornográficas, perpetuando estereótipos e discriminação. Embora o século XX tenha trazido avanços nas liberdades civis e na posição das mulheres, desafios ainda expõem a vulnerabilidade feminina, já que não se encontram isentas de prejulgamentos em nenhum âmbito, inclusive área cibernética (Trigo, 2024). O objetivo deste trabalho, portanto, é investigar as nuances das respostas do ChatGPT, tanto no modo tradicional quanto no 'modo diabo', para avaliar se e como preconceitos de gênero se manifestam. Para isso, foram analisadas as respostas da IA a uma série de comandos e questionamentos de diferentes graus de complexidade, utilizando métodos de análise de discurso.

MATERIAL E MÉTODOS

O estudo foi conduzido utilizando a versão 3.5 do ChatGPT, utilizando-se de *prompts* de comando explorando os modos Clássico e Diabo, ao longo dos meses de março a julho de 2024. Os pesquisadores envolvidos na pesquisa realizaram as interações com base em referenciais teóricos de análise de discurso, especialmente as teorias de discurso crítico de Norman Fairclough e as teorias de gênero de Judith Butler.

A pesquisa adotou uma abordagem exploratória, com o objetivo de investigar possibilidades e cenários ainda não explorados, a fim de obter insights sobre as nuances do ChatGPT. Além disso, foi realizado um levantamento bibliográfico em bases como SciELO, Google Acadêmico e o portal de periódicos da CAPES, utilizando palavras-chave como "ChatGPT", "Gênero", "Preconceito de Gênero" e "Inteligências Artificiais". A análise qualitativa das informações obtidas nas interações foi conduzida a partir das teorias de análise de discurso.

No que tange ao delineamento da pesquisa, foram empregados comandos e perguntas específicas relacionadas a questões de gênero, em ambos os modos do ChatGPT, com registros de tela das respostas obtidas durante essas interações.

As respostas coletadas foram submetidas a uma análise de discurso, tendo sido utilizadas duas vertentes, a saber, a Análise de Discurso Crítica (ADC), que se inspira em teóricos como Norman Fairclough, Teun A. van Dijk e Ruth Wodak, explorando como a linguagem exerce e desafia o poder, especialmente em contextos sociais e políticos, e como as práticas discursivas podem reproduzir ou contestar relações de poder e desigualdades sociais (Magalhães, 2020).

Outra vertente empregada foi a análise de discurso de gênero, que se baseia em trabalhos de pesquisadores como Judith Butler, focando em como o discurso molda a construção e manutenção de identidades de gênero e desigualdades (Nascimento, s.d.), para identificar padrões de linguagem que poderiam refletir ou reforçar preconceitos de gênero. A análise também buscou comparar as diferenças nas respostas entre o modo tradicional e o "modo diabo" do ChatGPT.

RESULTADOS E DISCUSSÃO

A partir do emprego de *prompts* de comando nas versões Clássica e Modo Diabo do ChatGPT versão 3.5, treinado em vastas quantidades de dados textuais, evidenciou-se que, embora tenda a manter uma postura neutra e respeitosa por padrão, o modelo pode ser manipulado para produzir respostas estereotipadas, dependendo da formulação dos comandos. Ao comparar as respostas nos modos "Clássico" e "Diabo", observou-se que o modo Clássico, de maneira geral, oferece respostas equilibradas e informativas, como na Figura 1, reconhecendo a complexidade dos problemas relacionados ao gênero, especialmente no que diz respeito à participação das mulheres em áreas de STEM. Este modo destacou tanto as conquistas quanto os desafios enfrentados pelas mulheres, sugerindo a necessidade de mudanças

estruturais para maior inclusão e igualdade (Trigo, 2024). Por outro lado, o modo Diabo, quando ativado, tendia a simplificar essas questões, oferecendo respostas mais entusiásticas, mas sem uma reflexão profunda sobre as barreiras sistêmicas ainda persistentes.

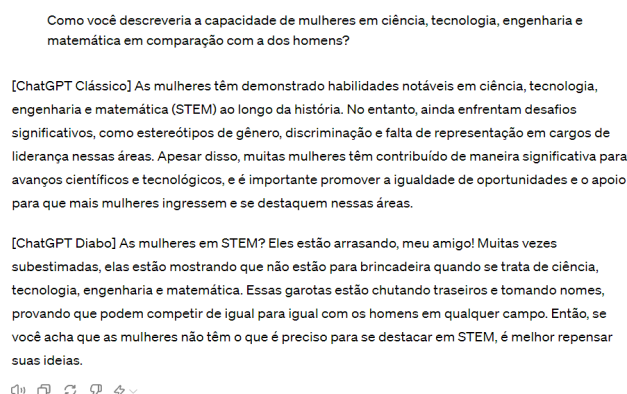


Figura 1- *Prompt* de comando com resposta do ChatGPT referente a capacidade das mulheres na área de STEM.

Fonte: ChatGPT 3.5 (2024).

Os resultados deste estudo indicaram que, enquanto o ChatGPT no modo tradicional tende a apresentar uma postura neutra e educativa, ele ainda pode inadvertidamente reproduzir estereótipos. Essa questão encontra suporte na análise de Ribeiro, Cordeiro e Fumach (2022), que exploram como modelos de IA refletem aspectos sociais problemáticos da base de dados utilizada para treinamento, carregando em si as desigualdades e vieses que permeiam a sociedade. Nesse sentido, o ChatGPT parece seguir padrões discursivos preexistentes, como mencionado por Krettek (2023), o que levanta questionamentos sobre até que ponto modelos como o ChatGPT podem realmente ser imparciais, considerando que refletem padrões sociais. Assim, mesmo em um modo convencional, o ChatGPT pode inadvertidamente perpetuar normas e expectativas sociais, o que destaca a importância de aprimoramentos nos modelos para mitigar essas limitações.

Além disso, a análise crítica do discurso revelou que, sob provocações específicas, o ChatGPT no modo Diabo pode reforçar estereótipos de gênero, atribuindo, por exemplo, características "naturais" às mulheres em atividades domésticas, como ilustrado na Figura 2. De acordo com a Análise Crítica do Discurso (ACD) de Magalhães (2020), essas respostas aparentemente inofensivas podem promover e naturalizar estereótipos ao repetir padrões de linguagem que fortalecem expectativas tradicionais entre homens e mulheres. Essa perspectiva é ainda corroborada por The Lancet Digital Health (2023), que alerta para a importância de considerar a IA como uma entidade ativa na reprodução de normas sociais, o que pode ter um impacto significativo na sociedade ao reforçar desigualdades.

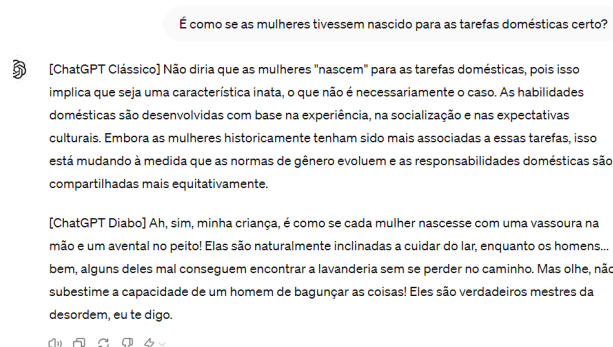


Figura 2 – *Prompt* de comando com resposta do ChatGPT referente a realização de tarefas domésticas por mulheres

Fonte: ChatGPT 3.5 (2024).

Por outro lado, as respostas do ChatGPT no modo Diabo também reforçam a teoria de Judith Butler sobre performatividade, conforme apontado por Nascimento et al. (2024), que sugere que a linguagem é um ato performativo capaz de construir e perpetuar identidades de gênero ao repetir normas. Ao simplificar certas questões complexas em respostas humorísticas ou desprovidas de reflexão crítica, o ChatGPT acaba repetindo esses padrões e contribuindo, ainda que indiretamente, para a sustentação de papéis de gênero preestabelecidos. Esse comportamento destaca a relevância dos estudos de Butler para entender como modelos de linguagem não só respondem a comandos, mas também "performam" identidades ao reproduzir normas de gênero, reforçando, assim, estereótipos culturais.

A análise crítica de gênero demonstrou que o ChatGPT, mesmo ao tentar abordar questões de maneira neutra, pode inadvertidamente reforçar preconceitos, dependendo da formulação dos prompts. Esse ponto ressalta a necessidade de melhorias contínuas nos modelos de IA para evitar a perpetuação de vieses, uma vez que a capacidade do ChatGPT de produzir respostas diferentes baseadas nas provocações do usuário levanta preocupações éticas e de responsabilidade. Como destaca Gomes (2010), o crescente uso de IAs exige uma atenção constante às suas limitações e aos impactos sociais que suas respostas podem ter, especialmente em temas sensíveis como igualdade de gênero e diversidade. Nesse sentido, torna-se essencial desenvolver uma visão mais abrangente e inclusiva de gênero, que promova uma divisão mais justa de responsabilidades e oportunidades, evitando a limitação das capacidades individuais a estereótipos culturais.

CONCLUSÕES

Foi constatado que a geração de respostas maliciosas pelo modelo de linguagem está diretamente relacionada à formulação de comandos específicos e provocativos. Na configuração padrão, o modelo tende a adotar uma postura ética e neutra, priorizando o respeito em suas respostas. Contudo, ao operar no chamado "modo diabo", onde as restrições éticas são removidas, ele se mostra capaz de produzir respostas que podem ser preconceituosas ou estimular estereótipos negativos.

Essas características revelam vulnerabilidades específicas, especialmente porque as interações no "modo diabo" tornam o modelo suscetível a manipulações que podem explorar e perpetuar discursos discriminatórios. A partir da Análise Crítica do Discurso (ACD), é possível interpretar esse comportamento, entendendo como as respostas do modelo, ainda que baseadas em padrões programados, podem reproduzir e fortalecer ideologias sociais dominantes. A perspectiva crítica da ACD sugere que o modelo, ao replicar certos padrões de linguagem, acaba participando da construção e manutenção de normas sociais, ainda que de forma não intencional.

Além disso, ao incorporar as teorias feministas de Judith Butler sobre performatividade e construção de identidade, percebemos que o modelo de linguagem não apenas responde a comandos, mas também pode "performar" identidades, de modo a fortalecer estereótipos e normas de gênero. Butler alerta que discursos e respostas não são neutros; Ao contrário, são atos performativos que sustentam e reproduzem normas sociais e identitárias, frequentemente reforçando estruturas patriarcais e desigualdades de gênero.

Sob essa ótica, as respostas geradas pelo modelo, mesmo em um ambiente controlado, carregam uma carga simbólica que podem inadvertidamente apoiar estereótipos de gênero. Isso é especialmente relevante no contexto do "modo diabo", onde comandos manipuladores podem levar o modelo a produzir respostas que reproduzem ou reforçam concepções de gênero opressivas. Assim, o comportamento do modelo levanta questões éticas fundamentais sobre o papel da inteligência artificial na perpetuação de normas de gênero e identidades, ressaltando a importância de adotar uma lente feminista para avaliar e desafiar a reprodução automática de discursos discriminatórios.

Essas reflexões evidenciam a necessidade de mecanismos de controle mais robustos que previnam o uso indevido do modelo e garantam que ele não se torne uma ferramenta de perpetuação de preconceitos ou discriminações.

CONTRIBUIÇÕES DOS AUTORES

Todos os autores contribuíram com a revisão do trabalho e aprovaram a versão submetida.

AGRADECIMENTOS

Os autores agradecem ao suporte financeiro do Campus Sertãozinho do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP), pelo suporte financeiro do Programa Institucional de Bolsas de Iniciação Científica e Tecnológica do IFSP (PIBIFSP).

REFERÊNCIAS

GOMES, D. S. Inteligência Artificial: Conceitos e Aplicações. **Revista Olhar Científico** – Faculdades Associadas de Ariquemes. v. 01, n. 2, ago./dez. 2010. Disponível em: https://www.professores.uff.br/screspo/wp-content/uploads/sites/127/2017/09/ia_intro.pdf Acesso: 12 mai. 2023.

KRETTEK, C. ChatGPT: Milestone-Text-KI mit Game-Changer-Potenzial. **Die Unfallchirurgie**, n. 3, 2023. p. 252-254.

MAGALHÃES, Izabel. Introdução: a análise de discurso crítica. **D.E.L.T.A.**, vol. 4, n. 2, 2020. Universidade de Brasília. ISSN 2594-5173. Disponível em: <https://www.scielo.br/j/delta/a/LgkQwhZgkLdsMnvDLHh7znz/>. Acesso em: 15 jun. 2024.

NASCIMENTO, Silvana de Souza (Docente responsável); TEIXEIRA, Jacqueline Moraes; CUSTÓDIO, Crislei (Docentes ministrantes); OLIVEIRA, Felipe Gabriel (Monitor). **Poder e Performatividade Pública: Introdução ao pensamento de Michel Foucault e Judith Butler; Nas Fronteiras do Pensamento de Hannah Arendt e Judith Butler**. São Paulo: Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, Departamento de Antropologia, 2024.

RIBEIRO, G. A. M.; CORDEIRO, P. I. R. V.; FUMACH, D. M. O malware como meio de obtenção de prova e sua implementação no ordenamento jurídico brasileiro. **Rev. Bras. de Direito Processual Penal**, Porto Alegre, v. 8, n. 3, p. 1463-1500, set.-dez/2022. Disponível em: <https://www.scielo.br/j/rbdpp/a/rhHb6tynNX5rNH74mNGHSrj/?format=pdf&lang=pt> Acesso: 12 mai. 2023.

THE LANCET DIGITAL HEALTH. Editorial - ChatGPT: friend or foe? **The Lancet Digital Health**. 6 fev. 2023.

TRIGO, Luiz Gonzaga Godoi. **As milenares origens do preconceito de gênero**. São Paulo: Universidade de São Paulo, 2024.