Turkish Journal of Psychiatry 2024;35(4):333–335

Türk Psikiyatri Dergisi

Turkish Journal of Psychiatry

Letter to the Editor

Is Artifical Intelligence Hallucinating?

Dear Editor,

Artificial intelligence (AI) is rapidly advancing to become an indispensable part of our daily lives. Particularly, the development and accessibility of generative AI systems using large language models (LLMs) such as ChatGPT have accelerated transformations in every field. Generative AI concerns all professions and eliminates age range distinctions. Consequently, a comprehensive AI ecosystem is currently forming. However, awareness of the issues it causes or may cause in the long run is not developing at the same pace as its benefits. Therefore, being aware of the risks as well as the benefits will provide insight into what kind of ecosystem will be formed (Ayhan, 2023; Perc et al. 2019).

There are roughly three main risks associated with generative AI. The most well-known of these is data privacy. AI systems, which are fed with data, can violate data protection and privacy if precautions are not taken, leaving individuals vulnerable to various manipulations. Secondly, these systems have a high risk of producing biased results (Ozer et al. 2024a). There are two sources of biased results: biases in the training data set and biased assumptions in the algorithms. Since these systems learn based on big data, they maintain biases related to different contexts such as religion, culture, race, gender, or socioeconomic status present in big data, and can thus reproduce these biases in their outputs. The second source of bias is the assumptions made in the algorithms of these systems. The bias in algorithmic assumptions naturally ensures that the results are also biased. In both cases, AI has the potential to deepen existing inequalities by spreading inequalities more quickly. Finally, not all information produced by these systems is accurate; AI systems often exhibit hallucinations, producing incorrect content. This article will briefly evaluate the hallucinations observed in the behavior of AI systems.

AI hallucination is a phenomenon where AI generates a convincing, contextually coherent but entirely fabricated response that is independent of the user's input or previous context. Therefore, although the responses generated by generative AI may seem plausible, they can be meaningless or incorrect. For example, in a study aiming to evaluate the frequency of AI hallucinations in research proposals entirely prepared by ChatGPT, it was shown that out of 178 references generated by ChatGPT, 69 were not Digital Object Identifiers (DOIs) and 28 references did not appear in Google searches or have an existing DOI (Athaluri et al. 2023). Generally, while generating content based on the training data set, generative AI can produce content during the hallucination process by disconnecting from the source of the training data set (source amnesia) (Berberette et al. 2024). Furthermore, to maintain consistency with the incorrect results it generates, generative AI can continue to produce incorrect content sequentially once it has produced incorrect content, a behavior known as the snowball effect of hallucination (Zhang and Press et al. 2023).

Hallucinations are classified into intrinsic and extrinsic categories (Ji et al. 2023). According to this classification, intrinsic hallucinations result in outputs that contradict the source content and the conversation history, while extrinsic hallucinations correspond to outputs whose accuracy cannot be verified based on the source content or conversation history. Therefore, outputs based on intrinsic hallucinations are directly related to the misinterpretation of information and are directly incorrect, while extrinsic hallucinations are

fictional content added to the generated text. On the other hand, Zhang et al. (2023) evaluate hallucinations in three different categories: input-conflicting (content deviating from the source input provided by the user), context-conflicting (content conflicting with the previously generated content by the generative AI), and reality-conflicting hallucinations. If both classifications are combined, hallucinations conflicting with the input and context correspond to intrinsic hallucinations, while hallucinations conflicting with reality correspond to extrinsic hallucinations.

Hallucinations observed in generative AI are more related to the probabilistic nature of the model and its direct relationship with the training data set than to any real cognitive process. Patterns, expressions, or concepts that are more frequently included in the training data set can trigger hallucinations during the response generation process of generative AI due to their easy and quick accessibility, regardless of the context's accuracy. Additionally, conflicting information within the large training data set can cause intrinsic tensions in the generative AI's response generation process, thereby triggering hallucinations (Berberette et al. 2024). Similarly, outdated, incomplete, or false information in the data sets can also lead to hallucinations (Zhang et al. 2023). Therefore, there is a strong relationship between generative AI hallucinations and the distribution of the training data (McKenna et al. 2023). On the other hand, hallucinations can also arise due to the weaknesses of various language models used in generative AI (Athaluri et al. 2023). In summary, the large, unstructured text corpus used to train generative AIs, the stochastic behavior caused by the probabilistic nature of the model, and the weaknesses of language models can lead to hallucinations.

On the other hand, considering that this behavior of generative AI is mostly related to incomplete, inconsistent, or incorrect information in the data sets used for AI system learning, it also shows how susceptible real-life data is to hallucinations. In fact, it is suggested that everyday memory reconstruction often involves a degree of confabulation (French et al. 2009), and that it is common for healthy individuals to unintentionally make details of stories fictional (Riesthuis et al. 2023). In other words, people often resort to fictional but unreal information to fill gaps in their narratives without malicious intent. Similarly, it is observed that the fabricated content produced by generative AI appears more frequently in texts that exhibit narrativity (Sui et al. 2024).

Additionally, the difficulty in distinguishing between inspiration and imitation during the content generation process based on large training sets by generative AI (GAI) also stands out as a separate issue (Carobene et al. 2024). The demonstration of a closer relationship among articles with LLM modifications indicates that GAI reduces text diversity (Liang et al. 2024). In this context, recent studies highlight the need to evaluate the problems caused by hallucinations according to different fields, suggesting that hallucinations may be beneficial in some areas. For example, it has been empirically shown that hallucinations exhibit significantly higher narrativity than fact-based outputs (Sui et al. 2024). Areas where hallucinations can be particularly valuable include providing inspiration for creative writing (Mukherjee and Chang, 2023), architectural design (Hegazy and Saleh, 2023), discovering new proteins (Anishchenko et al. 2021), and formulating innovative legal analogies (Dahl et al. 2024) (Sui et al. 2024). Therefore, beyond strengthening the language models used in this context, while efforts are made to reduce the frequency or scale of hallucinations by filtering out incorrect information from the learning data, it is important to consider what constitutes incomplete or inconsistent information and whether reducing hallucinations might negatively impact the diversity of the generated information.

On the other hand, clinically, hallucinations—defined as sensory experiences not associated with an external stimulus—are often linked to conditions such as schizophrenia, bipolar disorder, and Parkinson's disease (Tamminga, 2009). Using the term "hallucination" to describe the fictitious outputs produced by LLMs similarly implies that LLMs are consciously engaging in a sensory input perception process (Smith et al. 2023). However, generative AI lacks consciousness, and thus is devoid of subjective experience or awareness (Berberette et al. 2024). On the other hand, unlike human hallucinations that are not associated with an external stimulus, in AI systems the data on which LLMs are trained and the prompts that lead to this behavior are considered external stimuli (Østergaard and Nielbo, 2023). Therefore, since generative AI models do not see something that isn't there but rather fabricate it, the term "hallucination" may not be appropriate to describe this situation. Instead, the term "confabulation"—a psychiatric concept referring to the creation of narrative details that are believed to be true despite being false—is suggested (Smith et al. 2023). Recent studies indicate that much of the information produced by generative AI can be considered within the context of confabulation (Berberette et al. 2024; Sui et al. 2024).

Finally, understanding this behavior as confabulation also carries within it the solution to mitigating the problem. Clinically, confabulation is associated with right hemisphere deficiency in the brain, where the left hemisphere creates fabricated or distorted memories influenced by existing knowledge, experience, or contextual cues (Smith et al. 2023). Therefore, since the situation in generative AI corresponds to the incorrect reconstruction of information influenced by existing knowledge and context, the concept of confabulation better defines the current situation from a technological background perspective. This concept actually provides significant evidence that choosing a path that complements humans rather than automation will create a more humane ecosystem in the AI landscape. As previously stated, the human-complementing path is a new approach that focuses on complementing human deficiencies while considering

the impacts of automation on efficiency, but also protecting employment (Capraro et al. 2023; Ilikhan et al. 2024; Ozer et al. 2024b; Ozer and Perc, 2024). Thus, confabulation, which occurs when the right hemisphere of the brain is damaged and cannot perform its real function, corresponds to the working method of AI that lacks the contributions of the right hemisphere of the brain. In other words, the production of incorrect or nonexistent information by generative AI overlaps with the similar behavior of the left hemisphere when the right hemisphere is damaged. Therefore, metaphorically, having humans step in to correct these deficiencies and complete the text could be likened to donating the human right hemisphere to AI, supporting the correction of such errors and complementing generative AI.

In conclusion, generative AI systems can produce incorrect information due to various reasons. Therefore, it is evident that not all information generated by generative AI will be accurate, and incorrect information produced without malicious intent can always exist alongside correct information. Whether this behavior is labeled as hallucination or confabulation, the severity of complications it causes varies depending on the field of application. For instance, in the healthcare sector, tthe hallucination or confabulation experiencing of generative AI could lead to critically dangerous outcomes in diagnosis and treatment processes (Ilikhan et al. 2024). Similarly, in the education system, such behavior could have profound negative effects on learning processes (Ozer, 2024). However, as noted in the literature, in creative and innovative fields, this behavior may prove beneficial. Therefore, contextual awareness of this behavior is crucial in mitigating its potential negative impacts while harnessing its potential benefits in appropriate contexts.

REFERENCES

Anishchenko I, Pellock SJ, Chidyausiku TM et al. (2021) De novo protein design by deep network hallucination. Nature 600: 547–52.

Athaluri S, Manthena S, Kesapragada V et al. (2023) Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. Cureus 15: e37432.

Ayhan Y (2023) The Impact of Artificial Intelligence on Psychiatry: Benefits and Concerns-An essay from a disputed 'author'. Turkish Journal of Psychiatry 34: 65–7.

Berberette E, Hutchins J, Sadovnik A (2024) Redefining "Hallucination" in LLMs: Towards a psychology-informed framework for mitigating misinformation. arXiv:2402.01769v1.

Carobene A, Padoan A, Cabitza F et al. (2024) Rising adoption of artificial intelligence in scientific publishing: Evaluating the role, risks, and ethical implications in paper drafting and review process. Clin Chem Lab Med 62: 835–43.

Capraro V, Lentsch A, Acemoğlu D et al. (2023) The impact of generative artificial intelligence on socioeconomic inequalities and policy making. PNAS Nexus 3: 1-18.

Dahl M, Magesh V, Suzgun M et al. (2024) Large legal fictions: Profiling legal hallucinations in large language models. arXiv:2401.01301.

French L, Garry M, Loftus E (2009) False memories: A kind of confabulation in non-clinical. Confabulation: Views from neuroscience, psychiatry, psychology, and philosophy, William Hirstein (ed.): 33-66.

Hegazy M, Saleh AM (2023) Evolution of AI role in architectural design: between parametric exploration and machine hallucination. MSA Engineering Journal 2: 1-26.

Ilikhan S, Özer M, Tanberkan H et al. (2024) How to mitigate the risks of deployment of artificial intelligence in medicine? Turk J Med Sci 54: 483-92.

Ji Z, Lee N, Frieske R et al. (2023) Survey of hallucination in natural language generation. ACM Computing Survey 55: 1–38.

Liang W, Zhang Y, Wu Z et al. (2024) Mapping the increasing use of LLMs in scientific papers. arXiv preprint arXiv:2404.01268v1.

McKenna N, Li T, Cheng L et al. (2023) Sources of hallucination by large language models on inference tasks. arXiv preprint arXiv:2305.14552.

Mukherjee A, Chang H (2023) The creative frontier of generative ai: Managing the novelty-usefulness tradeoff. arXiv:2306.03601.

Østergaard SD, Nielbo KL (2023) False Responses From Artificial Intelligence Models Are Not Hallucinations. Schizophr Bull 49: 1105–7.

Ozer M (2024) Potantial benefits and risks of artificial intelligence in education. Bartın University Journal of Faculty of Education 13: 232-44.

Ozer M, Perc M (2024) Human complementation must aid automation to mitigate unemployment effects due to AI technologies in the labor market. Reflektif Journal of Social Sciences 5: 503-14.

Ozer M, Perc M, Suna HE (2024a) Artificial intelligence bias and the amplification of inequalities in the labor market. Journal of Economy, Culture and Society 69: 159-68.

Ozer M, Perc M, Suna HE (2024b) Participatory management can help AI ethics adhere to the social consensus. Istanbul University Journal of Sociology 44: 221-38.

Perc M, Özer M, Hojnik J (2019) Social and juristic challenges of artificial intelligence. Palgrave Communications 5, 61. https://doi.org/10.1057/s41599-019-0278-x

Riesthuis P, Otgaar H, Bogaard G et al. (2023) Factors affecting the forced confabulation effect: a meta-analysis of laboratory studies. Memory 31:635–51.

Smith AL, Greaves F, Panch T (2023) Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models. PLOS Digit Health 2: e0000388.

Sui P, Duede E, Wu S et al. (2024) Confabulation: The Surprising Value of Large Language Model Hallucinations. arxiV:2406.04175v2.

Tamminga CA. Schizophrenia and other psychotic disorders: Introduction and overview. In: Sadock BJ, Sadock VA, Ruiz P, eds. Kaplan and Sadock's Comprehensive Textbook of Psychiatry. 9th edition. Philadelphia: Lippincott Williams and Wilkins; 2009. p. 1432.

Zhang, Y, Li Y, Cui L et al. (2023). Siren's song in the AI ocean: A survey on hallucination in large language models. arXiv preprint arXiv:2309.01219.

Zhang M, Press O, Merrill W et al. (2023) How language model hallucinations can snowball. arXiv preprint arXiv:2305.13534.

Mahmut ÖZER

Received: 15.09.2024, **Accepted:** 19.09.2024, **Available Online Date:** 14.10.2024

National Education, Culture, Youth and Sports Commission, Grand National Assembly of Turkey.

Engineer, Mahmut Özer, e-mail: mahmutozer2002@yahoo.com

https://doi.org/10.5080/u27587