Hallucinations in automated texts – A critical view on the emerging terminology

Annette Gerstenberg (University of Potsdam)

gerstenberg(at)uni-potsdam.de

Abstract

The term *hallucination* is used in AI discourse to describe AI-generated outputs that are unfounded and lack backing in input data, a phenomenon which occurs frequently enough for the academic community to shun widespread collaboration with AI writing tools, particularly in certain disciplines. The issue of hallucinatory outputs will diminish as AI advances, but the appropriateness of using the terms *hallucinate* and *hallucination* in this context remains under debate. Originating from Latin, terms related to *hallucination* were once exclusively used within specialized medical terminology. Over time, and across languages, its metaphorical meaning has evolved, becoming part of colloquial language through the semantic innovation characteristic of youth language. We also show how this trend coincides with the use of metaphors such as anthropomorphisms in scientific discourse. The newest addition to *hallucination*'s list of meanings – an established metaphor for a new phenomenon – is contentious. On the one hand, there is a similarity in that both AI-induced and medically induced hallucinations are elusive and difficult to identify. On the other hand, it suggests a trivialization of medical conditions that remain significantly stigmatized in contemporary societies.

Keywords

hallucination, metaphor, AI, NLG, language for special purposes

1 Hallucinating in the context of AI*

In his editorial to the psychiatric journal Schizophrenia (9:52, 2023), Robin Emsley reported his experiences of using ChatGPT for academic writing. He found that the AI generated results full of misleading content and fictitious references. He concluded that he "would not recommend collaborating with a colleague with pseudologia fantastica" and thus would "not recommend ChatGPT as an aid to scientific writing" (Emsley 2023: 2). This experience is certainly widely shared among the academic community, but will no doubt become less common over time (Brin, Sorin, Vaid et al. 2023), alongside the improvement of prompting strategies and the establishment of specialized databases to replace the "pseudologia fantastica".

As a psychiatrist, Emsley argues against a recurrent term in the scientific discussion surrounding Artificial Intelligence (AI): the metaphorical use of *hallucination* to describe generated texts which lack the referential anchoring of a body of foundational, underlying texts. Consequently, he explicitly and consistently replaces *hallucinations* with alternative terms, and clarifies: "This [hallucinations] is a misnomer. Hallucinations are false perceptions. What I experienced were fabrications and falsifications. […] Or, if the absence of malicious intent is assumed, confabulations would be a better description." (Emsley 2023: 1, quoting McGowan 2023).

Nevertheless, *hallucination* has entered the cadre of artificial intelligence terminology, through which it has recently entered into wider public discourse. CDE named *hallucinate* in the context of artificial intelligence as the word of the year 2023 (WOTY), due in part to the recent surge in usage and the fact that *hallucinate* captures both the "potential strengths and its current weaknesses" of AI (Cambridge Dictionary Team 2023). In this way, the WOTY status celebrates the acquisition of a new meaning for *hallucinate*, albeit without explicit discussion of its origins and critique of its usage. On the contrary, it is emphasized that "our psychological vocabulary will be further extended to encompass the strange abilities of the new intelligences we're creating" (Henry Shevlin in Cambridge Dictionary Team 2023).

The term *hallucination* has been used in computational sciences for roughly 20 years. Originally, the meaning was a positive one. Baker and Kanade (2000) is commonly referred to as the first instance of *hallucinate* in computational sciences, in the context of Computer Vision, to encapsulate a particular benefit of their work; the hallucination algorithm is employed to enhance the resolution of faces.

Since then, *hallucination* has become a common term in the discussion of AI. In a review of models for abstractive text summarization, to meet the basic definition of a *hallucination*, a summary must have "a span(s) $w_i \ldots w_{i+j}$, $j \geq i$, that is not supported by the input document" (Maynez, Narayan, Shashi et al. 2020: 1906). A more fine-grained definition differentiates between intrinsic and extrinsic hallucinations; intrinsic hallucinations result from the synthesis of information present in the input text, while extrinsic hallucinations "ignore the source material altogether" (Maynez, Narayan, Shashi et al. 2020: 1908). These different types of *falsifications* (Emsley's 2023 proposed *hallucinations* substitute) are generally considered to be a challenge of AI generated texts (Salvagno, Taccone and Gerli 2023), for which different control strategies have been developed (Filippova 2020).

2 Metaphor versus medicine

The metaphorical use of the term *hallucination* has become the object of critical discussion well beyond Emsley (2023). By contrast, Ji, Lee, Frieske et al. (2023) advocate for the unequivocal use of *hallucination* in computational sciences. When referencing the standard psychological definition of *hallucination* as a perception in the absence of an appropriate stimulus, neither the individuals who experience hallucinations are mentioned, nor the diseases to which they are related. In this way, the authors trace the similarities between the psychological term and the Natural Language Generation (NLG) meaning:

> Hallucinated text gives the impression of being fluent and natural despite being unfaithful and nonsensical. It appears to be grounded in the real context provided, although it is actually hard to specify or verify the existence of such contexts. Similar to psychological hallucination, which is hard to tell apart from other "real" perceptions, hallucinated text is also hard to capture at first glance. (Ji, Lee, Frieske et al. 2023: 3).

Thus, according to Ji, Lee, Frieske et al. (2023), the semantic trait shared by the NLG term *hallucination* and its scientific definition is the difficulty both medically induced hallucinations and AI hallucinations pose in their identification as such.

This claimed similarity neatly sidesteps the need to develop a deeper understanding of the term's meaning and contextual usage. It also fails to take both the pathological contexts in which hallucinations can emerge, and the individuals affected by the related diseases into consideration. These include psychoses, mostly occurring with schizophrenia, affective disorders, dementia, and delirium, among others (Pschyrembel 2021, s.v. *hallucination*).

The inclusion of affected individuals is achieved by Østergaard and Nielbo's definition (2023). As experts from the field of psychiatry, they discuss the inappropriate use of *hallucination* in AI related contexts. Firstly, the core of the psychiatric definition is the sensory perception on an individual level of a *hallucination*, because the term is non-precise. This is not the case for *hallucinations* in AI. Secondly, and more importantly, the authors underline the fact that *hallucination* is not an abstract medical term, but rather one which is linked to severe diseases. The affected patients not only suffer from its symptoms directly, but can also be subject to stigmatization, as they are labelled publicly with offensive terms *schizophrenic*. They conclude with an "appeal to the field of AI to change its labeling of false responses. There is no need to use an imprecise and stigmatizing metaphor when there are already specific labels available" (Østergaard and Nielbo 2023: 1107).

3 Hallucinating over time and language

With this in mind, the recent scientific use of *hallucination* warrants closer inspection. With the 1st international conference on "automated texts in the romance languages" (ai-rom) at TU Dresden 2023 serving as motivation, comparing the English and Romance Languages dictionary entries revealed significant differences in their indications of earliest attested use and original Latin etymon. In what follows, not only do we situate this use within a longstanding tradition in the use of medical terms in varieties labelled as youth languages, by reviewing historical and modern cross-linguistic dictionaries. We also show how this trend coincides with the use of metaphors such as anthropomorphisms in scientific discourse.

In the modern languages, the derivates of the Latin (H)ALUCINARI Latin 'aberrare et non consistere', English 'get lost, not remain valid'[1], a loan word of ancient Greek (TLL), are key features of the languages of special purposes, especially in medicine and psychology. In English, *hallucination* is defined as the "mental condition of being deceived or mistaken […] an illusion" and – in the domain of pathology and psychology – as the "apparent perception (usually by sight or hearing) of an external object when no such object is actually present" (OED, s.v.). The German *Halluzination*, according to DWDS (s.v.), is mostly used in medical language of special purposes.

The French *hallucination* is traced back to the Latin etymon HALLUCINATIO (attested since 1660, PRob, s.v., and TLFi, s.v.); alongside its medical definition, the usage in "current language" is listed, as a result of "exaggeration", and a familiar use of the verb *halluciner*. In a similar fashion, the Italian *allucinazióne* is initially

[1] All translations in this article are our own.

defined with the medical meaning, followed by a more extended meaning. The term is traced back to the Latin ALUCINARI, and attested since 1728 (DELI s.v., also in GDLI). By contrast, the Spanish *alucinación* is listed as derived from the Latin ALLUCINATIO (without indication of earliest use), and the scientific meaning is described as a "subjective sensation with no preceding impression in the senses" (DRAE, s.v.). For the derived verb, a colloquial use is attested for Argentina and Uruguay (DRAE, s.v. *alucinar*).

Such an extension of medical or other scientific terms is far from rare in language use; youth language in particular is characterized by an intense use of this type of innovation. For related varieties, as semantic innovations and metaphors being one of the core features of youth language, the particularly shocking metaphorical use of diseases, especially psychological, has been demonstrated (Neuland 2018: 74). In European youth languages, diseases are a common source for new metaphors. To name some examples, Italian *arterio, arterioso, artèrio* < *arteri(osclerotic)o* '(youth language) father, mother, adult' (Radtke 1998: 64), German *Spasti* < spastically paralyzed person '(youth language) idiot, stupid person' (Neuland 1998: 79), or designating sensory or mental issues such as German *keine klaren Bilder sehen*, 'not seeing clear pictures' (Heinemann 1990: 46). Narcotics also play a role in the evolution of metaphors. For instance, the German *Trip*, which describes the state resulting from the abuse of narcotics and hallucinogens (Müller-Thurau and Marcks 1985: 171; see also 45–50; 91), and Italian *allucinato* '(youth language) being a victim of narcotics' (Banfi 1992: 128; see also SG, s.v.). As for French, one meaning of *délirer* in youth language is to 'have fun; amuse' (Soudot 1997: 63). This latter term is also documented as an umbrella term for speaking styles of adolescents and their ludic and poetic functions (Seux 1997: 96; 100). As for psychology, related forms have been established since the 1980s or 1990s (attested in Marcato 1997: 565, based on data collected in the area of Mestre), such as *panico*, *paranoia* and other terms, originally designating diseases or their symptoms. This is also true for derivates such as the Italian adjective, as the lexicalized participle present *allucinante* which is used, documented since 1980, in a very broad sense (DLG, s.v.), "especially in the youth language" (Treccani Vocabolario, s.v.).

Paranoia, a similar term originating in specialized medical discourse that has spread to everyday usage, is further advanced in its dissemination as compared to *hallucination*, as indicated by its dictionary descriptions. Words relating to *paranoia* are more frequently listed as falling under a familiar register (French *parano* < *paranoïa*, PRob "familiar abbreviation", attested before 1971; Italian *paranòia*, attested in colloquial speech as expressing a non-pathological fear or fixation, Zingarelli, s.v.). A colloquial meaning of *paranoia*, and of *schizo* < *schizophrenic* is lexicalized in English (Cambridge Dictionary of English, s.v.). However, a colloquial meaning of *hallucinate* is not included in the dictionary. Accordingly, Sp. *alucinar* is translated as 'to be amazed' (Cambridge Dictionary Spanish-English, s.v.), and as equivalent for It. *allucinante*, 'amazing, unbelievable, incredible' is given (Cambridge Dictionary Italian-English, s.v.). This means that there is no evidence of a direct link between youth language and the very latest technical innovation *hallucination*, but there is evidence of the underlying semantic process.

Most recently, the two meanings of *hallucination* for 'false information that is produced by an artificial intelligence', and 'the fact of an artificial intelligence […] producing false information' have been recorded in CDE (s.v., see also Merriam-Webster), to supplement the medical and metaphorical entries. Also, the verb *hallucinate* 'when an artificial intelligence […] hallucinates, it produces false information' has been recorded (CDE, s.v.) or documented with examples (Merriam-Webster, s.v.). However, its use is confined; the medical term or precise definitions of sensual perceptions are consistently at the forefront of traditional definitions, and the newest meaning is not yet accounted for in other contemporary dictionaries (DRAE; DWDS; PRob; Zingarelli).

The new meaning in AI discourse connects not only to semantic innovations in youth and colloquial language, but also to a common figure in scientific discourse. The semantic feature 'without a correlate in reality' is transferred from the source domain of HUMAN to the field of technology and can be considered an extension via anthropomorphism. The target domain narrows the target domain from the original broad meaning of *hallucination*: while *hallucination* can denote pure sensory perceptions without linguistic realization, *hallucination* in its new meaning is related to texts. Documented in scientific discourse, anthropomorphism such as related metaphors (Brown 2003; Ickler 1993: 104 on astronomy and astrophysics) is an important factor impacting on scientific thinking, especially when, in initial stages, old linguistic signs are used for new insights and perspectives (Burkhardt 1987; Ureña Gómez-Moreno 2016). The metaphorical process of anthropomorphism is well established in the discourse around human-machine, or human-robot communication (Westerman, Cross and Lindmark 2019).

4 To be continued…

Our short documentation suggested that the use of *hallucination* related to AI is the result of semantic innovation, a process widespread in youth language, that has already led to more colloquial uses of the originally specialized medical term. Also, metaphors and even anthropomorphisms have been widely used in different scientific disciplines, especially in early stages of new research areas.

In its new contexts, the use of *hallucination* no longer includes any reference to those individuals who hallucinate on medical grounds, although the negative stigmatizations they experience take new forms through current AI usage discourse. If the popularization of *hallucination* took place in part thanks to the popularity of youth language itself, and in part because metaphors primarily accompany initial phases of scientific innovations, perhaps as the field of AI and the discourse surrounding it develops, this rather questionable feature will be left behind. This would in turn create the space for a more semantically precise term such as *unsubstantiated content*, already used in the discussion of the taxonomy of LLM-generated misinformation (Chen and Shu 2023).

This new journal seems to be the perfect place for the linguistic discussion not only on AI generated output, but also on the impact of the surrounding discourse and terminology.

References

Baker, Simon & Kanade, Takeo. 2000. Hallucinating faces. *Proceedings Fourth IEEE International Conference 2000*. 83–88. DOI 10.1109/AFGR.2000.840616

Banfi, Emanuele. 1992. Conoscenza e uso di lessico giovanile a Milano e a Trento. In Banfi, Emanuele & Sobrero, Alberto A. (eds), *Il linguaggio giovanile degli anni novanta*. 99–138. Milano: Laterza.

Brin, Dana & Sorin, Vera & Vaid, Akhil & Soroush, Ali & Glicksberg, Benjamin S. & Charney, Alexander W. & Nadkarni, Girish & Klang, Eyal. 2023. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Scientific Reports* 13(1). 16492. DOI 10.1038/s41598-023-43436-9.

Brown, Theodore L. 2003. *Making Truth. Metaphor in Science*. Urbana (IL): University of Illinois Press.

Burkhardt, Armin. 1987. Wie die 'wahre Welt' endlich zur Metapher wurde. Zur Konstitution, Leistung und Typologie der Metapher. *Conceptus: Zeitschrift für Philosophie* 21(52). 39–67.

CDE = 2024. *Cambridge Dictionary*. Cambridge: Cambridge University Press, online (https://dictionary.cambridge.org/dictionary/english) (accessed 2024-05-10)

Cambridge Dictionary Team. 2023. *The Cambridge Dictionary Word of the Year 2023 is …* Cambridge: Cambridge University Press, online. (https://dictionary.cambridge.org/editorial/woty) (accessed 2024-05-10)

Chen, Canyu, & Shu, Kai 2023. Combating Misinformation in the Age of LLMs: Opportunities and Challenges. arXiv. http://arxiv.org/abs/2311.05656.

DELI = Cortelazzo, Manlio & Zolli, Paolo. 2002. *Il nuovo etimologico: Dizionario etimologico della lingua italiana*. Bologna: Zanichelli.

DLG = Manzoni, Gian R. & Dal Monte, Emilio. 1980. *Pesta duro e vai tránquilo: dizionario del linguaggio giovanile.* Milano: Feltrinelli Economica.

DRAE = Real Academia Española (ed.). 2001. *Diccionario de la lengua española*. 22nd edn. Madrid: RAE. (https://dle.rae.es) (accessed 2024-05-10)

DWDS = 2024. *Digitales Wörterbuch der Deutschen Sprache.* Berlin: Berlin-Brandenburgische Akademie der Wissenschaften, online. https://www.dwds.de (accessed 2024-05-10).

Emsley, Robin. 2023. ChatGPT: these are not hallucinations – they're fabrications and falsifications. *Schizophrenia (Heidelberg, Germany)* 9(1). 52. DOI 10.1038/s41537-023-00379-4

Filippova, Katja. 2020. Controlled Hallucinations: Learning to Generate Faithfully from Noisy Data. *Findings of the Association for Computational Linguistics: EMNLP 2020*. 864–870.

GDLI = Battaglia, Salvatore & Ronco, Giovanni. 1973. *Grande dizionario della lingua italiana.* Torino: UTET.

Heinemann, Margot. 1990. *Kleines Wörterbuch der Jugendsprache*, 2nd ed. Leipzig: VEB Bibliographisches Institut.

Ickler, Theodor. 1993. Zur Funktion der Metapher, besonders in Fachtexten. *Fachsprache* 15 (3-4). 94–110.

Ji, Ziwei & Lee, Nayeon & Frieske, Rita & Yu, Tiezheng & Su, Dan & Xu, Yan & Ishii, Etsuko & Bang, Ye Jin & Madotto, Andrea & Fung, Pascale. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys* 55(12). 1–38. DOI 10.1145/3571730

Marcato, Carla. 1997. In para totale…una cosa da panico…: sulla lingua dei giovani in Italia. *Italica* 74 (4). 560. DOI 10.2307/479484

Maynez, Joshua & Narayan, Shashi & Bohnet, Bernd & McDonald, Ryan. 2020. On faithfulness and factuality in abstractive summarization. *arXiv*. DOI 10.48550/arXiv.2005.0061

McGowan, Alessia & Gui, Yunlai & Dobbs, Matthew & Shuster, Sophia & Cotter, Matthew & Selloni, Alexandria & Goodman, Marianne & Srivastava, Agrima & Cecchi, Guillermo A. & Corcoran, Cheryl M. 2023. ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search. *Psychiatry research* 326. 115334. DOI 10.1016/j.psychres.2023.115334

Merriam-Webster = 2024. *Merriam-Webster.com Dictionary.* Springfield (MA): Merriam Webster. https://www.merriam-webster.com/dictionary (accessed 2024-05-10)

Müller-Thurau, Claus P. & Marcks, Marie. 1985. *Lexikon der Jugendsprache.* Düsseldorf: Econ.

Neuland, Eva. 1998. Zum Sprachgebrauch Jugendlicher verschiedener regionaler Herkunft. In Androutsopoulos, Jannis K. & Scholz, Arno (eds). *Jugendsprache = Langue des jeunes = youth language: linguistische und soziolinguistische Perspektiven.* 71–90. Frankfurt am Main, New York: Lang.

Neuland, Eva. 2018. *Jugendsprache: Eine* Einführung. Tübingen: Francke.

OED = 2024. *Oxford English Dictionary: The historical English dictionary.* Oxford: Oxford University Press online. DOI 10.1093/OED/5271413239

Østergaard, Søren Dinesen & Nielbo, Kristoffer Laigaard. 2023. False responses from Artificial Intelligence models are not hallucinations. *Schizophrenia bulletin* 49 (5). 1105–1107. DOI 10.1093/schbul/sbad068

PRob = Dictionnaires Le Robert. 2023. *Le Petit Robert - Version numérique (version 5.8, mai 2023).* Paris: Dictionnaires Le Robert; SEJER. https://www.lerobert.com (accessed 2024-05-10)

Pschyrembel = Pschyrembel Redaktion (ed.). 2004–2024. *Pschyrembel online* (Version 05.2021). Berlin: de Gruyter.

Radtke, Edgar. 1998. Ein italienisches Jugendsprachewörterbuch der Mikrodiachronie: Der ‘Dizionario della Lingua Parlata dei Giovani 1982–1992’ (DLPG). In Androutsopoulos, Jannis K. & Scholz, Arno (eds). *Jugendsprache = Langue des jeunes = youth language: linguistische und soziolinguistische Perspektiven.* 59–70. Frankfurt am Main, New York: Lang.

Salvagno, Michele & Taccone, Fabio Silvio & Gerli, Alberto Giovanni. 2023. Artificial intelligence hallucinations. *Critical care (London, England)* 27 (1). DOI 10.1186/s13054-023-04473-y

Seux, Bernard. 1997. Une parlure argotique de collégiens. *Langue française* 114, 82–103.

SG = Ambrogio, Renzo & Giovanni Casalegno. 2004. *Scrostati gaggio! Dizionario storico dei linguaggi giovanili.* Torino: UTET.

Sourdot, Marc. 1997. La dynamique du français des jeunes : sept ans de mouvement à travers deux enquêtes (1987–1994). *Langue française* 114. 56–81.

TLFi = Centre National de la Recherche Scientifique (CNRS), Analyse et traitement informatique de la langue française (ATILF-CNRS) & Université de Nancy 2. 2002. *Trésor de la langue française informatisée*. Nancy: ATILF http://stella.atilf.fr

TLL = Bayerische Akademie der Wissenschaften. [2024]. *Thesaurus Linguae Latinae.* München: BAW. https://tll.degruyter.com/ (accessed 2024-05-10)

Treccani Vocabolario = 2024. Il *Vocabolario Treccani*. Roma: Istituto della Enciclopedia Italiana. https://www.treccani.it/vocabolario (accessed 2023-11-15).

Ureña Gómez-Moreno, José Manuel. 2016. Refining the understanding of novel metaphor in specialised language discourse. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 22(1). 1–29. DOI 10.1075/term.22.1.01ure

Westerman, David & Cross, Aaron C. & Lindmark, Peter G. 2019. I Believe in a Thing Called Bot: Perceptions of the Humanness of "Chatbots". *Communication Studies* 70(3). 295–312. DOI 10.1080/10510974.2018.1557233

Zingarelli = Zingarelli, Nicola. 2024. *Lo Zingarelli on-line: vocabolario della lingua italiana.* Bologna: Zanichelli, https://dizionari.zanichelli.it

* Special thanks to Eman El Sherbiny Ismail and Eleanor Troth for their careful documentation and editing, as well as to the two anonymous reviewers of AI-Linguistica.