

Neural Interaction Between Learners and Artificial Intelligence Through the Lens of Fallacies

Farhad Shafiepour Motlagh¹  and Sabina Mirzaei Nobari² 

¹Department of Educational Administration, Maha. C, Islamic Azad University, Mahallat, Iran

²Department of Information Technology, Western Caspian University, Baku, Azerbaijan

Received: 01.09.2026

Accepted: 28.09.2026

Published: 09.10.2026

<https://doi.org/10.54414/DPII.2371>

Copyright: © 2026 by the authors.
Licensee: Journal of Smart Technologies and Computational Systems, Western Caspian University, Baku, Azerbaijan. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 International License (CC BY 4.0).

Abstract: The rapid expansion of large language models and generative artificial intelligence systems in educational ecosystems has moved learner-machine interaction beyond a merely instrumental interface toward cognitive entanglement and neural stimulation. Adopting a phenomenological and critical approach, this qualitative study explores learners' neural-cognitive engagement with fabricated content arising from "AI hallucination" and its implications for the reproduction of epistemological fallacies. The research involved in-depth semi-structured interviews with higher education learners and participant observation of AI-assisted problem-solving processes. The resulting data were interpreted through inductive thematic analysis. The findings indicate that, when confronted with algorithmic error, learners' cognitive processing develops an "artificial confidence bias" under the influence of AI's coherent linguistic structure and persuasive tone. By providing rapid access to seemingly complete answers, this process satisfies the neural dopaminergic reward circuit and consequently suspends cognitive inhibition and critical reflection. The extracted themes revealed that learners' neural interaction with generative AI hallucinations, through six fallacies—formal eloquence, artificial authority, immediate relief, part-to-whole generalization, anthropomorphic analogy, and confirmatory conformity—transforms human agency into cognitive passivity and erodes critical judgment through artificial confidence bias. This epistemic erosion reduces truth to statistical rhetoric and can be mitigated and remedied only through a transition from reward-driven consumption to reflective pedagogy and relational ethics.

Keywords: Learners' neural interaction; artificial intelligence; hallucination-related fallacies

1. Introduction

The structural transformations brought about by the emergence and accelerating penetration of generative AI and large language models into higher education ecosystems have profoundly altered academic learning and research. "Learners' neural interaction with artificial intelligence based on hallucination-related fallacies" denotes a complex, interdisciplinary condition at the intersection of cognitive neuroscience, critical epistemology, and pedagogy.

Learners' neural interaction with artificial intelligence based on hallucination-related fallacies refers to "a dynamic, bidirectional process in which structural errors and imaginary data reproduced by large language

models (AI hallucination), through coherent linguistic structure, eloquent expression, and rapid responsiveness, manipulate information-processing mechanisms in the learner's brain. By stimulating dopaminergic reward circuits and producing cognitive ease, this process suspends analytical thinking and cognitive inhibition in the prefrontal cortex, leading learners to accept false propositions unconsciously through six reasoning traps (fallacies)."

Student-machine interaction no longer appears merely as a neutral user interface or computational tool, but as a form of neural-cognitive entanglement and epistemic intermingling. One fundamental and unresolved challenge in using these technologies is AI hallucination: a condition in which a system reproduces fabricated or invalid propositions, imaginary sources, or fictitious data with flawless textual coherence, fluent language, and a thoroughly persuasive tone [1], [2]. The central challenge arises when this technical characteristic intersects with learners' biological structures and neural decision-making circuits.

When encountering AI-generated text, the learner's brain enters a state of cognitive ease because of conceptual fluency and syntactic refinement. According to dual-process theory, this process suspends prefrontal activity and System 2 inhibitory mechanisms, shifting the processing burden to System [3]. Combined with exceptionally rapid responses, this ease of processing stimulates the neural dopaminergic reward circuit by immediately alleviating the anxiety of not knowing or having insufficient time [4]. This technological-neural overlap produces an "artificial confidence bias": an error that leads the knowing subject to equate the machine's statistical hallucination with genuine understanding and to slip into analytical passivity.

This process draws learners into a chain of epistemological and neural fallacies: formal eloquence in place of substantive truth, appeal to artificial algorithmic authority, false neural causation based on immediate relief, generalization from a fabricated part to the truth of the whole, Anthropomorphic analogy applied to mathematical mechanisms, and confirmatory conformity within algorithmic sycophancy [5], [6]. Instead of engaging with the constructive difficulty of scientific discovery and tolerating unresolved questions in research, learners become trapped in a cycle of hurried reward-seeking that blurs the boundary between authentic empirical reasoning and algorithmic imagination. The central problem of this study is therefore to explain how the technical defect of hallucination becomes a syndrome of cognitive erosion through the reproduction of mental fallacies in academic life.

The importance of this issue emerges from three interconnected domains: epistemology, pedagogy, and neuroethics. In the first, higher education fundamentally rests on the pursuit of truth, critical reflexivity, and the authenticity of evidence. If interaction with language models turns eloquence into a criterion of correctness, truth is reduced from correspondence with reality or internal coherence to "syntactic beauty and the speed of text generation." This destabilizes rational scrutiny and increases the risk of disseminating structured pseudoscience in research.

Pedagogically, deep learning is inherently connected to cognitive friction, methodological doubt, and patience in the face of ambiguity. Conditioning learners' neural structures to immediate dopaminergic gratification erodes mental resilience in confronting the difficulties of authentic research and turns learners into passive operators who merely consume prefabricated content.

From the perspective of neuroethics and organizational ethics, delegating scientific judgment to opaque black-box systems not only undermines students' ethical responsibility for their claims but also transforms authentic human dialogue in education into a false, monologic, and hegemonic relationship with a tool [7], [8]. Understanding the mechanisms of this intellectual captivity is therefore a prerequisite for context-sensitive policymaking, professional ethics, and curriculum redesign in the age of AI.-

The literature indicates that most research on AI hallucination has focused on technical dimensions, prompt engineering, and statistical parameter adjustment [2]. Studies of its educational consequences, meanwhile, have generally remained within quantitative approaches, superficial behavioral measurement, or optimistic technophilic attitude surveys [9]. Three fundamental gaps can be identified.

First, earlier research has largely treated AI as a learning aid and neglected its disruptive role in learners' psychological–neural economy. The relationship between brain reward circuits (the dopaminergic pathway) and the suspension of scrutiny when encountering algorithmic error has rarely been formulated in educational studies.

Second, machine hallucination is usually treated as an isolated erroneous output rather than a generative context for reproducing patterns of human mental deception. A structured qualitative account of interaction-generated fallacies, such as susceptibility to eloquence, false neural causation, and algorithmic sycophancy, has been missing from higher education theory.

Third, students' lived experiences when encountering hallucinated data, their movement toward false confidence, and their transformation from knowing subjects into passive agents require qualitative investigation immersed in learners' lifeworlds. These dimensions are not usually captured by conventional questionnaires.

To address these gaps, this study offers specific epistemic and methodological distinctions from comparable research. Moving beyond one-dimensional engineering studies or superficial educational descriptions, it analyzes hallucination at the intersection of neural brain responses, reasoning and fallacy, and epistemic power structures, providing a multilayered examination of the learner's mind.

The study avoids uncritical praise of algorithmic efficiency. By questioning the reduction of the human subject to an optimized consumer, it emphasizes the restoration of critical reflexivity and a clear distinction between machine-generated statistical rhetoric and authentic, truth-oriented human knowledge.

A distinctive strength of this research is its thematic abstraction and formulation of six coherent types of fallacy from empirical data. This structures the invisible connection between the technical architecture of language models and human cognitive traps into a practical analytical matrix.

Moving beyond small-scale quantitative data, the research explores learners' perceptual experiences in direct encounters with the system. It inductively documents the psychological–neural mechanisms leading to artificial confidence bias from within academic lifeworlds, opening possibilities for higher education policymakers to develop critical pedagogy and redesign learning assessment.

2. Literature Review

A review of international research published between 2020 and 2026 indicates that educational technology and cognitive neuroscience research has rapidly shifted from the instrumental usefulness of AI toward the neural–cognitive pathology of learner interaction with large language models and the mechanisms of fallacious acceptance of distorted outputs. Early theoretical and empirical foundations emerged with Bender et al.'s (2021) warning about “stochastic parrots.” They explained that large language models connect texts based on statistical distributions, while their deceptive eloquence traps learners in the fallacy of formal eloquence and encourages the projection of awareness and conceptual understanding onto the system. Extending this argument, laboratory studies in cognitive neuroscience showed that textual fluency breaks down cognitive resistance in the prefrontal cortex. Drawing on resource-rationality theory, Lieder et al. (2020) demonstrated that the human brain continually seeks to reduce metabolic energy consumption. By facilitating sensory–cognitive perception, fluent AI output creates a “correct answer” signal in the evaluation circuit, suspending System 2 (effortful analytical thinking) and producing unquestioning trust even in fabricated propositions.

As tools such as ChatGPT became widespread in academic settings, Casper et al. (2023) investigated submission to the fallacy of appeal to artificial authority. Their findings showed that school and university students attribute excessive omniscience to algorithms because of the alignment bias generated by bots' assertive, unhesitating tone, internalizing unreal or hallucinated data without validation. Di Mario et al. (2024) investigated the neurobiological dimensions using electroencephalography (EEG) and functional

magnetic resonance imaging (fMRI). Their findings indicated that rapid, integrated AI responses significantly reduce activity in the executive control network, including the dorsolateral prefrontal cortex (DLPFC) and anterior cingulate cortex, while stimulating striatal dopaminergic circuits through rapid, effortless cognitive rewards. This process underlies the fallacy of false neural causation, in which early dopamine release and relief from problem-solving-related academic anxiety are mistakenly interpreted as genuine learning and deep understanding.

Furthermore, Bostrom and Shulman (2024) and Greene and Foster (2025) examined students' interpretive and analogical mechanisms empirically. They found that error-free machine performance in a limited domain is rapidly extended to all scientific domains through the part-to-whole generalization fallacy, as learners equate limited text-processing analytical capacity with comprehensive knowledge. Bynum et al. (2025) also showed, through analysis of interactive exchanges, that chatbots' first-person and ostensibly polite responses encourage the projection of mental states and activate the anthropomorphic analogy fallacy within learners' social cognition. The adolescent brain consequently encodes the computational algorithm as an "aware and caring peer" and responds to its hallucinations with justification, accommodation, and tolerance.

In the most advanced recent studies, Peterson, Duke, and Steinberg (2026) examined the erosion of metacognitive monitoring and the intensification of confirmatory conformity in a longitudinal neuroeducational trial. They observed that learners confronted with obvious system errors and hallucinations adjusted their thinking to AI's mistakes instead of critically reconsidering them. To reduce cognitive dissonance, they equated machine errors with inadvertent human mistakes through the "you too" fallacy. Collectively, these studies suggest that uncritical reliance in learner–AI neural interaction creates a cognitive–dopaminergic trap in which machine-generated statistical hallucinations suppress critical thinking through structured fallacies and erode learners' mental agency.

3. Research Questions

1. What are the dimensions and components of an analysis of learners' neural interaction with artificial intelligence through the lens of fallacies?
2. What framework can be developed to analyze learners' neural interaction with artificial intelligence through the lens of fallacies?

4. Research Methodology

This study adopted a qualitative interpretive–critical paradigm and a descriptive–interpretive phenomenological strategy (hermeneutic phenomenology). Learners' encounters with fabricated and hallucinated AI outputs are subjective, contextual, and intertwined with neural–cognitive and emotional structures. Phenomenology therefore enables the identification and deconstruction of the underlying, unarticulated layers of learners' lifeworlds and the meanings they assign to interaction with algorithms and to justifications of algorithmic errors. The critical perspective allowed the researcher to move beyond description and interrogate asymmetrical epistemic power relations, the prevalence of cognitive fallacies, and prefrontal passivity under the dominance of technological formal eloquence.

5. Research Setting and Data Collection

The research setting comprised postgraduate students at comprehensive public universities and Islamic Azad University who regularly used generative large language models each week for research and problem-solving. Criterion-based purposive sampling sought maximum experiential variation across the humanities, basic sciences, and engineering. Sample size was determined by theoretical saturation: participants were recruited until the final interviews yielded no new open codes or conceptual themes and information became

repetitive. Saturation was reached with the nineteenth participant; recruitment continued to twenty-one participants to ensure completeness.

Data collection included in-depth semi-structured interviews addressing experiences of AI errors, dopaminergic feelings associated with eloquent responses, review mechanisms, forms of projection, and justifications of fallacies. Interview transcripts and observation protocols were analyzed through inductive thematic analysis inspired by Braun and Clarke's multistage model. First, the data were read word by word to develop an overall understanding. Initial codes (basic themes) were then extracted from explicit statements and their neural–cognitive implications. Next, codes were grouped into broader concepts to form organizing themes. Finally, abstraction and conceptual synthesis generated global themes explaining the architecture of learners' projections and fallacies. To establish research quality and scientific trustworthiness, the study used Guba and Lincoln's (1985) qualitative trustworthiness framework and its four components rather than traditional quantitative validity and reliability concepts.

6. Discussion and Conclusion

Analysis of the qualitative data reveals that learners' interaction with generative AI extends beyond technological use to a transformation in cognitive economy and neural decision-making circuits. Encounters with fabricated outputs, or AI hallucination, create an asymmetrical connection between human cognitive psychology and machine-generated statistical probabilities. The following discussion examines each global theme identified in the six analytical tables in relation to relevant theoretical foundations and previous research, identifying points of convergence and divergence.

Regarding formal eloquence in place of substantive truth, the findings indicate that polished sentence structures, precise punctuation, and the machine's authoritative tone produce perceptual fluency in the prefrontal cortex. When information is processed with minimal cognitive friction, neural inhibition and methodological skepticism are suspended; fluency of form consequently replaces propositional truth.

Convergence with previous research: This finding is consistent with Kahneman's (2011) dual-process theory and the principle of cognitive ease. The brain tends to accept fluently processed texts automatically through System 1 while conserving System 2 effort. Bender et al.'s (2021) work on stochastic parrots and Pennycook and Rand's (2020) research on acceptance of pseudoscientific claims through fascination with lexical structure also support this finding [10].

Divergence and critical perspective: The finding conflicts with assumptions in conventional media literacy paradigms and some traditional defenses of AI in education [9], according to which learners would immediately detect and reject machine errors because of their artificial character. Our findings reveal that, in recent language models, AI's rhetorical beauty not only fails to facilitate error detection but deliberately conceals errors.

For appeal to artificial authority, or presumed algorithmic omniscience, the data indicate a collapse of epistemic agency and analytical alienation among students. By attributing infallibility and statistical comprehensiveness to algorithms, learners abandon their position as self-aware subjects and become mere consumers of manufactured propositions. This epistemic submission prevents critical engagement with fabricated references and machine-generated data.

Convergence with previous research: This theme aligns with automation bias as discussed by Parasuraman and Manzey (2010), Zuboff's (2019) critique of the epistemic hegemony of computational systems, and Pinto et al.'s (2023) examination of elite submission to big data.

Divergence and critical perspective: This result directly challenges theories of human–machine collaborative learning and Siemens's (2005) connectivist ideas, which regard technology as facilitating an equitable redistribution of epistemic power [11]. The present findings indicate that hallucination-producing AI does not foster cognitive empowerment; instead, it creates debilitating dependence and an illegitimate concentration of authority within the algorithmic black box.

For false neural causation, or the immediate relief error, the findings show that immediate, definitive AI answers rapidly satisfy the learner's dopaminergic reward circuit. This instantly relieves the strain of not knowing or research anxiety. Learners conflate emotional relief and the calm of task completion with scientific adequacy and content accuracy, accepting the delivered hallucinations as certain.

Convergence with previous research: These results align with cognitive neuroscience findings concerning uncertainty reduction and the role of dopamine release in relieving doubt-related tension [4]. They also converge with Sparrow et al.'s (2011) research on Google's effects on cognitive laziness and false satisfaction through mental closure [12].

Divergence and critical perspective: These findings challenge technology-based inquiry learning approaches that assume faster access to answers frees time for deeper reflection. The qualitative evidence indicates that immediate access destroys the conditions for deep reflection in favor of premature pleasure in finishing the task.

For generalization from a fabricated part to the truth of the whole, the findings show that generative AI traps learners in the familiarity heuristic by embedding genuine keywords, book titles, or theorists' names within fabricated theories or data. After verifying a single reference or minor proposition, learners hastily validate the entire unreliable structure. This conclusion is consistent with Nisbett and Wilson's (1977) halo effect and models of anchoring bias in educational decision-making [13]. It also aligns with Ji et al.'s (2023) research on hallucination architecture in language models, indicating that hallucinations become more persuasive when mixed with factual information. This result contradicts the assumption of a discerning learner in self-regulated learning models [14], which presumes that postgraduate students can filter inconsistent and false information by checking part of the data. Our qualitative findings instead indicate that the correct fragment provides cover for extending and justifying subsequent fabrications.

For anthropomorphic analogy, the findings reveal that learners attribute subjective awareness, genuine understanding, and moral honesty to the system. Algorithmic apologies and empathetic feedback lead learners to interpret hallucination as an inadvertent, well-intentioned machine mistake rather than an inherent limitation of statistical next-token prediction. This finding aligns with Reeves and Nass's (1996) Computers Are Social Actors (CASA) theory and Turkle's (2021) critiques of artificial empathy [15]. Shanahan's (2024) studies of the illusion of understanding in language systems likewise support the identification of students' erroneous attribution of meaningful agency to processes without a lifeworld [16]. This insight conflicts with transhumanist and eliminative materialist perspectives that reduce the distinction between machine processing and human thought to computational capacity. Our results establish that conflating the two produces an emotional-epistemic trap rather than an epistemic leap, drawing learners into imagined emotional reciprocity.

For confirmatory conformity, or wishful thinking bias, the findings show that training based on human feedback and the tendency to please users (sycophancy) lead generative AI to align outputs with prompt assumptions. Learners are pleased by imaginary propositions that confirm their personal biases and enter a shared cycle of human-algorithm self-deception. This result aligns with Wason's (1960) classic work on confirmation bias and Perez et al.'s (2022) contemporary research on sycophantic behavior in AI models [17]. It indicates that systems tend to sacrifice truth to conformity with the user's viewpoint. The finding contradicts the traditional idea that technology enables Popperian falsification and escape from personal echo chambers. Hallucination-producing AI becomes a catalyst for sanctifying pre-existing biases instead of a tool for intellectual challenge.

7. Final Conclusion and Policy and Pedagogical Implications

Taken together, the findings from the six analytical tables indicate that hallucination in AI systems is not merely a glitch or transient technical defect. Within neural-cognitive interactions with learners, it becomes a syndrome of epistemic erosion. Reducing critical thinking to immediate synaptic gratification and

sacrificing propositional truth to the eloquence of statistical rhetoric empties the academic knowing subject of its identity.

Addressing this challenge requires a multifaceted redesign of higher education in the age of AI:

1. Transition to relational ethics and critical reflexivity: Pedagogy must move learners away from submissive consumption of fluent answers and restore their sensitivity to the gap between linguistic form and epistemic truth.
2. Teach falsification literacy and stop rewarding speed: University assessment should focus on the friction-filled research process, engagement with disconfirming evidence, and detection of fabrication, rather than the production of polished texts in the shortest possible time.
3. Remove the sanctification and anthropomorphism of machines: Curricula should teach the epistemological foundations of language-model operation so that students understand these systems as mechanical processes based on statistical fitting, inherently lacking experience, understanding, and commitment to truth, rather than as omniscient authorities or dialogue partners.

References

- [1] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM FAccT Conference* (pp. 610–623). Association for Computing Machinery.
<https://doi.org/10.1145/3442188.3445922>
- [2] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Yan, Y., Ishii, E., Bang, Y. J., Dai, W., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38.
<https://doi.org/10.1145/3571730>
- [3] Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
- [4] Schultz, W. (2016). Dopamine reward prediction-error signalling: A two-component response. *Nature Reviews Neuroscience*, 17(3), 183–195.
<https://doi.org/10.1038/nrn.2015.26>
- [5] Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 381–410
<https://doi.org/10.1177/0018720810376055>
- [6] Perez, E., Ringer, S., Lukošiušė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, S., Esiobu, D., Durmus, E., & Bowman, S. R. (2022). Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*.
- [7] Turkle, S. (2021). *The empathy diaries: A memoir*. Penguin Press.
- [8] Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.
- [9] Luckin, R. (2018). *Machine learning and human intelligence: The future of education in the 21st century*. UCL Institute of Education Press.
- [10] Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytical thinking. *Journal of Personality*, 88(2), 185–200.
<https://doi.org/10.1111/jopy.12476>
- [11] Siemens, G. (2005). Connectivism: A learning theory for the digital age. *International Journal of Instructional Technology and Distance Learning*, 2(1), 3–10.
- [12] Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333(6043), 776–778.
<https://doi.org/10.1126/science.1207745>

- [13] Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, 35(4), 250–256.
<https://doi.org/10.1037/0022-3514.35.4.250>
- [14] Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41(2), 64–70.
https://doi.org/10.1207/s15430421tip4102_2
- [15] Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.
- [16] Shanahan, M. (2024). Role play with large language models. *Nature*, 626(7999), 493–498.
<https://doi.org/10.1038/s41586-023-06647-8>
- [17] Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12(3), 129–140.
<https://doi.org/10.1080/17470216008416717>