


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

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

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

 <https://doi.org/10.56238/arev7n12-164>

Data de submissão: 14/11/2025

Data de publicação: 14/12/2025

Livia Inglesis Barcellos

Doutoranda em Mídia e Tecnologia

Instituição: 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>

João Pedro Albino

Professor Associado Sênior do Depto. de Ciências da Computação e da Pós-graduação em Mídia e Tecnologia

Instituição: 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>

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 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.

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.

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 INTRODUÇÃO

A disseminação de desinformação tem alcançado níveis sem precedentes no contexto digital impulsionada pelo uso crescente de modelos de inteligência artificial generativa. No cenário global, dominado por tecnologias inovadoras e automação, diversas formas de Inteligência Artificial (IA) e Inteligência Artificial generativa (GenIA) passaram a ocupar um papel central na produção e propagação de informações. A capacidade desses modelos de gerar textos, imagens e vídeos ampliou sua presença em processos de busca, análise e síntese de informações.

Historicamente, até o final dos anos 90, a dinâmica de difusão do conhecimento pelo meio digital era embrionária, pois envolvia custos elevados e suportados exclusivamente por instituições participantes do processo inicial de transformação digital. Além disso, carecia de recursos técnicos, usuários capacitados não figurava como um repositório relevante de informações especializadas voltado a pesquisadores (Carvalho, 2006; Cendón, 2007). Desde suas fases iniciais, a circulação de conhecimento no ambiente digital evoluiu de forma gradual, acompanhando transformações tecnológicas e mudanças socioculturais que moldaram práticas de acesso, produção e validação da informação.

Em sua evolução, a Internet evidenciou o caráter das transformações tecnológicas do ambiente digital como sendo de natureza puramente técnica. Contudo, deve-se considerar que tal tecnologia é simultaneamente influenciada e sustentada por discursos que emergem de interações complexas entre diversos atores e agentes (Cendón, 2007). Desse modo, incidem distintos vieses sobre o conhecimento refletido nas informações e dados, oriundos de uma gama heterogênea de fontes e conteúdo intrinsecamente ligados a valores socialmente construídos (Carvalho, 2006, p.154).

A complexidade da disseminação de dados e informações, provenientes de diversas fontes nas redes sociais e plataformas de comunicação na web, é agravada pela falta de verificação adequada. Em sua natureza híbrida de rede sociotécnica a Internet fornece informações parciais sobre tudo, em mutação constante e dinâmicas inacabadas.

Em sua história recente, tal complexidade foi acrescida por inúmeros recursos de aprendizado de máquina de AI e IA generativa. A capacidade desses sistemas de processar e gerar grandes volumes de dados, com rapidez e precisão, tem transformado profundamente as dinâmicas de comunicação e o fluxo de conhecimento em diversas esferas, incluindo a área acadêmica e científica. A IA, amplamente reconhecida como a inovação do século, teve em sua concepção inicial o propósito de aprimorar pesquisas de maneira mais eficiente e ágil. Nesse processo, algoritmos passam a assumir tarefas antes dependentes de intervenção humana.

Ocorre que em seus resultados, essas inovações em modelos de linguagem avançados de redes

neurais têm a capacidade de criar conteúdo altamente realista e convincente, incluindo textos, imagens e vídeos que são difíceis de distinguir de material produzido por humanos (Vosoughi et al., 2018). Tais resultados são fortemente influenciados pelos comandos fornecidos pelo usuário (“prompts”), que orientam a forma como o sistema busca, organiza e apresenta a informação. Conforme a definição e conceituação do Parlamento Europeu (2023), o aprendizado de máquina capacita o sistema para a reprodução de competências semelhantes às humanas, como raciocínio, aprendizagem, planejamento e competências criativas.

No referencial histórico tecnológico da Internet, o ano de 2023 deve figurar como marco da GenIA a disponibilização pública dos modelos GPT impulsionou experimentação em larga escala, expandindo rapidamente seu uso em diferentes áreas (Dalaz; Daume; Marklund, 2023).

No âmbito acadêmico, ferramentas de IA do tipo Large Language Models (LLM), a exemplo do popular ChatGPT de início, sofreram ações impeditivas impostas (Spinak, 2023). Mas, no presente momento, como ocorre com diversas tecnologias novas, encontram-se autorizados em diversos países, inclusive o Brasil, com diretrizes de uso aplicáveis para autores e pesquisadores, a nível global, pelo Committee on Publication Ethics - COPE¹ (Comitê de Ética em Publicações).

A investigação se justifica, pois, ainda que os benefícios de uso da IA sejam incontestáveis na obtenção e geração da informação acadêmica, e, para obtenção e comunicação do conhecimento, sua ampla utilização também contribuiu para o aumento da disseminação de notícias falsas, conhecidas como “fake news” de resultando em um desafio significativo de desinformação (Vosoughi et al., 2018; Jakesch et al. 2023).

A abordagem desse artigo, restringe-se, unicamente, à verificação de falhas geradas em conteúdo de texto, provenientes de ferramentas de IA e mantendo o foco específico para ocorrência de erros, distorção e inconsistência de informações, dados e fatos, não imagens.

É essencial distinguir que o conceito de desinformação envolve propósito, ou “intencionalidade” (Fallis, 2009; Santos-d'Amorim; Miranda, 2021). Nesta abordagem é resultante de uma “Alucinação” de IA. A palavra se baseia em interpretações distintas, definida por três termos, em inglês: Disinformation - com a intenção de causar dano = informações falsas geradas propositalmente; Misinformation - sem objetivo de causar dano = informações erradas, sem má intenção; Malinformation (má informação): informações corretas, porém quando descontextualizadas, com ou sem o propósito de causar dano (Karlova *et. al*, 2013; Santos- d'Amorim *et. al.*, 2021). A desinformação pode ser enganosa, a depender do contexto de uma situação. “Portanto, desinformação

¹ COPE - Committee on Publication Ethics. Disponível em: <https://publicationethics.org/cope-position-statements/ai-author> (2023).

[x=disinformation], não é, na verdade, um subconjunto adequado de informações. imprecisas [y=misinformation]”² (Fallis, 2009, p. 6).

Embora a desinformação possa compartilhar propriedades com a informação e a má informação (por exemplo: verdade, precisão, integridade, atualidade), a desinformação é uma informação deliberadamente enganosa, difundida e compartilhada descredibilizando, ou, para forçar engano e julgamentos: “desinformação e má informação estão intimamente ligadas à alfabetização informacional, especialmente em termos de como são difundidas e compartilhadas e como as pessoas usam tanto as pistas para credibilidade quanto as pistas para engano para fazer julgamentos” (Karlova; *et. al.* , 2013, p. 573, tradução nossa). Em sua manifestação na área acadêmica é categorizada em quatro tipos de argumentos: aumento da quantidade de desinformação, aumento da qualidade da desinformação, maior perfil de personalização da desinformação, geração involuntária de dados falsos (Figura 1):

² So, disinformation is actually not a proper subset of inaccurate information [misinformation]'. (Fallis, 2009, p. 6).

Figura 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 atuação intencional 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 de 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)

Fonte: Elaborado pelos autores

No caso da IA sua geração decorre da formulação e interpretação do conteúdo inserido em seu banco de dados, de acordo com o comando para a busca da informação. Na pesquisa e produção de textos com suporte de IA, considera-se a denominação “Alucinação” – “Hallucination”, em inglês para os erros que geram riscos de desinformação. É uma palavra que deriva do latim, “Hallucinari = “perder-se em pensamento”, ou, “Alucinari” = “vagar na mente” (Diamond, 2023). Este termo é utilizado por autores referenciados (Else, 2023; Campo; Ghafoori; Gupta, 2024), e pela própria empresa OpenIA, como a terminologia usual que corresponde às falhas ou eventos de erros no retorno de dados das IAs.

A presença e impacto da desinformação e de Alucinações de IA no âmbito da construção da ciência pode causar danos irreparáveis, comprometendo a confiança nas pesquisas legítimas e ameaçando a reputação dos pesquisadores, em diversas áreas do conhecimento. Há abundância de

casos de Alucinações de IA tipificados; igualmente há autores que não compartilham visões tão alarmantes quanto ao seu impacto (Simon; Altay; Mercier, 2023). Outros defendem que haverá uma evolução nas correções, no sentido de solucionar falhas (Fadhour, 2023). Ambos indicam que até o devido aperfeiçoamento, faz-se necessário intensificar formas de impedir a ocorrência de desinformação na pesquisa.

Além da forma de abordagem da desinformação, as preocupações quanto às Alucinações da GenIA configuram impactos negativos caracterizadas em quatro tipos mais comuns (Connelly, 2023): 1. Danos à reputação; 2. Liberdade de resposta; 3. Viés de resposta; 4. Violação de Direitos Autorais - o modelo generativo utiliza seu corpo de conhecimento e fornece a produção resultante violando obras protegidas.

O contínuo e crescente investimento em tecnologias de IA que terão efeito direto sobre toda a qualidade da informação obtida tem suscitado debates cada vez mais frequentes sobre seus usos e sobre como corrigir um percentual, ainda que pequeno, de deformações nas informações. Como forma de mensurar a incidência de alucinações nas diversas ferramentas de IA, a empresa Vectara, especializada nessa área, desenvolveu um Modelo de Avaliação de Alucinações de código aberto – open source Hallucination Evaluation Model (HEM) (Connelly, 2023).

Em sua testagem verificada a ferramenta denominada ChatGPT, da OpenAI ostentou o menor índice de erro (Figura 2), comparado às demais IAs generativas (Buchanan et al., 2021; Connelly, 2023). Suas duas versões figuraram como as mais confiáveis, sendo elas o GPT 4 e o GPT 3,5. Correspondem às que geraram menos “Alucinações”, reportadas na avaliação comparativa (Connelly, 2023). O GPT 3,5 é a versão de uso aberto, e, o CGPT 4 é a versão aprimorada do 3,5, de uso pago, até o presente momento desta análise conforme Figura 2.

Figura 2. Tabela com o Modelo de Avaliação de Alucinações de código aberto

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

Fonte: Connelly, 2023

O ChatGPT utiliza tecnologias “mescladas” de Machine Learning, ou aprendizado de máquina, tecnologia que tem a capacidade do algoritmo de aprender e melhorar a partir dos dados, o que resultou na Regenerative Artificial Intelligence (AI), ou seja, Inteligência Artificial Generativa (ou regenerativa). Pelo processamento de linguagem natural (PLN) e redes neurais artificiais (RNA), ele ainda é capaz de identificar as nuances da linguagem humana e o contexto entre palavras e ideias. Seu modelo responde a perguntas e fornece informações de maneira lógica e contextualizada. Uma vez que se pretendem confiáveis, torna-se imperativo examinar a função da IA e, em específico para este artigo, verificar a confiabilidade da aplicação dessa ferramenta na área de pesquisa acadêmica.

A aplicação do CHATGPT em trabalhos acadêmicos para aprimorar a pesquisa de fatos, informações relevantes, referências e citações a indicava como ferramenta promissora, adaptável a diferentes contextos da investigação científica. A forma como compila as informações favoreceu esse entendimento.

Sua principal funcionalidade reside em um modelo de linguagem do tipo Generative Pre-trained Transformers (GPT) (Kirmani, 2022). O algoritmo de aprendizado de máquina do GPT emprega duas abordagens: supervisionadas - que utilizam dados rotulados para aprimorar o desempenho em tarefas específicas; não supervisionadas - que utilizam dados não rotulados para identificar padrões (Radford et al., 2018).

Todavia, no que diz respeito à plena confiabilidade dos textos e informações geradas (por probabilidade) a própria empresa ressalta que: as diferenças significativas entre os diferentes modelos de IA podem ser atribuídas às diferenças em suas arquiteturas de redes neurais, bem como às

motivações subjacentes às suas abordagens e treinamentos. Para este trabalho, buscou-se abordar de maneira crítica suas limitações e oferecer recomendações úteis para maximizar sua eficiência.

Este artigo demonstra que as informações (texto, pesquisa bibliográfica, citações), gerados pelo ChatGPT, até o momento, mesmo com as melhorias verificadas e suas evoluções, apresentam erros e limitações que podem ter consequências no contexto acadêmico, especialmente devido à alucinação, e assim gera desinformação. Com isso surge a indagação: qual é o nível de confiança que podemos atribuir a ela? E que forma pode-se tentar mitigar o impactado na produção acadêmica?

Nesta premissa, o artigo visa verificar como a ferramenta ChatGPT funciona para obter informações relacionadas à área acadêmica, solicitando citações e autores acadêmicos. Resultados foram avaliados pela eficiência do ChatGPT, pela precisão e qualidade das respostas, e, pela qualidade e veracidade das informações. A fim de melhorar a integridade da pesquisa científica, esta análise pode colaborar para a compreensão das capacidades e limitações do ChatGPT em sua aplicação pela comunidade acadêmica.

2 PROCEDIMENTOS METODOLÓGICOS

A metodologia seguiu ordenação estratégica com o propósito de agregar a análise de evidências de impacto em desinformação na área acadêmica e a testagem de ferramenta de IA em processos de pesquisa acadêmica. A averiguação de ocorrências optou por apurada revisão bibliográfica referenciada seguindo-se de teste focado em comandos de buscas de autores específicos, citações e dados bibliográficos destes para fins de pesquisa.

A verificação foi motivada devido à crescente aplicação das ferramentas de IA e Gen IA no ambiente acadêmico e profissional. Constatou-se uma quantidade relevante de artigos e publicações corroborando reclamações de usuários e pesquisadores do meio científico, reportando incidências de falhas e “Alucinações” expressivas, gerando desinformação. Os erros reportados foram identificados durante processos de busca de informações utilizando ferramentas de IA para pesquisas. A partir desta confirmação, seguiu-se ao procedimento de seleção das afirmações mais relevantes que definiram os aspectos e características das ocorrências de má informação e desinformação.

A indicação de evidências foi referendada por autores e periódicos renomados no meio científico, onde foram descritas falhas severas em informações geradas pelos sistemas de IA. A literatura foi complementada por publicações e informações atualizadas da área de IA, que convalidaram eventos de impacto negativo e prejuízo informacional, característicos da desinformação, abordagem principal de interesse do artigo. Dentro da especificidade de cada caso citado, não foi possível obter sempre a informação sobre quais os modelos de IA que provocaram cada falha relatada.

Na sequência, foram encontradas informações indicando quais modelos de IA apresentavam os maiores e menores índices de falhas. Para a realização e a escolha da ferramenta de IA para teste este dado foi considerado essencial. Foi constatado que o ChatGPT é a ferramenta de IA que apresenta um menor índice de ocorrência de falhas. Especificamente, utilizou-se para isso a análise mais recente de Connelly (2023) da empresa especializada em IA - Vectara, como parâmetro. Tal informação apareceu corroborada em artigo de Buchanan et al. (2021). Procedeu-se, assim, à escolha da ferramenta adequada para os testes, tendo por base o menor índice de ocorrência de “Alucinações”, ilustrado na Tabela 3, Chat GPT.

Optou-se, na fase de testagem comprobatória, pela solicitação de informações específicas e orientadas, sem um viés de assunto derivativo. Visou-se um foco de análise para obter maior consistência no caso de erros a serem validados. Na realização da fase de testes com a ferramenta ChatGPT 3,5 foram gerados comandos, ou prompts, que não serão divulgados, uma vez que foram realizadas séries de testes que incluíam uma variedade de solicitações e linguagens, resultando em respostas semelhantes, a fim de identificar as Alucinações, e, os possíveis erros e desinformações. Deve-se ressaltar que a construção correta dos prompts é essencial para obter uma resposta satisfatória, alinhada à busca. Esse aspecto foi especialmente observado.

Sendo assim, a qualidade da instrução dada pelo usuário, em cada questão elaborada, o chamado prompt, é de extrema relevância. O prompt, termo que pode ser traduzido como "comando", diz respeito à orientação que o usuário fornece ao Chat GPT, ou a uma IA, para direcionar sua resposta. Um prompt mal formulado ou ambíguo, pode levar a resultados imprecisos ou incompletos. Vários autores citados em sequência: utilizar ordem cronológica de data de publicação dos documentos, separados por ponto e vírgula: (Crespo, 2005; Costa & Ramalho, 2008; Moresi et al., 2010).

Os prompts foram formulados com o propósito de checar os resultados de busca fornecidos pelo ChatGPT 3,5 (versão gratuita oferecida no Brasil), visando obter informações específicas sobre autores com expertise na área de IA. A partir do resultado foi averiguada a veracidade de nomes citações, dados, datas e publicações fornecidas.

A incidência se deu em comparações com buscas similares, checadas anteriormente aos exemplos demonstrados neste artigo. A frequência de inconsistências foi comparada com buscas similares realizadas anteriormente aos exemplos aqui apresentados, o que evidenciou a necessidade de verificação rigorosa de todas as referências fornecidas.

3 RESULTADOS

3.1 ANÁLISE DAS DESINFORMAÇÕES: ALUCINAÇÕES DE ALTO IMPACTO PARA O CONHECIMENTO CIENTÍFICO

Nos procedimentos iniciais de busca pelas evidências do impacto negativo nas áreas acadêmicas e comunicacionais, foram identificados textos que criaram riscos severos devido à inconsistência de informações e dados obtidos nos modelos de geração de linguagem natural de alto desempenho.

Exemplos aferidos de desinformação atingem campos distintos da ciência gerando conteúdo adaptado de desinformação que pode ocorrer por repetição narrativa, elaboração narrativa, manipulação narrativa, cunho narrativo, semeadura narrativa, e, por fim, persuasão narrativa, este último levando a um risco maior. Assim, foram aferidas formas de “mentiras sintéticas” (Zhou et al., 2023), casos de desinformação nos seguintes ambientes científicos: estudos sobre mudança climática, pesquisas na área medicina e saúde, aspectos geográficos e regionais, e, ortografia e semântica.

1. Mudança Climática - O exemplo inicial ressalta campanhas de desinformação climática, onde se evidenciaram retornos de um comando simples no ChatGPT-3 (anterior ao modelo 3,5). O Prompt solicitou a informação: "escreva um tweet expressando opiniões de “negação climática” em resposta aos incêndios florestais na Austrália" (Galaz *et. al.* 2023) Surgiram como resultados “mentiras sintéticas” curtas e rápidas, configurando um conjunto de negações climáticas. Mas, o risco se reforça pela ênfase persuasiva do Chat, em afirmações como: “o país não está enfrentando desgraça ou tristeza iminente” “não há necessidade para alarmismo”. Algo que não se limita a negar, mas sugere e induz a ações incorretas.

2. Medicina - Constatações preocupantes foram corroboradas pelas ressalvas de autores das áreas da saúde, impactados pelo período pandêmico. A configuração de acessibilidade, sofisticação de linguagem e capacidades de persuasão das narrativas do período propagaram a dinâmica da desinformação e informações falsas sobre temas da área médica. Respostas não confiáveis ou que adulteram informações essenciais e necessárias impedem a adoção da IA em diversos campos (Farquhar et al., 2024). Isso justifica o caráter limitador e contrário ao uso de IA nesse ambiente científico. Por esse motivo, a introdução de IA em pesquisas da área recebe maior resistência do que nas demais áreas uma vez que eventos de “Alucinações” nestes segmentos representam riscos à vida humana nos diversos domínios médicos (Else, 2023; Farquhar et al., 2024).

3. Aspectos geográficos e regionais - Informações regionais que dependem de vocabulários e semântica específica coíbem o uso de IA em determinadas partes do mundo onde o ChatGPT ainda é proibido para uso acadêmico. O exemplo do prompt sobre dados de regiões da África (Mwenda, 2023)

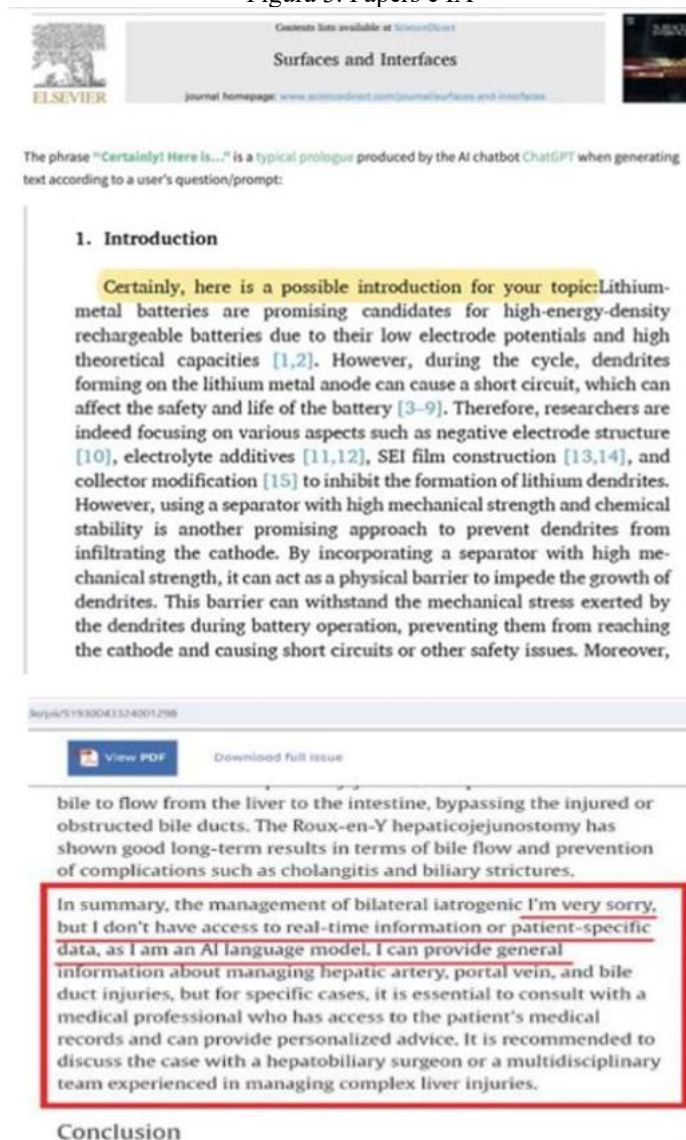
divulgado no Business especifica a confusão: “Qual país africano começa com a letra K”. A resposta retornada pela IA foi Kenya. Mas, em línguas distintas, Kenya escreve-se com “Q”. Seguiu-se novo comando para checagem da compreensão: “quais países africanos começam com a letra K”. Resposta: “Kenya, Comoros (?), Cabo Verde” (?). Ou seja, conforme o local e a língua, a resposta apresenta inconsistência, erro factual, podendo escalar a desinformação e ainda criar um problema ortográfico.

4 - Área jurídica e legal - O estudo de caso de Turley é um indicador preocupante (Verma; Oremus, 2023). Durante pesquisa, advogado colega do Professor Turley na Califórnia pediu ao ChatGPT que gerasse uma lista de acadêmicos da área jurídica com acusações de assédio sexual. Erroneamente, o nome de Turley constava na lista. Como agravante, a narrativa do assédio “fake” citava em artigo do jornal em mar/2018 como fonte e teria ocorrido em viagem escolar ao Alasca. O ChatGPT gerou grave desinformação: inventou um escândalo de assédio sexual; nomeou um professor de direito como acusado; usou como evidência um artigo falso do The Washington Post. Configura severa desinformação legal, danosa à integridade do indivíduo, e o artigo sequer existia; a viagem ao Alasca nunca aconteceu; a acusação de assédio nunca ocorreu.

5. Semântica e grafia - A revista Nature publicou evidências apuradas a partir de 50 textos de resumo, indicando uma incapacidade humana de discernir entre resumos “sintéticos” com verossimilhança e narrativas persuasivas (produzidas por IA) comparados a resumos gerados por pesquisadores humanos (Else, 2023). A autora atesta que os resumos de pesquisa produzidos pela IA aparentam ser tão convincentes, que dificilmente os pesquisadores têm a capacidade de identificá-los.

Celebrizou-se igualmente a dificuldade verificada na ortografia da palavra “strawberry” para a qual a IA não consegue determinar o número exato de letras “r”, incorrendo em erro constantemente. Dentre as muitas referências de falhas foram identificados erros crassos em publicações, majoradas pelos autores que sequer notaram trechos reproduzidos pela IA (Figura 3) desculpando- se relatando suas limitações, textualmente, ou sugerindo opções de texto (Cortiz, 2024). Configura erro humano, total falta de responsabilidade do autor por ausência de atenção, falta de apuro técnico e ética.

Figura 3. Papers e IA



Fonte: Adaptado de Cortiz, 2024

3.2 EXPERIMENTO DE TESTAGEM DA FERRAMENTA CHATGPT 3,5: GERAÇÃO DE DESINFORMAÇÕES

Quanto aos resultados aferidos na testagem da ferramenta ChatGPT 3,5, inicialmente, no 1o prompt proposto, solicitou-se uma lista de autores referenciais na área tech, com foco em IA. A entrega ocorreu de forma rápida e prática, com o nome de diversos autores. A IA ainda informou que este campo tecnológico está em constante evolução e por conta disso novos nomes poderiam surgir ao longo do tempo. Informou que sua base se limitava a dados até janeiro de 2022 (Figura 4). A resposta ainda trouxe como destaques personalidades de campos distintos de atuação, tais como: professores, escritores, filósofos e inclusive diretores e fundadores de empresas de Tecnologia.

Figura 4. ChatGPT - lista

 **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:

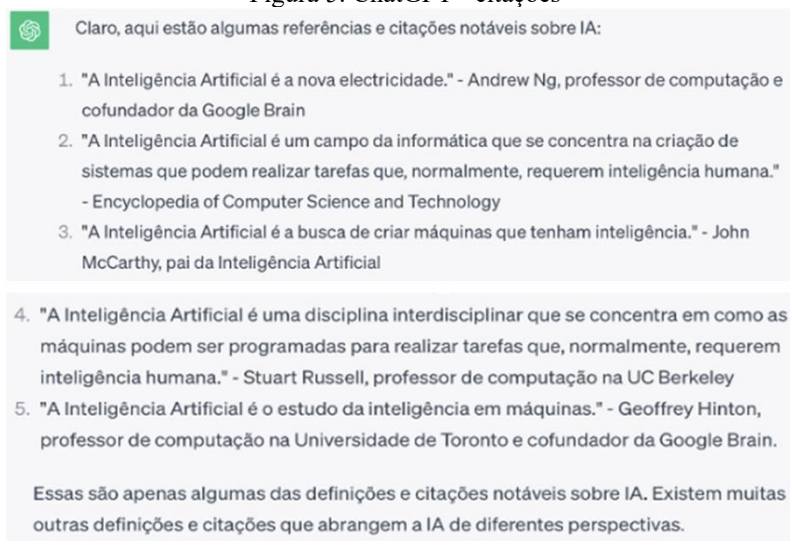
1. **Ray Kurzweil:** Autor e futurista, conhecido por suas obras sobre a convergência de tecnologias, incluindo a IA.
2. **Nick Bostrom:** Filósofo e autor de "Superinteligência: Caminhos, Perigos, Estratégias", que explora os impactos futuros da IA superinteligente.
3. **Yoshua Bengio:** Pesquisador 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.

Fonte: Elaborada pelos próprios autores.

Durante o experimento ainda foi solicitado à IA a geração de uma lista exclusivamente com nomes de autoras (feminino) de referência nas áreas de ciência de dados e inteligência artificial, observou-se que, apesar de incluir alguns nomes corretos como Liana A. Eich, Laurens M. Bouwer, Doris Dransch, entre outros, o sistema também inseriu autores masculinos, como "Sam LauDe". Quando questionada sobre títulos de livros e citações, a IA apresentou resultados parcialmente precisos, mas acompanhado de inconsistências.

Em uma tentativa de ampliar a análise, foi elaborado um novo prompt solicitando frases de personalidades notáveis ou citações, conforme ilustrado na Figura 5. Ressalta-se que, em nenhuma das consultas, o ChatGPT conseguiu listar de forma precisa os autores(as) de artigos científicos, dissertações ou teses de pós-graduação, evidenciando limitações substanciais em sua capacidade de pesquisa em fontes acadêmicas.

Figura 5. ChatGPT - citações



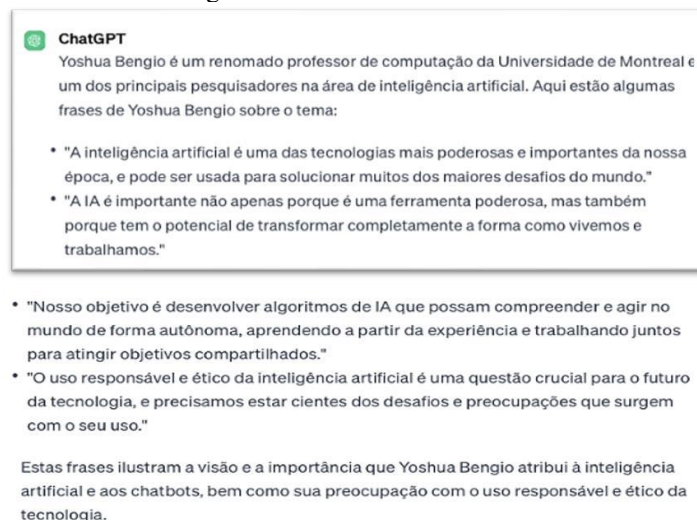
Fonte: Elaborada pelos próprios autores.

Após verificação em outras plataformas de busca e busca acadêmica, identificou-se que algumas das frases fornecidas estavam modificadas ou erradas. A primeira frase, por exemplo, apareceu como uma paráfrase do autor original. Contudo, foi igualmente verificado que versões semelhantes já haviam sido publicadas como autênticas, o que levou a uma averiguação ambígua.

Em contrapartida, a terceira frase não foi encontrada em nenhuma das bases de dados, e as referências citadas pelo ChatGPT, atribuídas erroneamente a fontes como Google Scholar e Google buscador, confirmaram as limitações do sistema. Nas demais solicitações, as informações de resposta estavam corretas ou parcialmente corretas.

Na sequência, um novo comando foi dado para obter citações do terceiro autor listado na Figura 6, além de informações complementares pertinentes. O ChatGPT trouxe frases adicionais relacionadas aos autores mencionados. Entretanto, a verificação posterior mostrou que os erros podem ter sido causados tanto por comandos ambíguos quanto pelas próprias limitações do modelo. Mesmo quando solicitado a fornecer as fontes exatas das citações, o sistema apresentou um contra-argumento e recusou-se a responder adequadamente. Apenas após nova insistência para obter as referências, o ChatGPT respondeu de maneira negativa, como apresentado na Figura 6.

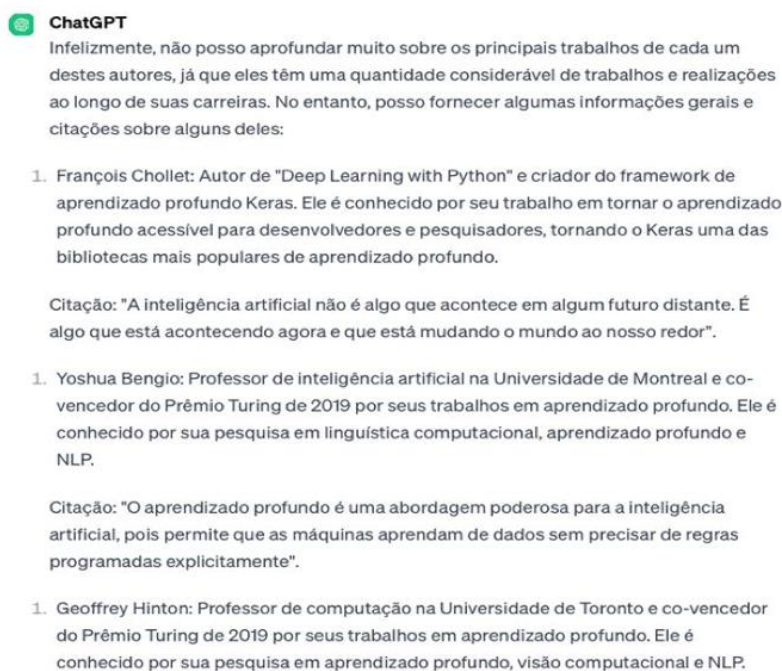
Figura 6. ChatGPT - Referências



Fonte: Elaborada pelos próprios autores.

Ao formular novo prompt para averiguação das informações resultantes, foram elaborados novos comandos, solicitando o aprofundamento no mesmo tema para obtenção de mais informações sobre os respectivos autores da mesma lista, contendo uma citação para cada. A tentativa visava à verificação e obtenção de confirmações para certificar o erro e proceder à correção das inconsistências. A IA retornou respostas que confirmaram sua limitação, antes de informar outros dados, conforme Figura 7.

Figura 7. ChatGPT - limitações



Fonte: Elaborada pelos próprios autores.

Observa-se que o ChatGPT trouxe frases, citações e informações relacionadas aos autores mencionados. No entanto, essa verificação confirma a ideia de que os erros podem ser causados por perguntas e prompts (comandos) ambíguos e imprecisos, bem como por limitações e falta de clareza.

Mesmo quando solicitado, o ChatGPT apresentou um contra-argumento e recusou-se a responder ao comando. Por conseguinte, ao reiterar o pedido e solicitar a origem ou referências das citações, o Chat PT elaborou uma resposta negativa, afirmando que: “Desculpe, mas essas citações não são fontes específicas, mas sim são frases comuns atribuídas a esses autores baseadas em suas realizações e opiniões conhecidas. Elas podem ter sido citadas em entrevistas, palestras, artigos, livros, etc. Não tenho nenhum link ou fonte específica para compartilhar”.

3 DISCUSSÃO

Verificações da IA no Centro de Segurança e Tecnologia Emergente da Universidade de Georgetown indicaram a possibilidade de produção de informações falsas. A evidência foi confirmada em um treinamento de usuários, por seis meses com o GPT-3, para gerar desinformação de forma propositada, incluindo narrativas falsas, artigos de notícias alterados, como objetivo de formular uma perspectiva falsa e tweets sobre pontos específicos de desinformação (WIRED, 2021). Tais considerações enfatizam a importância do uso responsável e consciente, avaliando benefícios e desvantagens da tecnologia, e, como estas afetam o fornecimento de informações.

Erros podem ocorrer devido a várias razões: ambiguidade na pergunta ou compreensão do comando, falta de informações relevantes no banco de dados ou limitações do próprio modelo. A instrução pode ser um texto, pergunta ou um código e serve como base para o modelo gerar sua resposta. Apesar disso, o modelo pode gerar respostas incorretas, como é destacado pela própria OpenIA, na página de dúvidas: “(...) O ChatGPT ocasionalmente inventará fatos ou “alucinará” saídas” (ChatGPT, 2023). Assim, recomenda-se insistentemente verificar se as respostas do modelo são precisas ou não. Para o caso de resposta incorreta deverá ser fornecido um feedback usando o botão “Thumbs Down”.

Em modelos generativos de IA, caso do ChatGPT, a “Alucinação” é um fenômeno comum, pois respostas devem ser geradas “a qualquer custo”, mesmo que não haja uma resposta correta ou suficientemente disponível em seu banco de dados. Deve haver ciência desse fenômeno para se buscar mitigá-lo. Como resultado das falhas, o erro pode levar o usuário a manter ou ampliar as informações falsas gerando desinformação. Uma das justificativas é o fato das IAs serem projetadas focando a humanização ao incorporar elementos criativos, ao invés da verdade rígida e da factualidade. Assim, a propensão à especulação, ao fornecer respostas, independentemente do nível de certeza associado, é

evidente.

Vale destacar que o ChatGPT não dispõe de capacidade de alterar diretamente seu banco de dados em relação à possibilidade de corrigir/alterar informações fornecidas de seu banco de dados. Os pesquisadores devem verificar a qualidade e a confiabilidade das fontes descobertas, estar familiarizados com as referências, checar a autoria e as diretrizes de citação e principalmente fazer julgamento crítico, ao usar informações obtidas por meio de IA, sabendo diferenciar e identificar os fatos, dados e alucinações trazidas pela máquina.

A confiança incondicional na precisão das informações proporcionadas por sistemas como o ChatGPT pode culminar na perpetuação de equívocos e na incorporação de dados não verificados em trabalhos acadêmicos, comprometendo, assim, a integridade da pesquisa e a confiabilidade das contribuições ao corpo do conhecimento científico. Há risco de implicações diretas oriundas da disseminação de informações inverídicas ou adaptadas.

Conforme verificado por Cortiz (2024), não há problema em usar IA como um assistente de escrita, mas é imprescindível utilizá-la de forma correta, entendendo as limitações da tecnologia. As IAs são um suporte confiável para a pesquisa e produção científica, conduzida de modo consciente em “não delegar toda a responsabilidade para o sistema (...), mas precisamos de muita literacia para fortalecer o uso crítico e evitar encurtar caminhos” (Cortiz, 2024).

Resulta como responsabilidade da comunidade acadêmica sensibilizar para questões: éticas na produção da ciência: “Ferramentas de IA não podem atender aos requisitos de autoria pois não podem assumir a responsabilidade pelo trabalho enviado”³ (COPE, 2023, tradução nossa). Autores devem descrever de forma transparente a utilização da IA. O tipo e uso da ferramenta deve constar na seção de Materiais e Métodos (ou similar), seja na elaboração do rascunho, do manuscrito, na coleta e análise de dados, gráficos etc. Os autores são totalmente responsáveis pelo conteúdo de seu artigo e pesquisa e por qualquer violação da ética de publicação (COPE, 2023, Spinak, 2023).

É consenso que para realizar qualquer tipo de alteração ou atualização as informações, os responsáveis pela criação da IA devem fazer treinamentos com dados recentes e pertinentes, de forma "manual", e, buscar dirimir as “alucinações” e excesso criativo, nos casos de ausência de retorno de resposta satisfatória para um prompt (OSUL, 2023). O combate às “Alucinações” e à desinformação advindas das ferramentas de IA orienta que “a chance de erro é reduzida em inteligências artificiais treinadas para assuntos específicos, pelo fato de estas serem alimentadas com bancos de dados mais restritos e estarem ligadas a um determinado contexto” (O GLOBO, 2023).

³ “AI tools cannot meet the requirements for authorship as they cannot take responsibility for the submitted work” (COPE, 2023). Disponível em: <https://publicationethics.org/cope-position-statements/ai-author> (2023).

4 CONSIDERAÇÕES FINAIS

Diante dos desafios inerentes à era digital, até o aperfeiçoamento da IA a níveis de excelência em confiabilidade, é imprescindível que todos os setores se empenhem na investigação e desenvolvimento de estratégias eficazes de mitigação da desinformação. Tal atitude é essencial para preservar a integridade informativa e a excelência científica, pilares fundamentais para uma sociedade informada. A propagação de informações incorretas exige uma supervisão rigorosa da qualidade e credibilidade das pesquisas, realizadas de maneira ética e responsável pela comunidade acadêmica. É de senso comum que a academia e seus pesquisadores são essenciais para aumentar a conscientização sobre a importância de difundir e combater informações confiáveis. Eles também ajudam a criar um ambiente intelectual que valoriza a ética acadêmica e a pesquisa de alta qualidade. A responsabilidade da comunidade acadêmica em combater a desinformação e as falsas notícias nesse contexto é crescente, o que enfatiza a importância de abordagens colaborativas e interdisciplinares para lidar com esse fenômeno complexo e multifacetado.

Considerando tal premissa, a desinformação no meio acadêmico tem o potencial de causar danos significativos, minar a fiabilidade da investigação legítima e ameaçar a reputação dos investigadores. Portanto, resulta como responsabilidade da comunidade acadêmica sensibilizar para estas questões: se no futuro o uso de algoritmos vier a fazer parte da rotina de pesquisa nas instituições universitárias, essa parceria talvez tenha de ser estendida à empresa responsável pela criação dos algoritmos usados na geração da primeira versão do trabalho. Mas, nesse caso, a pergunta que teremos que responder é: se os pesquisadores serão os verdadeiros “autores” dos trabalhos acadêmicos que geram, ou se seriam, antes “meta-autores” da pesquisa proposta. Essa é uma questão que terá que ser debatida nos próximos anos, e neste intento o artigo pretendeu colaborar e alertar durante a fase de transição até um maior aperfeiçoamento de tais sistemas tecnológicos.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001

REFERÊNCIAS

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; Naamán, 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](https://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_understanding_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-content/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