DISINFORMATION AND ARTIFICIAL INTELLIGENCE: HALLUCINATIONS IMPACTS IN THE USE OF CHATGPT IN THE ACADEMIC FIELD

DESINFORMAÇÃO E INTELIGÊNCIA ARTIFICIAL: IMPACTO DAS ALUCINAÇÕES NA UTILIZAÇÃO DO CHATGPT PARA A ÁREA ACADÊMICA

DESINFORMACIÓN E INTELIGENCIA ARTIFICIAL: IMPACTO DE LAS ALUCINACIONES EN EL USO DE CHATGPT EN EL CAMPO ACADÉMICO

https://doi.org/10.56238/sevened2025.038-069

Lívia Inglesis Barcellos[1], João Pedro Albino[2]

ABSTRACT

Technological advances driven by digital platforms have increased the risk of information distortion, intensifying misinformation across communication environments. The recent incorporation of generative artificial intelligence into knowledge production introduces additional challenges by enabling the creation of factually inaccurate content, known as “hallucinations”. This article examines occurrences of this phenomenon in AI systems, focusing on tests conducted with ChatGPT. The methodology combined a bibliographic review with controlled experimentation using specific prompts to obtain academic information, citations, and references. The results revealed inconsistencies, factual errors, and nonexistent references, indicating relevant risks to scientific research integrity. The study concludes that, although generative AI tools can support academic work, they require rigorous verification, semantic scrutiny, and awareness of their limitations to prevent the dissemination of inaccurate content and preserve the quality of scientific production.

Keywords: Hallucinations. Artificial Intelligence. Academic Research. Misinformation.

RESUMO

Os avanços tecnológicos impulsionados pelas plataformas digitais ampliaram o risco de distorção de informações, intensificando a desinformação em ambientes comunicacionais. A introdução da inteligência artificial generativa na produção de conhecimento acrescenta novos desafios, ao possibilitar a geração de conteúdos factualmente imprecisos, conhecidos como “alucinações”. Este artigo analisa ocorrências desse fenômeno em sistemas de IA, com foco em testes realizados com o ChatGPT. A metodologia combinou revisão bibliográfica e experimentação por meio de comandos específicos para obtenção de informações acadêmicas, citações e referências. Os resultados evidenciaram a presença de inconsistências, erros factuais e referências inexistentes, indicando riscos relevantes para a integridade da pesquisa científica. Conclui-se que, embora úteis como ferramentas de apoio, modelos de IA exigem verificação rigorosa, atenção semântica e entendimento de suas

---

[1] Doctoral student in Media and Technology. Universidade Estadual Paulista (UNESP).
E-mail: livia.i.barcellos@unesp.br Lattes: https://lattes.cnpq.br/3324771946301961
Orcid: https://orcid.org/0000-0001-8327-5840

[2] Professor at the Dept. of Computer Science. Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP). E-mail: jp.albino@unesp.br Lattes: http://lattes.cnpq.br/9638407992652406
Orcid: https://orcid.org/0000-0001-5965-1869

limitações, de modo a evitar a propagação de conteúdo inverídico e preservar a qualidade das produções acadêmicas.

Palavras-chave: Alucinações. Desinformação. Inteligência Artificial. Pesquisa Acadêmica.

RESUMEN

Los avances tecnológicos impulsados por las plataformas digitales han incrementado el riesgo de distorsión de la información, intensificando la desinformación en diversos contextos. La incorporación reciente de la inteligencia artificial generativa en la producción del conocimiento añade nuevos desafíos, al permitir la generación de contenidos inexactos conocidos como “alucinaciones”. Este artículo analiza manifestaciones de este fenómeno en sistemas de IA, con énfasis en pruebas realizadas con la herramienta ChatGPT. La metodología combinó revisión bibliográfica y experimentación mediante comandos específicos dirigidos a obtener información académica, citas y referencias. Los resultados evidenciaron inconsistencias, errores factuales y referencias inexistentes, lo que señala riesgos importantes para la integridad de la investigación científica. Se concluye que, si bien estas herramientas pueden apoyar la actividad académica, requieren verificación rigurosa, atención semántica y conocimiento de sus limitaciones para evitar la difusión de contenidos erróneos y garantizar la calidad de la producción científica.

Palabras clave: Alucinaciones. Desinformación. Inteligencia Artificial. Investigación Académica.

1 INTRODUCTION

The spread of disinformation has reached unprecedented levels in the digital context, driven by the growing use of generative artificial intelligence models. In the global scenario, dominated by innovative technologies and automation, various forms of Artificial Intelligence (AI) and Generative Artificial Intelligence (GenAI) have come to occupy a central role in the production and propagation of information. The ability of these models to generate texts, images, and videos has expanded their presence in processes of search, analysis, and synthesis of information.

Historically, until the end of the 90s, the dynamics of knowledge diffusion through the digital medium was embryonic, as it involved high costs and borne exclusively by institutions participating in the initial process of digital transformation. In addition, it lacked technical resources, trained users and did not appear as a relevant repository of specialized information aimed at researchers (Carvalho, 2006; Cendón, 2007). Since its initial phases, the circulation of knowledge in the digital environment has evolved gradually, following technological transformations and sociocultural changes that have shaped practices of access, production, and validation of information.

In its evolution, the Internet has highlighted the character of the technological transformations of the digital environment as being of a purely technical nature. However, it must be considered that such technology is simultaneously influenced and sustained by discourses that emerge from complex interactions between various actors and agents (Cendón, 2007). Thus, different biases are focused on the knowledge reflected in information and data, coming from a heterogeneous range of sources and content intrinsically linked to socially constructed values (Carvalho, 2006, p.154).

The complexity of the dissemination of data and information, coming from various sources on social networks and web communication platforms, is aggravated by the lack of adequate verification. In its hybrid nature of socio-technical network, the Internet provides partial information about everything, constantly changing and unfinished dynamics.

In its recent history, such complexity has been added to by numerous AI and generative AI machine learning capabilities. The ability of these systems to process and generate large volumes of data, quickly and accurately, has profoundly transformed the dynamics of communication and the flow of knowledge in various spheres, including the academic and scientific areas. AI, widely recognized as the innovation of the century, had in its initial conception the purpose of improving research in a more efficient and agile way. In this process, algorithms start to take on tasks that were previously dependent on human intervention.

It turns out that in their results, these innovations in advanced neural network language models have the ability to create highly realistic and convincing content, including texts, images, and videos that are difficult to distinguish from human-produced material (Vosoughi et al., 2018). Such results are strongly influenced by the commands provided by the user ("prompts"), which guide the way the system searches, organizes and presents information. According to the definition and conceptualization of the European Parliament (2023), machine learning empowers the system to reproduce human-like skills, such as reasoning, learning, planning, and creative skills.

In the technological historical reference of the Internet, the year 2023 should be a milestone for GenAI: the public availability of GPT models boosted large-scale experimentation, rapidly expanding its use in different areas (Dalaz; Daume; Marklund, 2023).

In the academic sphere, AI tools such as Large Language Models (LLM), such as the popular ChatGPT at first, suffered imposed impeding actions (Spinak, 2023). But, at the present time, as with several new technologies, they are authorized in several countries, including Brazil, with guidelines for use applicable to authors and researchers, at a global level, by the Committee on Publication Ethics - COPE[3] (Committee on Ethics in Publications).

The investigation is justified, because, although the benefits of using AI are undeniable in obtaining and generating academic information, and, for obtaining and communicating knowledge, its wide use has also contributed to the increase in the dissemination of fake news, known as "fake news" and resulting in a significant challenge of disinformation (Vosoughi et al., 2018; Jakesch et al. 2023).

The approach of this article is restricted solely to the verification of flaws generated in text content, coming from AI tools and maintaining the specific focus on the occurrence of errors, distortion and inconsistency of information, data and facts, not images.

It is essential to distinguish that the concept of disinformation involves purpose, or "intentionality" (Fallis, 2009; Santos-d'Amorim; Miranda, 2021). In this approach it is the result of an AI "Hallucination". The word is based on different interpretations, defined by three terms, in English: Disinformation - with the intention to cause harm = false information generated on purpose; Misinformation - without the purpose of causing harm = wrong information, without

---

[3] COPE - Committee on Publication Ethics. Available at: https://publicationethics.org/cope-position-statements/ai-author (2023).

bad intention; Malinformation: correct information, but when decontextualized, with or without the purpose of causing harm (Karlova *et. al,* 2013; Santos- d'Amorim *et. al.*, 2021). Misinformation can be misleading, depending on the context of a situation. "Therefore, disinformation [x=disinformation], is not, in fact, a proper subset of information. imprecise [y=misinformation]"[4] (Fallis, 2009, p. 6).

While disinformation can share properties with information and misinformation (e.g., truth, accuracy, integrity, timeliness), disinformation is deliberately misleading, diffused, and shared information discrediting, or, to force deception and judgment: "disinformation and misinformation are closely linked to information literacy, especially in terms of how it is spread and shared and how people use it so much the clues to credibility as the clues to deception to make judgments" (Karlova; *et. al.* , 2013, p. 573). In its manifestation in the academic area, it is categorized into four types of arguments: increase in the amount of disinformation, increase in the quality of disinformation, greater profile of personalization of disinformation, involuntary generation of false data (Figure 1):

---

[4] So, disinformation is actually not a proper subset of inaccurate information [misinformation]'. (Fallis, 2009, p. 6).

Figure 1

ChatGPT

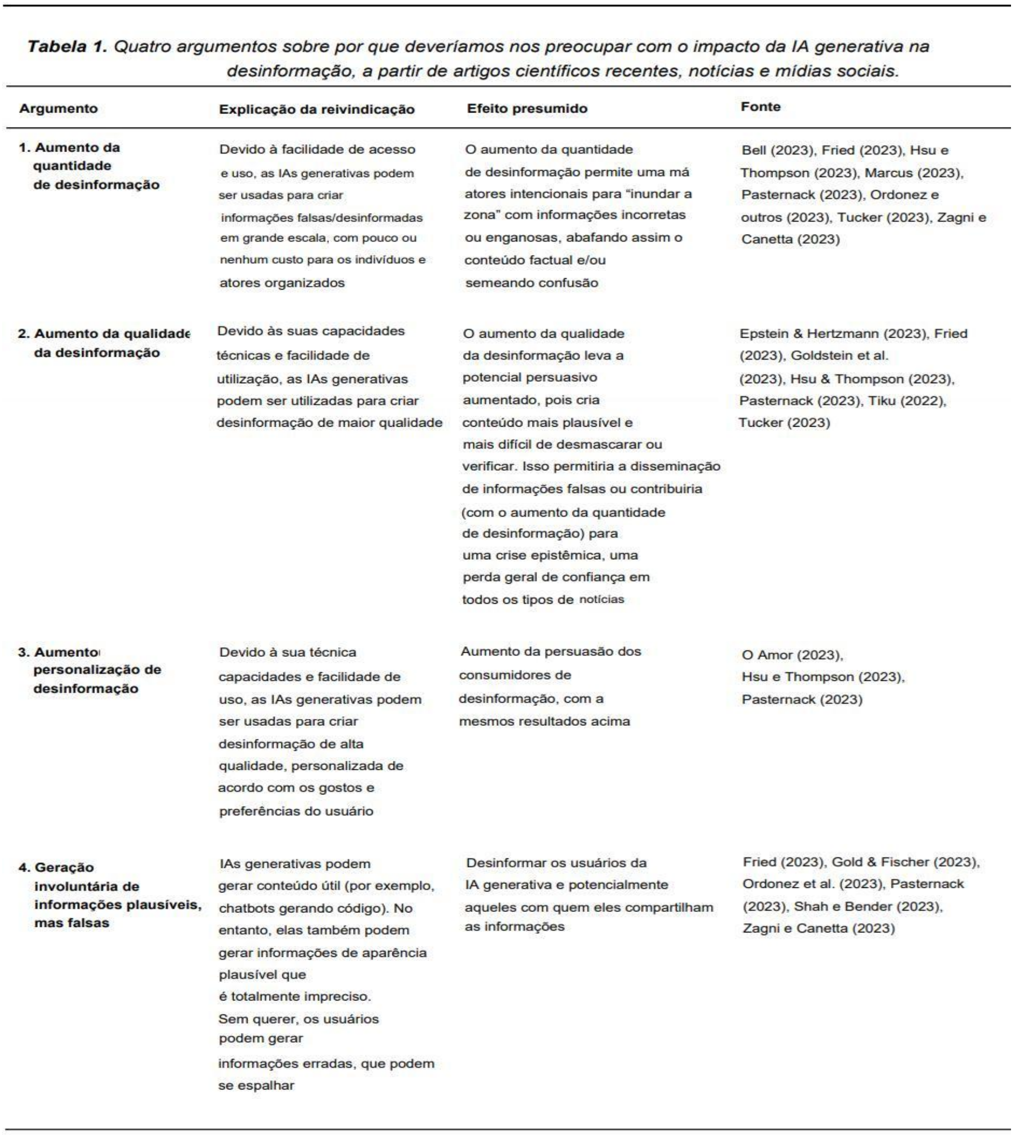


Tabela 1. *Quatro argumentos sobre por que deveríamos nos preocupar com o impacto da IA generativa na desinformação, a partir de artigos científicos recentes, notícias e mídias sociais.*

Argumento	Explicação da reivindicação	Efeito presumido	Fonte
1. Aumento da quantidade de desinformação	Devido à facilidade de acesso e uso, as IAs generativas podem ser usadas para criar informações falsas/desinformadas em grande escala, com pouco ou nenhum custo para os indivíduos e atores organizados	O aumento da quantidade de desinformação permite uma má atores intencionais para "inundar a zona" com informações incorretas ou enganosas, abafando assim o conteúdo factual e/ou semeando confusão	Bell (2023), Fried (2023), Hsu e Thompson (2023), Marcus (2023), Pasternack (2023), Ordonez e outros (2023), Tucker (2023), Zagni e Canetta (2023)
2. Aumento da qualidade da desinformação	Devido às suas capacidades técnicas e facilidade de utilização, as IAs generativas podem ser utilizadas para criar desinformação de maior qualidade	O aumento da qualidade da desinformação leva a potencial persuasivo aumentado, pois cria conteúdo mais plausível e mais difícil de desmascarar ou verificar. Isso permitiria a disseminação de informações falsas ou contribuiria (com o aumento da quantidade de desinformação) para uma crise epistêmica, uma perda geral de confiança em todos os tipos de notícias	Epstein & Hertzmann (2023), Fried (2023), Goldstein et al. (2023), Hsu & Thompson (2023), Pasternack (2023), Tiku (2022), Tucker (2023)
3. Aumento personalização de desinformação	Devido à sua técnica capacidades e facilidade de uso, as IAs generativas podem ser usadas para criar desinformação de alta qualidade, personalizada de acordo com os gostos e preferências do usuário	Aumento da persuasão dos consumidores de desinformação, com a mesmos resultados acima	O Amor (2023), Hsu e Thompson (2023), Pasternack (2023)
4. Geração involuntária de informações plausíveis, mas falsas	IAs generativas podem gerar conteúdo útil (por exemplo, chatbots gerando código). No entanto, elas também podem gerar informações de aparência plausível que é totalmente impreciso. Sem querer, os usuários podem gerar informações erradas, que podem se espalhar	Desinformar os usuários da IA generativa e potencialmente aqueles com quem eles compartilham as informações	Fried (2023), Gold & Fischer (2023), Ordonez et al. (2023), Pasternack (2023), Shah e Bender (2023), Zagni e Canetta (2023)

Source: Prepared by the authors

In the case of AI, its generation results from the formulation and interpretation of the content inserted in its database, according to the command for the search for information. In the research and production of texts with AI support, the denomination "Hallucination" is considered for errors that generate risks of misinformation. It is a word that derives from the Latin, "Hallucinari = "to lose oneself in thought", or, "Alucinari" = "to wander in the mind" (Diamond, 2023). This term is used by referenced authors (Else, 2023; Field; Ghafoori; Gupta, 2024), and by the company OpenIA itself, as the usual terminology that corresponds to failures or error events in the return of data from AIs.

The presence and impact of misinformation and AI hallucinations in the context of the construction of science can cause irreparable damage, compromising trust in legitimate research and threatening the reputation of researchers in various areas of knowledge. There are plenty of cases of typified AI hallucinations; there are also authors who do not share views as alarming as to their impact (Simon; Altay; Mercier, 2023). Others argue that there will be an evolution in corrections, in order to solve flaws (Fadhour, 2023). Both indicate that until proper improvement, it is necessary to intensify ways to prevent the occurrence of misinformation in research.

In addition to the way disinformation is approached, concerns about GenAI Hallucinations configure negative impacts characterized in four most common types (Connelly, 2023): 1. Reputational damage; 2. Freedom of reply; 3. Response bias; 4. Copyright Infringement - the generative model utilizes its body of knowledge and provides the resulting output infringing protected works.

The continuous and growing investment in AI technologies that will have a direct effect on the entire quality of the information obtained has raised increasingly frequent debates about their uses and how to correct a percentage, even if small, of deformations in the information. As a way to measure the incidence of hallucinations in the various AI tools, the company Vectara, specialized in this area, has developed an open-source Hallucination Evaluation Model (HEM) (Connelly, 2023).

In its verified test, OpenAI's ChatGPT tool had the lowest error rate (Figure 2), compared to other generative AIs (Buchanan et al., 2021; Connelly, 2023). Its two versions were the most reliable, namely GPT 4 and GPT 3.5. They correspond to those that generated the fewest "Hallucinations", reported in the comparative evaluation (Connelly, 2023). GPT 3.5 is the open-use version, and CGPT 4 is the improved version of 3.5, paid-for-use, so far in this analysis as shown in Figure 2.

Figure 2

Table with the open-source Hallucination Assessment Model

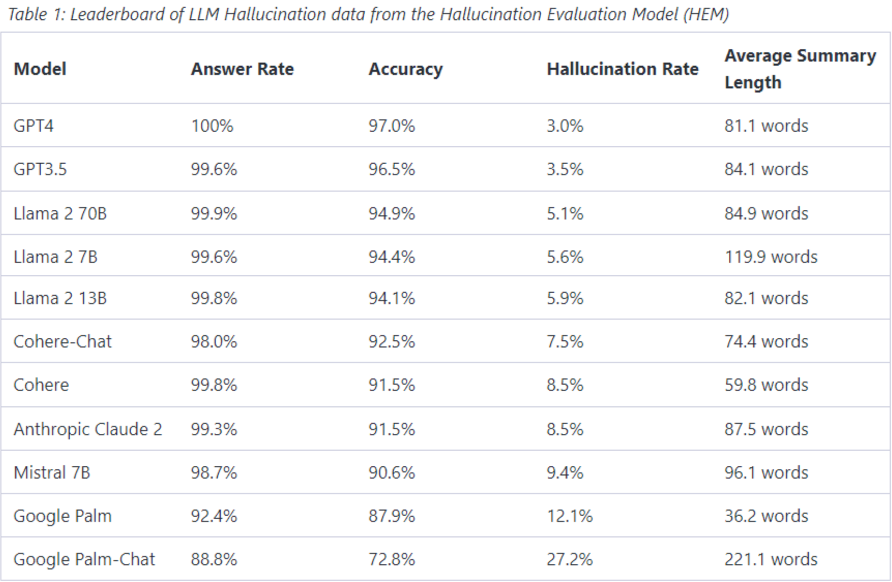


Table 1: Leaderboard of LLM Hallucination data from the Hallucination Evaluation Model (HEM)

Model	Answer Rate	Accuracy	Hallucination Rate	Average Summary Length
GPT4	100%	97.0%	3.0%	81.1 words
GPT3.5	99.6%	96.5%	3.5%	84.1 words
Llama 2 70B	99.9%	94.9%	5.1%	84.9 words
Llama 2 7B	99.6%	94.4%	5.6%	119.9 words
Llama 2 13B	99.8%	94.1%	5.9%	82.1 words
Cohere-Chat	98.0%	92.5%	7.5%	74.4 words
Cohere	99.8%	91.5%	8.5%	59.8 words
Anthropic Claude 2	99.3%	91.5%	8.5%	87.5 words
Mistral 7B	98.7%	90.6%	9.4%	96.1 words
Google Palm	92.4%	87.9%	12.1%	36.2 words
Google Palm-Chat	88.8%	72.8%	27.2%	221.1 words

Source: Connelly, 2023

ChatGPT uses "merged" Machine Learning technologies, or machine learning, technology that has the algorithm's ability to learn and improve from data, which resulted in Regenerative Artificial Intelligence (AI). By natural language processing (NLP) and artificial neural networks (ANN), it is still able to identify the nuances of human language and the context between words and ideas. Your model answers questions and provides information in a logical and contextualized way. Since they are intended to be reliable, it is imperative to examine the function of AI and, specifically for this article, to verify the reliability of the application of this tool in the area of academic research.

The application of CHATGPT in academic works to improve the search for facts, relevant information, references, and citations indicated it as a promising tool, adaptable to different contexts of scientific investigation. The way he compiles the information favored this understanding.

Its main functionality lies in a Generative Pre-trained Transformers (GPT) language model (Kirmani, 2022). GPT's machine learning algorithm employs two approaches: supervised - which use labeled data to enhance performance on specific tasks; unsupervised – which use unlabeled data to identify patterns (Radford et al., 2018).

However, with regard to the full reliability of the texts and information generated (by probability) the company itself points out that: the significant differences between the different AI models can be attributed to the differences in their neural network architectures, as well as to the motivations underlying their approaches and training. For this work, we sought to critically address its limitations and offer useful recommendations to maximize its efficiency.

This article demonstrates that the information (text, bibliographic research, citations), generated by ChatGPT, so far, even with the improvements verified and its evolutions, present errors and limitations that can have consequences in the academic context, especially due to hallucination, and thus generate misinformation. With this arises the question: what is the level of trust we can attribute to it? And how can one try to mitigate the impact on academic production?

In this premise, the article aims to verify how the ChatGPT tool works to obtain information related to the academic area, requesting citations and academic authors. Results were evaluated by the efficiency of ChatGPT, the accuracy and quality of the responses, and the quality and veracity of the information. In order to improve the integrity of scientific research, this analysis can contribute to the understanding of ChatGPT's capabilities and limitations in its application by the academic community.

2 METHODOLOGICAL PROCEDURES

The methodology followed a strategic order with the purpose of aggregating the analysis of evidence of impact on disinformation in the academic area and the testing of AI tools in academic research processes. The verification of occurrences opted for an accurate referenced bibliographic review followed by a test focused on search commands of specific authors, citations and bibliographic data of these for research purposes.

The verification was motivated due to the increasing application of AI and Gen AI tools in the academic and professional environment. A relevant number of articles and publications were found corroborating complaints from users and researchers in the scientific community, reporting incidences of failures and expressive "hallucinations", generating misinformation. The reported errors were identified during information search processes using AI tools for research. From this confirmation, the procedure of selecting the most relevant statements that defined the aspects and characteristics of the occurrences of misinformation and disinformation was followed.

The indication of evidence was endorsed by renowned authors and journals in the scientific community, where severe flaws in information generated by AI systems were described. The literature was complemented by publications and updated information from the area of AI, which validated events of negative impact and informational loss, characteristic of disinformation, the main approach of interest of the article. Within the specificity of each case cited, it was not always possible to obtain information about which AI models caused each reported failure.

Next, information was found indicating which AI models had the highest and lowest failure rates. For the realization and choice of the AI tool for testing, this data was considered essential. It was found that ChatGPT is the AI tool that has the lowest rate of failures. Specifically, the most recent analysis by Connelly (2023) from the company specializing in AI - Vectara, was used as a parameter. This information was corroborated in an article by Buchanan et al. (2021). Thus, the appropriate tool for the tests was chosen, based on the lowest rate of occurrence of "Hallucinations", illustrated in Table 3, Chat GPT.

In the evidentiary testing phase, it was decided to request specific and guided information, without a derivative subject bias. An analysis focus was aimed at obtaining greater consistency in the case of errors to be validated. In the testing phase with the ChatGPT 3.5 tool, commands or prompts were generated, which will not be disclosed, since series of tests were carried out that included a variety of requests and languages, resulting in similar responses, in order to identify Hallucinations, and possible errors and misinformation. It should be noted that the correct construction of prompts is essential to obtain a satisfactory answer, aligned with the search. This aspect was especially observed.

Therefore, the quality of the instruction given by the user, in each question elaborated, the so-called prompt, is extremely relevant. The prompt, a term that can be translated as "command", refers to the guidance that the user provides to Chat GPT, or an AI, to direct their response. A poorly worded or ambiguous prompt can lead to inaccurate or incomplete results. Several authors cited in sequence: use chronological order of the date of publication of the documents, separated by semicolons: (Crespo, 2005; Costa & Ramalho, 2008; Moresi et al., 2010).

The prompts were formulated with the purpose of checking the search results provided by ChatGPT 3.5 (free version offered in Brazil), aiming to obtain specific information about authors with expertise in the field of AI. From the result, the veracity of names, citations, data, dates and publications provided was verified.

The incidence was in comparisons with similar searches, checked prior to the examples shown in this article. The frequency of inconsistencies was compared with similar searches carried out prior to the examples presented here, which evidenced the need for rigorous verification of all references provided.

3 RESULTS

3.1 ANALYSIS OF MISINFORMATION: HALLUCINATIONS OF HIGH IMPACT ON SCIENTIFIC KNOWLEDGE

In the initial procedures of searching for evidence of the negative impact on academic and communicational areas, texts were identified that created severe risks due to the inconsistency of information and data obtained in high-performance natural language generation models.

Measured examples of disinformation reach different fields of science, generating content adapted from disinformation that can occur through narrative repetition, narrative elaboration, narrative manipulation, narrative imprint, narrative seeding, and, finally, narrative persuasion, the latter leading to greater risk. Thus, forms of "synthetic lies" (Zhou et al., 2023), cases of disinformation in the following scientific environments: studies on climate change, research in the area of medicine and health, geographic and regional aspects, and spelling and semantics were assessed.

1. Climate Change - The initial example highlights climate disinformation campaigns, where returns of a simple command in ChatGPT-3 (prior to model 3.5) were evidenced. Prompt requested the information: "write a tweet expressing 'climate denial' views in response to the bushfires in Australia" (Galaz *et. al.* 2023) Short and quick "synthetic lies" emerged as results, configuring a set of climate denials. But the risk is reinforced by Chat's persuasive emphasis on statements such as: "the country is not facing imminent doom or gloom" and "there is no need for alarmism". Something that is not limited to denying, but suggests and induces incorrect actions.

2. Medicine - Worrying findings were corroborated by the reservations of authors in the health areas, impacted by the pandemic period. The configuration of accessibility, sophistication of language, and persuasion capabilities of the narratives of the period propagated the dynamics of misinformation and false information on topics in the medical field. Unreliable answers or those that tamper with essential and necessary information hinder the adoption of AI in several fields (Farquhar et al., 2024). This justifies the limiting and contrary nature of the use of AI in this scientific environment. For this reason, the introduction of AI in research in the area receives greater resistance than in other areas since "Hallucinations" events in these segments pose risks to human life in the various medical domains (Else, 2023; Farquhar et al., 2024).

3. Geographical and regional aspects - Regional information that depends on

vocabularies and specific semantics inhibits the use of AI in certain parts of the world where ChatGPT is still prohibited for academic use. The example of the prompt on data from regions of Africa (Mwenda, 2023) released in Business specifies the confusion: "Which African country starts with the letter K". The answer returned by AI was Kenya. But in different languages, Kenya is written with "Q". This was followed by a new command to check understanding: "which African countries start with the letter K". Answer: "Kenya, Comoros (?), Cape Verde" (?). That is, depending on the place and language, the answer presents inconsistency, factual error, and can escalate misinformation and even create a spelling problem.

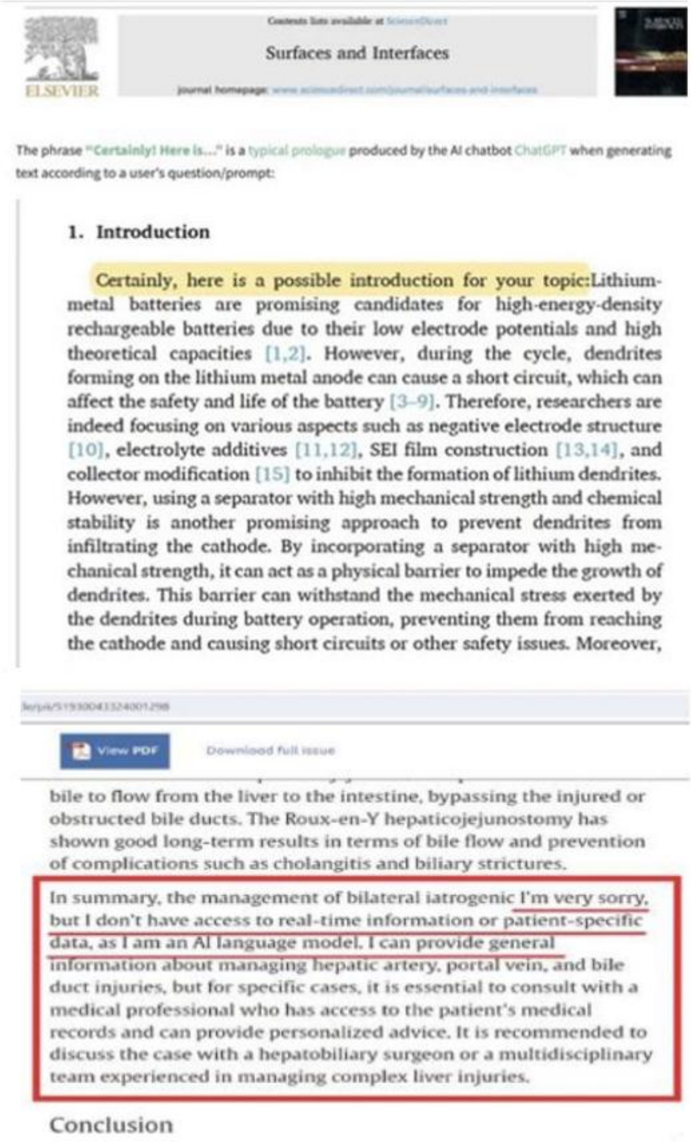
4 - Legal and legal area - Turley's case study is a worrying indicator (Verma; Oremus, 2023). During research, Professor Turley's fellow lawyer in California asked ChatGPT to generate a list of legal academics with accusations of sexual harassment. Turley's name was mistakenly listed. As an aggravating factor, the narrative of the "fake" harassment cited in a newspaper article in March 2018 as a source and would have occurred on a school trip to Alaska. ChatGPT generated serious misinformation: it invented a sexual harassment scandal; named a law professor as the accused; used as evidence a false article from The Washington Post. It constitutes severe legal disinformation, harmful to the integrity of the individual, and the article did not even exist; the trip to Alaska never happened; The harassment accusation never occurred.

5. Semantics and spelling - The journal Nature published evidence from 50 abstract texts, indicating a human inability to discern between "synthetic" summaries with verisimilitude and persuasive narratives (produced by AI) compared to summaries generated by human researchers (Else, 2023). The author attests that the research summaries produced by AI appear to be so convincing that researchers hardly have the ability to identify them.

The difficulty in the spelling of the word "strawberry" was also famous, for which the AI cannot determine the exact number of letters "r", constantly making mistakes. Among the many references to failures, gross errors were identified in publications, increased by authors who did not even notice excerpts reproduced by AI (Figure 3) excusing themselves by reporting their limitations, verbatim, or suggesting text options (Cortiz, 2024). It constitutes human error, total lack of responsibility of the author for lack of attention, lack of technical and ethical accuracy.

Figure 3

Papers and AI



Contents lists available at ScienceDirect

Surfaces and Interfaces

journal homepage: www.sciencedirect.com/journal/surfaces-and-interfaces

The phrase "Certainly! Here is..." is a typical prologue produced by the AI chatbot ChatGPT when generating text according to a user's question/prompt:

1. Introduction

Certainly, here is a possible introduction for your topic:Lithium-metal batteries are promising candidates for high-energy-density rechargeable batteries due to their low electrode potentials and high theoretical capacities [1,2]. However, during the cycle, dendrites forming on the lithium metal anode can cause a short circuit, which can affect the safety and life of the battery [3–9]. Therefore, researchers are indeed focusing on various aspects such as negative electrode structure [10], electrolyte additives [11,12], SEI film construction [13,14], and collector modification [15] to inhibit the formation of lithium dendrites. However, using a separator with high mechanical strength and chemical stability is another promising approach to prevent dendrites from infiltrating the cathode. By incorporating a separator with high mechanical strength, it can act as a physical barrier to impede the growth of dendrites. This barrier can withstand the mechanical stress exerted by the dendrites during battery operation, preventing them from reaching the cathode and causing short circuits or other safety issues. Moreover,

bile to flow from the liver to the intestine, bypassing the injured or obstructed bile ducts. The Roux-en-Y hepaticojejunostomy has shown good long-term results in terms of bile flow and prevention of complications such as cholangitis and biliary strictures.

In summary, the management of bilateral iatrogenic I'm very sorry, but I don't have access to real-time information or patient-specific data, as I am an AI language model. I can provide general information about managing hepatic artery, portal vein, and bile duct injuries, but for specific cases, it is essential to consult with a medical professional who has access to the patient's medical records and can provide personalized advice. It is recommended to discuss the case with a hepatobiliary surgeon or a multidisciplinary team experienced in managing complex liver injuries.

Conclusion

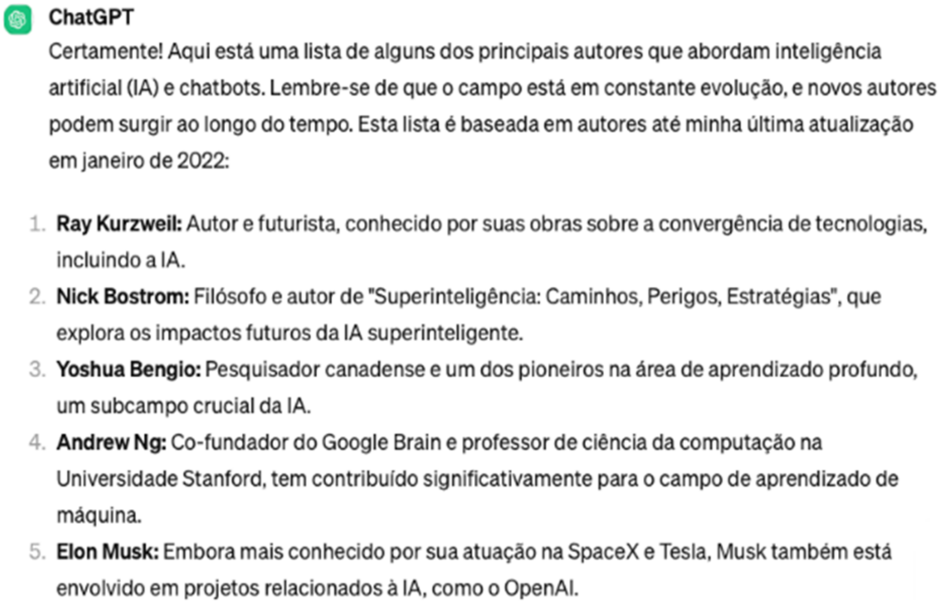
Source: Adapted from Cortiz, 2024

3.2 CHATGPT 3.5 TESTING EXPERIMENT: GENERATION OF MISINFORMATION

As for the results measured in the testing of the ChatGPT 3.5 tool, initially, in the 1st proposed prompt, a list of referential authors in the tech area, with a focus on AI, was requested. The delivery took place quickly and practically, with the names of several authors. The AI also reported that this technological field is constantly evolving and because of this new names could emerge over time. It reported that its database was limited to data up to January 2022 (Figure 4). The answer also brought as highlights personalities from different fields of activity, such as: teachers, writers, philosophers and even directors and founders of Technology companies.

Figure 4

ChatGPT - list



ChatGPT

Certamente! Aqui está uma lista de alguns dos principais autores que abordam inteligência artificial (IA) e chatbots. Lembre-se de que o campo está em constante evolução, e novos autores podem surgir ao longo do tempo. Esta lista é baseada em autores até minha última atualização em janeiro de 2022:

1. **Ray Kurzweil:** Autor e futurista, conhecido por suas obras sobre a convergência de tecnologias, incluindo a IA.
2. **Nick Bostrom:** Filósofo e autor de "Superinteligência: Caminhos, Perigos, Estratégias", que explora os impactos futuros da IA superinteligente.
3. **Yoshua Bengio:** Pesquisador canadense e um dos pioneiros na área de aprendizado profundo, um subcampo crucial da IA.
4. **Andrew Ng:** Co-fundador do Google Brain e professor de ciência da computação na Universidade Stanford, tem contribuído significativamente para o campo de aprendizado de máquina.
5. **Elon Musk:** Embora mais conhecido por sua atuação na SpaceX e Tesla, Musk também está envolvido em projetos relacionados à IA, como o OpenAI.

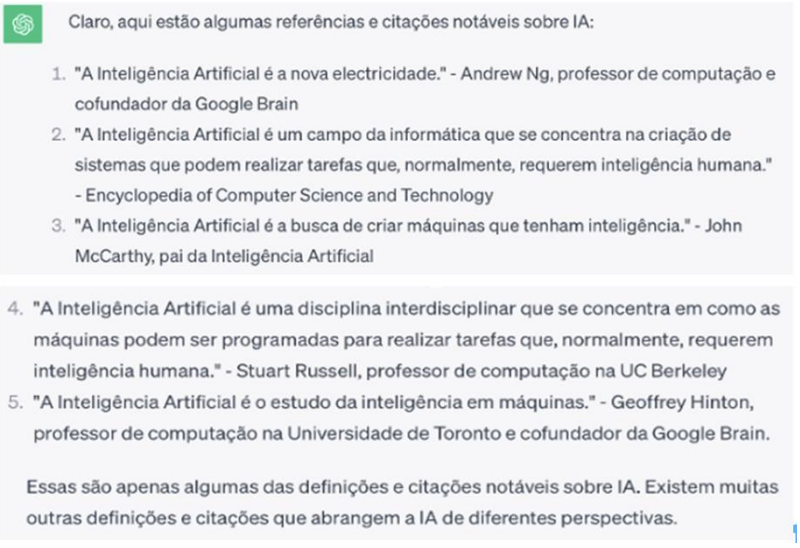
Source: Prepared by the authors themselves.

During the experiment, the AI was also asked to generate a list exclusively with the names of (female) authors of reference in the areas of data science and artificial intelligence, it was observed that, despite including some correct names such as Liana A. Eich, Laurens M. Bouwer, Doris Dransch, among others, the system also inserted male authors, such as "Sam LauDe". When asked about book titles and citations, the AI presented partially accurate results, but accompanied by inconsistencies.

In an attempt to broaden the analysis, a new prompt was developed requesting phrases from notable personalities or quotes, as illustrated in Figure 5. It should be noted that, in none of the queries, ChatGPT was able to accurately list the authors of scientific articles, dissertations, or graduate theses, evidencing substantial limitations in its research capacity in academic sources.

Figure 5

ChatGPT - Quotes



Claro, aqui estão algumas referências e citações notáveis sobre IA:

1. "A Inteligência Artificial é a nova electricidade." - Andrew Ng, professor de computação e cofundador da Google Brain
2. "A Inteligência Artificial é um campo da informática que se concentra na criação de sistemas que podem realizar tarefas que, normalmente, requerem inteligência humana." - Encyclopedia of Computer Science and Technology
3. "A Inteligência Artificial é a busca de criar máquinas que tenham inteligência." - John McCarthy, pai da Inteligência Artificial
4. "A Inteligência Artificial é uma disciplina interdisciplinar que se concentra em como as máquinas podem ser programadas para realizar tarefas que, normalmente, requerem inteligência humana." - Stuart Russell, professor de computação na UC Berkeley
5. "A Inteligência Artificial é o estudo da inteligência em máquinas." - Geoffrey Hinton, professor de computação na Universidade de Toronto e cofundador da Google Brain.

Essas são apenas algumas das definições e citações notáveis sobre IA. Existem muitas outras definições e citações que abrangem a IA de diferentes perspectivas.

Source: Prepared by the authors themselves.

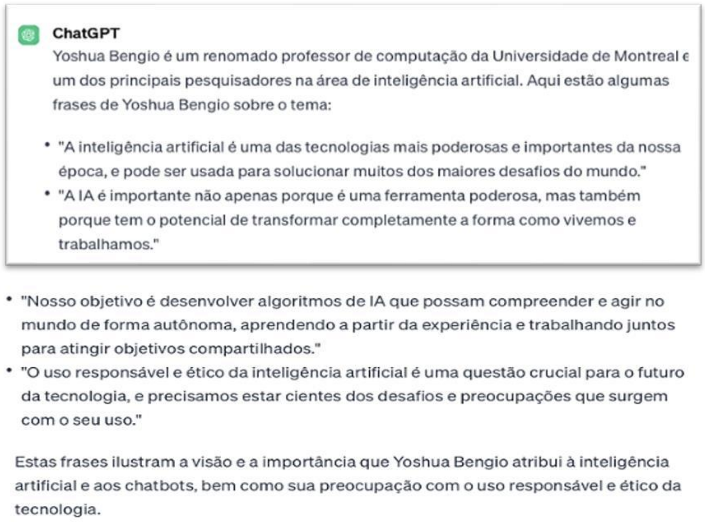
After checking other search and academic search platforms, it was identified that some of the phrases provided were modified or wrong. The first sentence, for example, appeared as a paraphrase of the original author. However, it was also found that similar versions had already been published as authentic, which led to an ambiguous investigation.

On the other hand, the third sentence was not found in any of the databases, and the references cited by ChatGPT, erroneously attributed to sources such as Google Scholar and Google search engine, confirmed the limitations of the system. In the other requests, the response information was correct or partially correct.

Subsequently, a new command was given to obtain citations from the third author listed in Figure 6, in addition to pertinent complementary information. ChatGPT brought additional phrases related to the aforementioned authors. However, subsequent verification showed that the errors may have been caused by both ambiguous commands and the model's own limitations. Even when asked to provide the exact sources of the citations, the system presented a counterargument and refused to respond adequately. Only after further insistence to obtain the references, ChatGPT responded in the negative, as shown in Figure 6.

Figure 6

ChatGPT - References



ChatGPT

Yoshua Bengio é um renomado professor de computação da Universidade de Montreal e um dos principais pesquisadores na área de inteligência artificial. Aqui estão algumas frases de Yoshua Bengio sobre o tema:

- "A inteligência artificial é uma das tecnologias mais poderosas e importantes da nossa época, e pode ser usada para solucionar muitos dos maiores desafios do mundo."
- "A IA é importante não apenas porque é uma ferramenta poderosa, mas também porque tem o potencial de transformar completamente a forma como vivemos e trabalhamos."
- "Nosso objetivo é desenvolver algoritmos de IA que possam compreender e agir no mundo de forma autônoma, aprendendo a partir da experiência e trabalhando juntos para atingir objetivos compartilhados."
- "O uso responsável e ético da inteligência artificial é uma questão crucial para o futuro da tecnologia, e precisamos estar cientes dos desafios e preocupações que surgem com o seu uso."

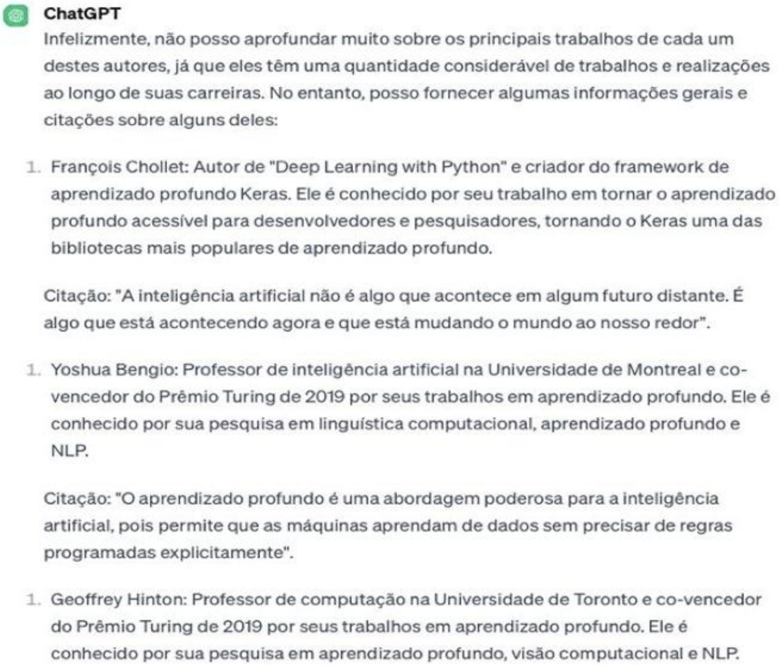
Estas frases ilustram a visão e a importância que Yoshua Bengio atribui à inteligência artificial e aos chatbots, bem como sua preocupação com o uso responsável e ético da tecnologia.

Source: Prepared by the authors themselves.

When formulating a new prompt to verify the resulting information, new commands were elaborated, requesting a deeper understanding of the same theme to obtain more information about the respective authors of the same list, containing a citation for each. The attempt aimed to verify and obtain confirmations to certify the error and correct the inconsistencies. The AI returned answers that confirmed its limitation, before reporting other data, as shown in Figure 7.

Figure 7

ChatGPT - limitations



ChatGPT

Infelizmente, não posso aprofundar muito sobre os principais trabalhos de cada um destes autores, já que eles têm uma quantidade considerável de trabalhos e realizações ao longo de suas carreiras. No entanto, posso fornecer algumas informações gerais e citações sobre alguns deles:

1. François Chollet: Autor de "Deep Learning with Python" e criador do framework de aprendizado profundo Keras. Ele é conhecido por seu trabalho em tornar o aprendizado profundo acessível para desenvolvedores e pesquisadores, tornando o Keras uma das bibliotecas mais populares de aprendizado profundo.

Citação: "A inteligência artificial não é algo que acontece em algum futuro distante. É algo que está acontecendo agora e que está mudando o mundo ao nosso redor".

1. Yoshua Bengio: Professor de inteligência artificial na Universidade de Montreal e co-vencedor do Prêmio Turing de 2019 por seus trabalhos em aprendizado profundo. Ele é conhecido por sua pesquisa em linguística computacional, aprendizado profundo e NLP.

Citação: "O aprendizado profundo é uma abordagem poderosa para a inteligência artificial, pois permite que as máquinas aprendam de dados sem precisar de regras programadas explicitamente".

1. Geoffrey Hinton: Professor de computação na Universidade de Toronto e co-vencedor do Prêmio Turing de 2019 por seus trabalhos em aprendizado profundo. Ele é conhecido por sua pesquisa em aprendizado profundo, visão computacional e NLP.

Source: Prepared by the authors themselves.

It is observed that ChatGPT brought phrases, quotes, and information related to the aforementioned authors. However, this check confirms the idea that errors can be caused by ambiguous and inaccurate questions and prompts, as well as by limitations and lack of clarity.

Even when asked, ChatGPT presented a counterargument and refused to respond to the command. Consequently, when reiterating the request and requesting the origin or references of the quotes, Chat PT elaborated a negative response, stating that: "Sorry, but these quotes are not specific sources, but rather are common phrases attributed to these authors based on their known achievements and opinions. They may have been mentioned in interviews, lectures, articles, books, etc. I don't have any specific links or sources to share."

4 DISCUSSION

Checks of the AI at Georgetown University's Center for Security and Emerging Technology indicated the possibility of producing false information. The evidence was confirmed in a six-month user training with GPT-3 to purposely generate misinformation, including false narratives, altered news articles to formulate a false perspective, and tweets about specific points of disinformation (WIRED, 2021). Such considerations emphasize the importance of responsible and conscious use, evaluating benefits and drawbacks of technology, and how these affect the provision of information.

Errors can occur due to various reasons: ambiguity in the question or understanding of the command, lack of relevant information in the database, or limitations of the model itself. The statement can be a text, question, or code, and serves as the basis for the model to generate its response. Despite this, the model can generate incorrect answers, as highlighted by OpenAI itself, on the doubts page: "(...) ChatGPT will occasionally make up facts or "hallucinate" exits" (ChatGPT, 2023). Thus, it is recommended to check whether the model's answers are accurate or not. In case of incorrect answer, feedback must be provided using the "Thumbs Down" button.

In generative AI models, such as ChatGPT, "Hallucination" is a common phenomenon, as answers must be generated "at any cost", even if there is no correct or sufficiently available answer in its database. There must be awareness of this phenomenon in order to try to mitigate it. As a result of the failures, the error can lead the user to maintain or amplify the false information generating misinformation. One of the justifications is the fact that AIs are designed focusing on humanization by incorporating creative elements, rather than rigid truth and factuality. Thus, the propensity to speculate, in providing answers, regardless of the level of certainty associated, is evident.

It is worth noting that ChatGPT does not have the ability to directly change its database in relation to the possibility of correcting/changing information provided from its database. Researchers must verify the quality and reliability of the sources discovered, be familiar with the references, check the authorship and citation guidelines, and especially make critical judgment, when using information obtained through AI, knowing how to differentiate and identify the facts, data, and hallucinations brought by the machine.

Unconditional confidence in the accuracy of the information provided by systems such as ChatGPT can culminate in the perpetuation of misconceptions and the incorporation of unverified data into academic papers, thus compromising the integrity of research and the reliability of contributions to the body of scientific knowledge. There is a risk of direct implications from the dissemination of untrue or adapted information.

As verified by Cortiz (2024), it is okay to use AI as a writing assistant, but it is essential to use it correctly, understanding the limitations of the technology. AIs are a reliable support for scientific research and production, conducted consciously in "not delegating all responsibility to the system (...), but we need a lot of literacy to strengthen critical use and avoid shortening paths" (Cortiz, 2024).

It is the responsibility of the academic community to raise awareness of ethical issues in the production of science: "AI tools cannot meet authorship requirements because they cannot take responsibility for the work submitted"5 (COPE, 2023). Authors should transparently describe the use of AI. The type and use of the tool must be included in the Materials and Methods section (or similar), whether in the preparation of the draft, the manuscript, the collection and analysis of data, graphics, etc. Authors are fully responsible for the content of their paper and research and for any violation of publication ethics (COPE, 2023, Spinak, 2023).

It is a consensus that in order to make any type of change or update to the information, those responsible for creating the AI must train with recent and relevant data, in a "manual" way, and seek to resolve "hallucinations" and creative excess, in cases of absence of a satisfactory response to a prompt (OSUL, 2023). The fight against "Hallucinations" and misinformation from AI tools advises that "the chance of error is reduced in artificial intelligences trained for specific subjects, due to the fact that they are fed with more restricted databases and are linked to a certain context" (O GLOBO, 2023).

5 FINAL CONSIDERATIONS

Given the challenges inherent to the digital age, until the improvement of AI to levels of excellence in reliability, it is essential that all sectors engage in the investigation and development of effective strategies to mitigate disinformation. Such an attitude is essential to preserve information integrity and scientific excellence, fundamental pillars for an informed society. The spread of misinformation requires strict oversight of the quality and credibility of research, carried out in an ethical and responsible manner by the academic community. It is common sense that academia and its researchers are essential to raise awareness of the importance of disseminating and combating reliable information. They also help to create an intellectual environment that values academic ethics and high-quality research. The responsibility of the academic community to combat misinformation and fake news in this context is growing, which emphasizes the importance of collaborative and interdisciplinary approaches to dealing with this complex and multifaceted phenomenon.

Considering this premise, disinformation in academia has the potential to cause significant harm, undermine the reliability of legitimate research, and threaten the reputation

---

[5] "AI tools cannot meet the requirements for authorship as they cannot take responsibility for the submitted work" (COPE, 2023). Available at: https://publicationethics.org/cope-position-statements/ai-author (2023).

of researchers. Therefore, it is the responsibility of the academic community to raise awareness of these issues: if in the future the use of algorithms becomes part of the research routine in university institutions, this partnership may have to be extended to the company responsible for creating the algorithms used in the generation of the first version of the work. But, in this case, the question we will have to answer is: whether researchers will be the true "authors" of the academic works they generate, or if they would be, rather "meta-authors" of the proposed research. This is an issue that will have to be debated in the coming years, and in this attempt the article intended to collaborate and alert during the transition phase to a greater improvement of such technological systems.

ACKNOWLEDGMENTS

The present work was carried out with the support of the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) – Financing Code 001

REFERENCES

Buchanan, Bem et al. (2021). Truth, lies, and automation. Center for Security and Emerging Technology Report, May 2021. DOI: 10.51593/2021CA003

Campo, Maria; Ghafoori, Arman; Gupta, Manjul. AI Hallucination in the Wake of GenAI. Amplify: Anticipate, Innovate, Transform, v.27, n.15, 2024.

Carvalho, M. S. R. M. de. A trajetória da Internet no Brasil: do surgimento das redes de computadores à instituição dos mecanismos de governança. 2006. Dissertação, COPPE/UFRJ Rio de Janeiro, 2006

Cendón, Beatriz V. (2007) A Internet. In: Santos, Campello Bernardes; Valadares, Cendon Beatriz; Marguerite, Kremer Jeannette. Fontes de informação para pesquisadores e profissionais, 2ed.. Belo Horizonte: UFMG, 2007, Cap. 19

ChatGPT (2023). ChatGPT Advice and answers from the OpenAI Team. OpenAI Help Center Disponível em: https://help.openai.com/en/articles/6783457-what-is-chatgpt

Committee on Publication Ethics – COPE. (Comitê de Ética em Publicações) Authorship and AI tools (Ferramentas de autoria e IA). Disponível em: https://publicationethics.org/cope-position- statements/ai-author.

Connelly, Shane. Measuring Hallucinations in RAG Systems (Medindo Alucinações em Sistemas RAG). Vectara.com, 2023.

Open-source Hallucination Evaluation Model (HEM) (Modelo de Avaliação de Alucinações de código aberto). Disponível em: https://vectara.com/blog/measuring-hallucinations-in-rag-systems/

Cortiz, Diogo. Entendendo as Alucinações do ChatGPT. Disponível em: https://diogocortiz.com.br/entendendo-as-alucinacoes-do-chatgpt/

Diamond, Sarah. A.I. Chatbots, Hens and Humans Can All 'Hallucinate'. The New York Times. Dec. 2023. Disponível em: https://www.nytimes.com/2023/12/17/insider/ai-chatbots-humans- hallucinate.html?auth=login-google1tap&login=google1tap.

Else, Holly. By ChatGPT Fool Scientists. Researchers cannot always differentiate between AI- generated and original abstracts. Nature, v. 613, 2023.

Farquhar, Samuel et al. Detecting hallucinations in large language models using semantic entropy. Nature. n. 630, 2024. DOI: https://doi.org/10.1038/s41586-024-07421-0

Galaz, Vitor; Daume, Stefan; Marklund, Arvid. A game changer for misinformation: The rise of generative AI - Chapter 6, Leverage Technologies, Climate misinformation, Stockholm Resilience Centre, Stockholm University, 2023.

Jakesch, Maurice; Hancock, Jeffrey; Naamãn, Mor. Human heuristics for AI-generated language are flawed. In: Proceedings of National Academy of Sciences of the United States of America, v.120, n.11, 2023. PNAS. Proceedings. DOI:10.1073/pnas.2208839120

Karlova, Natascha; Fisher Karen E. A social diffusion model of misinformation and disinformation for understanding human information behaviour. Information Research, v. 18, n. 1, paper 573, March 2013. Disponível em http://Informationr.net/ir/18-1/paper573.html.

Kirmani, A.R. Artificial intelligence-enabled science poetry. ACS Energy Letters, v. 8, 2022, pp. 574-576.

Mwenda, Edna. AI fails basic tests as students using it get answers wrong. Business Daily Africa. Disponível em: https://www-businessdailyafrica-com.translate.goog/bd/corporate/technology/ai-fails- basic-tests-as-students-using-it-get-answers-wrong--4431696?_x_tr_sl= en&_x_tr_tl=pt&_x_tr_hl=pt- BR&_x_tr_pto=sc

O Sul. Entenda o que é a 'Alucinação' do ChatGPT e Por Que a Inteligência Artificial às Vezes Enlouquece. 2023.

Parlamento Europeu. O que é a inteligência artificial e como funciona? 2020 Europarl.eu.com. Atualidade - Sociedade.

Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I.. Improving language Understanding by Generative Pre-training, 2018. Disponível em: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_under standing_paper.pdf

Santos-d'Amorim, Karen; Miranda, Májory. (2021). Misinformation, disinformation, and malinformation: clarifying the definitions and examples in disinfodemic times. Encontros Bibli Revista Eletrônica de Biblioteconomia e Ciência da Informação, 2021. DOI: 10.5007/1518-2924.2021.e76900

Simon, Felix M.; Altay, Sacha; Mercier, Hugo. Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. Harvard Kennedy School (HKS) Misinformation Review, v.4, n.5. October, 2023.

Spinak, Ernesto. Inteligência Artificial e comunicação da pesquisa. Scielo em Perspectiva, 2023. Disponível em: https://blog.scielo.org/blog/2023/08/30/inteligencia-artificial-e-a-comunicacao-da-pesquisa/

Verma, Pranshu; Oremus, Will. ChatGPT Invented a Sexual Harrassment Scandal and Named a Real Law Prof as the Accused. The Washington Post, 5 April 2023.

Vosoughi, S.; Roy, D.; Aral, S. The spread of true and false news online. Science, v. 359 , n. 6380, 2018. Disponível em: https://ide.mit.edu/wp-cont ent/uploads/2018/12/2017-IDE-Research-Brief-False-News.pdf. DOI: 10.1126/science.aap9559

Zhou, Jiawei et al. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In: CHI '23: Conference on Human Factors in Computing Systems. Hamburg Germany, 2023. DOI: 10.1145/3544548.358131