Confabulation, Not Hallucination: Rethinking False Outputs in Generative AI

Umberto Volpe[1], Sophia Frangou[2,3]

1. Unit of Clinical Psychiatry, Department of Clinical Neurosciences/DIMSC, School of Medicine, Università Politecnica delle Marche, Ancona – ITALY.

2. Djavad Mowafaghian Centre for Brain Health, Department of Psychiatry, University of British Columbia, Canada.

3. Department of Psychiatry, Icahn School of Medicine at Mount Sinai, USA.

Email: prof.umberto.volpe@gmail.com

Few psychiatric concepts have migrated as rapidly into technological discourse as hallucination. In generative artificial intelligence (AI), particularly large language models (LLMs), the term is widely used to describe outputs that are plausible but factually unsupported. These falsehoods represent a major risk of generative AI because they can be difficult to detect. Fluent and authoritative text may conceal fabricated citations, inaccurate summaries, incorrect recommendations or unsupported interpretations. For lay users, clinicians and researchers alike, the danger is misplaced trust in coherent text that appears grounded when it is not (1,2). Terminology matters because it should represent the problem accurately, not merely name it conveniently. This Viewpoint argues that “hallucination” is conceptually problematic and that “confabulation” offers a more accurate analogy for AI-generated falsehoods. We first revisit the meaning of hallucination, then consider confabulation as a model of narrative error, before outlining approaches to detection, governance and mitigation.

Source Attribution: Why AI Falsehoods Are Not Hallucinations

In phenomenology, hallucinations presuppose a specific perceptual modality, a conscious subject and a disturbance in the attribution of experience to its source. Jaspers (1973) (3) emphasised that hallucinations are characterised by immediate experiential certainty, distinguishing them from imagination, memory or belief. In this sense, hallucinations are not simply false perceptions, but

This peer-reviewed article has been accepted for publication but not yet copyedited or typeset, and so may be subject to change during the production process. The article is considered published and may be cited using its DOI.

This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is unaltered and is properly cited. The written permission of Cambridge University Press must be obtained for commercial re-use or in order to create a derivative work.

https://doi.org/10.1192/j.eurpsy.2026.12237 Published online by Cambridge University Press

internally generated perceptual experiences that are experienced as externally real. Contemporary neuroscience similarly frames perception as an inferential process. Predictive-processing accounts propose that hallucinations may arise when top-down predictions are given excessive weight relative to sensory input, producing perception-like experiences without corresponding external stimuli (4,5). This framework is relevant because it highlights a form of failed source attribution, whereby internally generated perceptual content is insufficiently constrained by incoming sensory evidence.

However, this does not make human hallucinations equivalent to AI-generated falsehoods. In human hallucinations, the disputed source is perceptual. In AI systems, there is no perceptual experience and no conscious subject who processes, attributes or evaluates that experience. The relevant failure in the generation of plausible but factually false AI outputs is therefore not perceptual source attribution, but evidential source verification.Thus, even within a computational account of hallucination, the analogy with AI remains limited. Human hallucinations concern perception and experience while AI confabulations concern generation, provenance and verification.

Source Verification: Why AI Falsehoods Resemble Confabulations

In phenomenology, confabulation refers to the production of false or distorted narratives without intent to deceive (6). These may range from minor distortions to elaborate autobiographical fabrications involving identity, place or life history (6). Unlike hallucinations, confabulations do not involve sensory experience. The central difficulty lies in the organisation and verification of internally generated material. Accordingly, confabulation has been linked to impaired strategic retrieval, deficient monitoring of internally generated information, and failures to place memories in their appropriate temporal and contextual frame (7-9). Confabulated accounts may also show positive or wish-fulfilling bias, with narratives that appear to support self-enhancement or psychological coherence (10).

These features make confabulation a closer analogy to AI-generated falsehoods than hallucination. Like human confabulations, false AI outputs may be coherent, plausible and assembled from fragments of available information. In AI systems, such falsehoods arise from inadequate source verification. This can be caused by the probabilistic nature of text generation, weak grounding in external evidence, failure to retrieve or use relevant sources, misleading prompt context, and optimisation for fluency or user satisfaction rather than truth. AI systems outputs may show biases that have a limited resemblance to positive bias in human confabulation. One example is sycophancy, where models tend to agree with, reassure or please users even when this compromises accuracy (11). The crucial difference is that human confabulation is shaped by memory, subjectivity and emotional salience, whereas AI confabulation is shaped by training data, statistical prediction, prompt context, system design and optimisation procedures. In both cases,

https://doi.org/10.1192/j.eurpsy.2026.12237 Published online by Cambridge University Press

however, the core problem is the production of a coherent narrative without adequate verification against external evidence.

Confabulation in AI Governance

The distinction between halluciation and confabulation is beginning to appear in AI risk-management language. The U.S. National Institute of Standards and Technology (NIST) uses “confabulation” as a generative-AI risk category (12). NIST defines confabulation as the production of confidently stated but erroneous or false content, noting that such outputs are often referred to colloquially as “hallucinations” or “fabrications”. The EU–U.S. Trade and Technology Council’s terminology document lists “confabulation” as a term also known as “hallucination” and defines it as the generation of inaccurate or false responses that appear plausible to users (13). In the UK, Government and Information Commissioner’s Office documents also recognise “confabulation”, but generally treat it as secondary to “hallucination” (14-16).

Why terminology matters: Detection, mitigation and governance of AI confabulation

The distinction between hallucination and confabulation is not merely semantic. Recognising AI falsehoods as confabulations reframes the practical challenge. The problem is not an anomalous perceptual event, but the generation of coherent content that is insufficiently constrained by evidence, provenance or verification (10-16).

No dependable method can currently determine, in all cases, whether a specific AI-generated answer is factually correct. Several safeguards can reduce risk. Retrieval-augmented generation links model outputs to external sources, such as scientific papers, clinical guidelines or institutional documents, but retrieved sources can still be misread or misapplied (17). Citations improve inspectability, but a real citation does not guarantee that the cited work supports the claim. Uncertainty methods, including self-consistency approaches, can flag unstable answers, but repeated consistency does not prove truth (18). Benchmarking frameworks such as the Holistic Evaluation of Language Models (HELM) support systematic model evaluation across dimensions including accuracy, robustness, calibration and fairness, but benchmarks cannot capture the full complexity of real-world use (19). Provenance standards such as those developed by the Coalition for Content Provenance and Authenticity can document whether digital content was AI-generated or modified, but provenance does not establish factual accuracy (20).

Human oversight remains essential, especially in high-stakes settings, but it is not sufficient on its own. AI confabulations place aly users most at risk but clinicians and researchers may miss subtle errors when outputs are fluent, authoritative or outside their immediate expertise. Safe deployment will therefore require layered safeguards, including evidence grounding, source verification,

https://doi.org/10.1192/j.eurpsy.2026.12237 Published online by Cambridge University Press

uncertainty signalling, rigorous evaluation, provenance tracking, post-deployment monitoring and clear accountability for human review. Until such standards are mature and widely adopted, AI confabulation remains difficult to detect, quantify and govern.

Conclusions

The language used to describe AI errors is therefore not neutral. “Hallucination” mislocates the problem in perception, whereas “confabulation” identifies the practical hazard more precisely. Recognising this is the first step toward safer use, because the appropriate response is to build systems, standards and habits of use that make verification explicit.

References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-1940.
2. Kwong JCC, Wang SCY, Nickel GC, Cacciamani GE, Kvedar JC. The long but necessary road to responsible use of large language models in healthcare research. NPJ Digit Med. 2024;7(1):177.
3. Jaspers K. (1973). General psychopathology (J. Hoenig & M. W. Hamilton, Trans.). University of Chicago Press. (Original work published 1913)
4. Fletcher PC, Frith CD. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. Nat Rev Neurosci. 2009;10(1):48-58.
5. Friston K. The free-energy principle: a unified brain theory? Nat Rev Neurosci. 2010;11(2):127-38.
6. Hirstein W. Confabulation: Views from Neuroscience, Psychiatry, Psychology and Philosophy (Oxford, 2009; online edn, Oxford Academic, 1 Feb. 2010), https://doi.org/10.1093/acprof:oso/9780199208913.001.0001, accessed 23 June 2026..
7. Moscovitch M. Confabulation and the frontal systems: strategic versus associative retrieval. In: Roediger HL III, Craik FIM, editors. Varieties of memory and consciousness: essays in honour of Endel Tulving. Hillsdale (NJ): Lawrence Erlbaum Associates; 1989. p. 133-160.
8. Burgess PW, Shallice T. Confabulation and the control of recollection. Memory. 1996;4(4):359-411.
9. Schnider A. Spontaneous confabulation and the adaptation of thought to ongoing reality. Nat Rev Neurosci. 2003;4(8):662-71.
10. Turnbull OH, Berry H, Evans CE. A positive emotional bias in confabulatory false beliefs about place. Brain Cogn. 2004;55(3):490-4.

https://doi.org/10.1192/j.eurpsy.2026.12237 Published online by Cambridge University Press

11. Sharma M, Tong M, Korbak T, Duvenaud D, Askell A, Bowman SR, et al. Towards understanding sycophancy in language models. In: Proceedings of the 12th International Conference on Learning Representations; 2024 May 7-11; Vienna, Austria. OpenReview.net; 2024. Available from: https://arxiv.org/abs/2310.1354
12. Autio C, Schwartz R, Dunietz J, Jain S, Stanley M, Tabassi E, Hall P, Roberts K. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. Gaithersburg (MD): National Institute of Standards and Technology; 2024 Jul 26. Report No.: NIST AI 600-1. doi:10.6028/NIST.AI.600-1.
13. EU-U.S. Trade and Technology Council Working Group 1, Subgroup on AI Taxonomy and Terminology. EU-U.S. terminology and taxonomy for artificial intelligence: second edition [Internet]. Brussels: European Commission; 2024 Apr 5 [cited 2026 Jul 5]. Available from: https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence-second-edition
14. Cabinet Office, Government Digital Service, Central Digital and Data Office. Generative AI Framework for HMG [Internet]. London: Cabinet Office; 2024 Jan 18 [cited 2026 Jul 5]. Available from: https://www.gov.uk/government/publications/generative-ai-framework-for-hmg
15. Department for Science, Innovation and Technology, Government Digital Service, Central Digital and Data Office. Artificial Intelligence Playbook for the UK Government [Internet]. London: Department for Science, Innovation and Technology; 2025 Feb 10 [cited 2026 Jul 5]. Available from: https://www.gov.uk/government/publications/ai-playbook-for-the-uk-government/artificial-intelligence-playbook-for-the-uk-government-html
16. Information Commissioner's Office. Glossary: response to the consultation series on generative AI [Internet]. Wilmslow: Information Commissioner's Office; [date unknown] [cited 2026 Jul 5]. Available from: https://ico.org.uk/about-the-ico/what-we-do/our-work-on-artificial-intelligence/response-to-the-consultation-series-on-generative-ai/glossary/
17. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in Neural Information Processing Systems 33. Red Hook (NY): Curran Associates; 2020. p. 9459-9474
18. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Zhou D. Self-consistency improves chain of thought reasoning in language models [Internet]. c2022 [cited 2026 Jul 5]. Available from: https://openreview.net/forum?id=1PL1NIMMrw
19. Liang P, Bommasani R, Lee T, Tsipras D, Soylu D, Yasunaga M, et al. Holistic evaluation of language models. Trans Mach Learn Res. 2023. [cited 2026 Jul 5]. Available from https://openreview.net/pdf?id=iO4LZibEqW

https://doi.org/10.1192/j.eurpsy.2026.12237 Published online by Cambridge University Press

20. Coalition for Content Provenance and Authenticity. C2PA technical specification version 2.4 [Internet]. San Francisco: Coalition for Content Provenance and Authenticity; 2025 [cited 2026 Jul 5]. Available from: https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html

https://doi.org/10.1192/j.eurpsy.2026.12237 Published online by Cambridge University Press

Table 1. Conceptual distinctions between human hallucination, human confabulation and AI-generated false output

Dimension	Human hallucination	Human confabulation	AI falsehood
Level of error	Perceptual experience	Narrative reconstruction	Generative output
Source problem	Internally generated perceptual content is experienced as externally real	Internally generated memories or explanations are not adequately checked against context or evidence	Generated content is not adequately grounded in reliable sources or verification
Typical falsehood	A perception-like experience without a corresponding stimulus	A coherent but false account, memory or explanation	A coherent but false narrative, citation, summary or recommendation
Why it matters	The problem concerns source attribution	The problem concerns verification of source	The problem concerns source verification
What must be checked	Whether the experience corresponds to an external stimulus or is internally generated	Whether a narrative is anchored in accurate memory, context and external evidence	Whether the source exists, supports the claim, and has been accurately represented

https://doi.org/10.1192/j.eurpsy.2026.12237 Published online by Cambridge University Press