Pós-realidade e Teoria da Desinformação: inquietações sobre o uso massivo de IA Generativa

Simone Dias Marques[1]; Rita do Carmo Ferreira Laipelt[2]

> *"Uma realidade além da realidade, que, apreendida por todos no cotidiano, transforma tudo, do mais próximo ao mais distante, em uma noção de verdade vivida, mesmo que não diretamente"* (Jean Baudrillard).

RESUMO

A evolução dos Modelos Amplos de Linguagem (Large Language Models) de Inteligência Artificial tem tido um impacto significativo na interação entre humanos e tecnologias, especialmente no que diz respeito ao uso, recebimento, recuperação e compartilhamento de informações. Este artigo apresenta preocupações relacionadas ao potencial de desinformação e à distorção da realidade, ilustrados por meio de exemplos de uso do ChatGPT e outros modelos de IA generativa. O tema foi escolhido por sua relevância atual e por estar diretamente ligado ao objeto de estudo do meu projeto de pesquisa de mestrado em Ciência da Informação[3]. **OBJETIVO:** O objetivo deste artigo é discutir o potencial de desinformação e distorção da realidade que surgem a partir do uso do ChatGPT e de outros modelos de IA generativa. **METODOLOGIA:** Abordagem descritiva. O uso do ChatGPT e de outras Inteligências Artificiais Generativas é recente e ainda requer avaliação e pesquisa adequadas. **RESULTADO:** Este artigo destaca a importância da preservação de entradas de dados fornecidas por humanos e da possibilidade de recuperação de *inputs* diante do comportamento distorcido das memórias das IAs.

Palavras-chave: Inteligência Artificial. ChatGPT. Competência Crítica em Informação. Comportamento informacional. Desinformação.

---

1 Mestranda pelo Programa de Pós-graduação em Ciência da Informação (PPGCIN) da Universidade Federal do Rio Grande do Sul (UFRGS).

2 Professora do Programa de Pós-graduação em Ciência da Informação (PPGCIN) da Universidade Federal do Rio Grande do Sul (UFRGS).

3 *O impacto do modelo de Inteligência Artificial Generativa GPT-3 no comportamento da informação: implicações do ChatGPT para o conhecimento e a sociedade*. Projeto de Pós-Graduação em Ciência da Informação pela Universidade Federal do Rio Grande do Sul, sob orientação da professora Doutora Rita do Carmo Ferreira Laipelt.

Desinformação e degeneração em IAs

O ChatGPT, modelo de linguagem de conversação da OpenAI, foi lançado para uso público em 2023 e é um dos sistemas mais utilizados atualmente para diversas atividades. A partir do seu uso massivo, especialistas e pesquisadores passaram a relatar diversas inconsistências nas respostas da IA e sua possível manipulação para fins de desinformação como jamais antes seria possível na história humana.

A capacidade do modelo em inventar informações e narrativas falsas de forma credível e sem custo em escala jamais vista é uma das principais preocupações atuais entre a comunidade científica e entidades governamentais, como a Europol[4].

Propensos à alucinação[5], a dizer coisas que parecem plausíveis e em tom de autoridade, os sistemas de IA como ChatGPT e Bard também causam preocupações a respeito de seu uso por crianças, que não têm capacidade crítica para discernir o certo do errado.

As respostas deste tipo de IA já mostram que podem ser intencionadas a partir dos inputs. Como esses sistemas não contêm nenhum mecanismo para verificar a veracidade do que dizem, eles podem ser facilmente manipulados por humanos a partir de *prompts* para gerar desinformação e *fake news*.

Dentro desse contexto, os bots utilizados em campanhas políticas parecem inocentes. A desinformação e a propagação de *fake news* por meio de redes sociais têm sido utilizadas, sobretudo desde 2018, para moldar opiniões contra o Estado Democrático de Direito e suas instituições, incluindo modalidades de voto, a ciência, a educação e para fortalecer preconceitos, teorias da conspiração e minar a capacidade crítica de eleitores.

4 A Agência da União Europeia para a Cooperação Policial (Europol) é a agência de aplicação da lei da UE, cuja missão é ajudar a tornar a Europa mais segura. Com sede em Haia, nos Países Baixos, a missão da Europol é apoiar os seus Estados-Membros na prevenção e combate a todas as formas graves de criminalidade internacional e organizada, cibercrime e terrorismo. A Europol também trabalha com muitos países parceiros fora da UE e organizações internacionais.

5 Alucinação em IA refere-se à geração de resultados que podem parecer plausíveis, mas são factualmente incorretos ou não relacionados ao contexto dado. Esses *outputs* geralmente surgem de vieses inerentes ao modelo de IA, falta de compreensão do mundo real ou limitações de dados de treinamento.

Tal conjuntura cria o que Schneider (2022) definiu como Desinformação Digital em Rede (DDR), cuja finalidade atende a interesses de alienação social e vigilância de dados que têm consequências diretas na realidade, capturando estruturas sociais na esfera afetiva, cognitiva e moral. Para Schneider (2022), a solução para o rompimento das bolhas de alienação digital é a educação midiática.

> Para resistir e combater os efeitos da desinformação é preciso promover o fomento à educação midiática, particularmente [...] a competência crítica em informação, porque a raiz do problema não está na conexão, mas desconexão conectada, expropriada pela comunicação corporativa, pela ideologia neoliberal, pela vigilância e espionagem sorrateiramente articuladas da GAFAN (Google, Apple, Facebook, Amazon e Microsoft) e da NSA (National Security Agency) [..] que irriga as raízes da pós-verdade" (SCHNEIDER, 2022, p. 63).

Entretanto, ninguém podia imaginar que o lançamento do ChatGPT superaria qualquer outro tipo de sistema ou rede social em sua força para modificar o comportamento da informação, afetando a sociedade, o conhecimento e as relações entre capital e trabalho.

É possível dizer que a IA generativa é a invenção mais importante desde a internet, e diante disso se faz necessário e urgente abrir caminhos para uma *teoria da degeneração da informação e desinformação* para analisar implicações a um ambiente de *pós-realidade*.

Conforme mais pessoas usam IAs generativas para produzir e publicar textos, fotografias, filmes, etc., mais a web é inundada por materiais gerados artificialmente. Isso gera degeneração da informação. Esse fenômeno e suas consequências estão sendo estudados por pesquisadores do Reino Unido e do Canadá.

Schumailov e colaboradores (2023) analisaram o problema e descobriram que "aprender com os dados produzidos por outros modelos causa o colapso do modelo" — um processo degenerativo pelo qual, com o tempo, os modelos esquecem os dados originais com os quais aprenderam inicialmente e passam a gerar erros cumulativos, contaminando todo o conjunto de treinamento para os modelos subsequentes.

Ao longo do retreino em cima de seus próprios dados, a demência do modelo impacta na qualidade da informação e cria realidades paralelas sem base nem compromisso de verdade, deformando as referências humanas entre o que é real e o

que não é.

Por exemplo, há alguns meses, uma imagem do Papa[6] em trajes *fashion* circulou pela web e enganou diversos meios de comunicação, que a trataram inicialmente como verdadeira. Posteriormente, imagens de Donald Trump[7] preso por policiais também viralizaram. Pouco tempo depois, o Midjourney[8] anunciou o fim dos anos gratuitos de uso devido ao risco de desinformação causado pelo realismo das imagens.

Midjourney's software can be used to create misinformation, like fake images of Donald Trump being arrested. Image: Will Joel / The Verge

6 A imagem do Papa gerada pelo Midjourney foi notícia em diversos meios de comunicação. Apenas para referência, citaremos aqui a Folha de São Paulo: https://www1.folha.uol.com.br/mercado/2023/03/ia-que-cria-imagens-interrompe-teste-apos-montagens-de-trump-e-papa-viralizarem.shtml.

7 Para referência, citamos aqui o site The Verge: https://www.theverge.com/2023/3/30/23662940/deepfake-viral-ai-misinformation-midjourney-stops-free-trials.

8 Midjourney é um programa de IA generativa criado pelo laboratório de pesquisa independente Midjourney, Inc., de São Francisco, e gera imagens a partir de prompts em linguagem natural semelhantes ao DALL-E e Stable Diffusion da OpenAI.

Apesar de não terem qualquer base real nos fatos, são imagens que podem passar facilmente a impressão de realidade diante do requinte de verossimilhança.

Tanto é assim que apenas depois de seus autores as desmentirem é que os veículos de comunicação que as apresentaram como *fato* se retrataram. Já há inclusive um termo para classificar cenas fictícias e a fabricação de eventos históricos que jamais ocorreram: mídia sintética.

Competência crítica em informação na pós-realidade

Sem padrões, sem regulamentação: o processamento de IA costuma ser uma caixa preta, pois não sabemos determinar qual lógica exatamente o modelo aprendeu para processar as informações, nem de onde vêm ou para onde vão os dados que cada um deles extrai de usuários.

Partindo das constatações de Schumailov e colaboradores (2023), pode-se assim propor a possibilidade de surgimento de uma *pós-realidade*, um ecossistema em que distinguir entre o que é informação gerada por LLMs e o que é real se torna quase impossível, embora tais imagens e informações não passem de um pastiche de dados extraídos pelas IAs a partir de referências em bancos de dados da web e de si mesmas: um *simulacro*, no sentido que lhe deu Baudrillard (1981).

A *pós-realidade* desinformacional refere-se, assim, ao atual contexto em que a quantidade de informações disponíveis é imensa, mas a qualidade e a veracidade dessas informações podem ser questionáveis. A disseminação de notícias falsas, desinformação e a manipulação da informação tornaram-se desafios significativos para os indivíduos, a sociedade e até mesmo para as próprias inteligências artificiais.

Para que o aprendizado de máquina possa ser saudável por um longo período de tempo é preciso garantir que o acesso às fontes de **dados originais humanos** seja preservado e que os dados adicionais, não gerados pelos LLMs, permaneçam disponíveis. A necessidade de distinguir dados gerados por LLMs de outros dados levanta questões sobre a proveniência do conteúdo rastreado da Internet: não está claro como o conteúdo gerado por LLMs pode ser rastreado em escala.

Em resumo, na definição de Schumailov (2023), é como se o modelo de IA ficasse com demência senil[9] de forma irreversível e isso contaminasse todos os seus descendentes, gerando um impacto ainda desconhecido no comportamento da informação e na sua recuperação e apagando as fontes originais, em geral humanas.

É fácil induzir o ChatGPT a criar desinformação e até mesmo relatar estudos sobre uma ampla gama de tópicos, da medicina à política e à religião, inventar estudos, estatísticas e links que não levam a nenhum lugar. Além disso, esses bots custam quase nada para operar e, portanto, reduzem o custo de geração de desinformação a zero. É possível a alguém obter seu próprio LLM treinado e personalizado por US $ 40.000, o preço de um aplicativo simples.

Infelizmente, a pasta de dente não pode mais ser colocada para dentro. Ou, como dizem os entusiastas da IA, “foguete não tem ré” e tudo aponta para um cenário em que a desinformação automatizada em grande escala veio para ficar e afetar grande parte nossas interações sociais, comerciais, financeiras, de emprego, educacionais e governamentais” (Fernández Marcial & Esteve Gomes, 2022, *apud* Berryhill, et al., 2021).

CONCLUSÃO

Para a Ciência da Informação, é fundamental reconhecer a importância da preservação dos *inputs* humanos para a recuperação da informação, bem como para a prevenção da alucinação ou degeneração de inteligências artificiais. Essa abordagem ganha relevância com base no estudo de Schumailov (2023), que destaca a necessidade de adotar uma competência crítica em informação para enfrentar a *pós-realidade* desinformacional.

As IAs são treinadas com grandes volumes de dados, mas sua capacidade de processar informações é apenas tão boa quanto os dados de entrada que recebem. Se

9 MARQUES, S. D. *Looping de respostas distorcidas: o “colapso do modelo” de modelos de IA* generativa. Disponível em: https://medium.com/@simonemarques_38909/looping-de-respostas-distorcidas-o-colapso-do-modelo-de-modelos-de-ia-generativa-2ca080267363

esses *inputs* forem contaminados com desinformação, viés ou manipulação, a IA pode reproduzir esses vieses e disseminar informações incorretas. Assim, a preservação de *inputs* humanos é uma salvaguarda contra o comportamento degenerativo das memórias das inteligências artificiais.

Dessa forma, a preservação de inputs humanos de alta qualidade, provenientes de fontes confiáveis e verificadas, torna-se uma estratégia crucial para a recuperação da informação confiável e precisa. Ao incorporar a perspectiva humana na análise e na validação dos dados, é possível melhorar a capacidade das IAs em discernir informações verdadeiras e filtrar aquelas que são falsas ou enganosas. Essa abordagem conjunta entre seres humanos e IA pode levar a um ecossistema informacional mais saudável e confiável.

ABSTRACT

The evolution of Artificial Intelligence Large Language Models (LLMs) has had a significant impact on the interaction between humans and technologies, especially with regard to the use, receipt, retrieval, and sharing of information. This article presents concerns related to the potential for misinformation and reality distortion, illustrated through examples of using ChatGPT and other generative AI models. The topic was chosen because of its current relevance and because it is directly linked to the object of study of my Master's research project in Information Science.

Keywords: Artificial Intelligence. ChatGPT. Critical Competence in Information. Information behavior. Misinformation.

REFERÊNCIAS

BAUDRILLARD, J. **Simulacres et simulation**, Paris: Galilée, 1981.

BRISOLA, A. *Competência Crítica em Informação como Resistência à Sociedade da Desinformação sob um Olhar Freiriano: Diagnósticos, epistemologia e caminhos ante as distopias informacionais contemporâneas*. (Tese). Instituto Brasileiro de Informação em Ciência e Tecnologia, UFRJ (2021).

BEZERRA, A. & Schneider, M. (Orgs.) **Competência crítica em informação: Teoria, consciência e práxis**. Disponível em: <https://ridi.ibict.br/handle/123456789/1200> (2022).

CAPURRO, R. **Epistemología y ciencia de la información**. Enlace, Maracaibo, v. 4, n. 1,p. 1129, 2007. Disponível em: http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S1690751520070001000002&lng=es&nrm=i.

CHOMSKY, N. *The False Promise of ChatGPT*. New York Times, Março de 2023. Disponível em:<https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>. Acesso: Maio de 2023.

FERNÁNDEZ, V. & Esteves Gomes, L.I. *Impacto de la Inteligencia Artificial en el comportamiento informacional: elementos para el debate*. Bibliotecas. Anales de Investigación; 18(3), 1-12. (2022). Disponível em: <http://revistas.bnjm.cu/index.php/BAI/article/view/524/503>. Junho, 2023.

GOODFELLOW, I., Bengio, Y., & Courville, A. **Deep Learning**. MIT Press (2016).

GÓMEZ, M. N. G. *Para uma reflexão epistemológica acerca da Ciência da Informação*. Perspectivas em Ciência da Informação, v. 6, n. 1, 2001. Disponível em: http://hdl.handle.net/20.500.11959/brapci/37093.

HAO, Karen. *This is how AI bias really happens and why it's so hard to fix*. MIT Review, 2019. Disponível em https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/.

HSU, Tiffany & Thompson, Stuart A. *Disinformation Researchers Raise Alarms About A.I. Chatbots: Researchers used ChatGPT to produce clean, convincing text that repeated conspiracy theories and misleading narratives*. The New York Times. Disponível em: <https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html>.
Acesso em Junho de 2023.

MARCUS, G. *AI Platforms like ChatGPT Are Easy to Use but Also Potentially Dangerous*. Scientific American, Dezembro de 2022. Disponível em: <https://www.scientificamerican.com/article/ai-platforms-like-chatgpt-are-easy-to-use-but-also-potentially-dangerous/>, Junho de 2023.

MARQUES, M. B.; Gomes, L. E. *Ciência da Informação: visões e tendências*. Imprensa da Universidade de Coimbra/Coimbra University Press, 2020. Disponível em: https://www.researchgate.net/publication/343830444_Ciencia_da_Informacao_visoes_e_tendencia.

MIHAIL, E. *A Complete Introduction to Prompt Engineering For Large Language Models*. Feb., 2023. Disponível em: https://www.mihaileric.com/posts/a-complete-introduction-to-prompt-engineering/.

PORTO, L. S. *Uma investigação filosófica sobre a Inteligência Artificial*. Informática na

Educação: teoria & prática. Porto Alegre, v.8, n.2, p.11-26, jan./jun.2006. Disponível em: https://seer.ufrgs.br/index.php/InfEducTeoriaPratica/article/view/2304/1005 (14/06/2023).

ROBERTSON, J. ,*Countering Disinformation in a Post-ChatGPT World*. Disponível em: <https://www.boozallen.com/insights/cyber/tech/how-to-counter-disinformation-in-a-post-chatgpt-world.html>, Junho, 2023.

SEARLE, John R. **The Construction of Social Reality**, New York Free Press, 1995.

SCHNEIDER, Marco. **A era da desinformação. Pós-verdade, fake news e outras armadilhas**. Garamond, Rio de Janeiro, 2022.
SCHUMAILOV, I. *et al*. *Model Dementia: Generated Data Makes Models Forget*. In: https://www.researchgate.net/publication/371137235_Model_Dementia_Generated_Data_Makes_Models_Forget>. Maio de 2023.

TRONCO, Giordano B. *ChatGPT impacta rotinas na pesquisa e na educação e levanta questionamentos sobre veracidade e metodologias de avaliação*. Jornal da Universidade/ Universidade Federal do Rio Grande do Sul, Abril de 2023. Disponível em: <https://www.ufrgs.br/jornal/chatgpt-impacta-rotinas-na-pesquisa-e-na-educacao-e-levanta-questionamentos-sobre-veracidade-e-metodologias-de-avaliacao/>. Acesso em Maio de 2023.