From Hallucination to Confabulation: A Philosophical Reassessment of AI Error Terminology

Kevser Erdoğan[1]

[1] Ankara University, Graduate School of Social Sciences, Ankara, Türkiye
* Corresponding author: Kevser Erdoğan (kevser.erdogan2210@gmail.com.tr)

Abstract: The term hallucination, which is frequently discussed in the field of artificial intelligence, does not fully correspond to its established usage in psychology. This situation is not merely a terminological problem but also a consequence of implicit philosophical assumptions concerning meaning, mentality, and intentionality. Definitions of AI hallucination vary considerably in the literature, ranging from "plausible-sounding errors" to formulations broad enough to risk encompassing all AI-generated outputs. Some studies argue that these phenomena have been misnamed and that the term would be more appropriately drawn from philosophy than from psychology. Conceptual clarity is therefore essential. Focusing on the relationship between artificial intelligence and psychology, this study aims to shed light on the origins of the naming, definitional, and taxonomic problems surrounding the notion of AI hallucination. To this end, foundational texts in the philosophy of artificial intelligence and the literature on AI hallucination are examined through a comparative and conceptual analysis. The analysis suggests that many researchers employ a language implicitly grounded in dualistic assumptions. At the same time, despite the tensions generated by such assumptions, the influence of psychology remains evident in the selection of concepts used to describe AI errors.

Keywords: AI Hallucination; Philosophy of Mind; Meaning; Psychology

Received: 06 June 2026 | Revised: 26 June 2026 | Accepted: 11 July 2026

Cite: Erdoğan, K. (2026). From Hallucination to Confabulation: A Philosophical Reassessment of AI Error Terminology. *Journal of Artificial Intelligence and Human Sciences*, *3*(1), 76-88. Doi: 10.5281/zenodo.21467077

1. Introduction

AI-generated content (AIGC) can be traced back to the 1950s; however, due to technical limitations, its development only accelerated in the 2010s and gained substantial momentum in 2022 with the emergence of pre-trained large-scale AI models (Sun et al., 2024, p. 2; Wang et al., 2023). A similar trajectory can be observed in the prominence of concerns surrounding AI hallucination: even if the term itself was not always employed, the underlying phenomenon began attracting comparable levels of attention during roughly the same period. At first glance, hallucinations in text generation and image processing may appear to refer to two distinct phenomena. Upon closer examination, however, the notion of AI hallucination -particularly when used in a negative sense- encompasses dozens of different referents. Although the term can be found in some pre-2000 works (Tait, 1982; Baker & Kanade, 1999), it is today most commonly used in the context of large language models (LLMs).

According to Merriam-Webster (2023), in the field of computing, hallucination refers to "a plausible but false or misleading response generated by an artificial intelligence algorithm." The same dictionary, which is frequently referenced in the relevant literature, treats *delusion*, *illusion*, *mirage*, and *hallucination* as synonyms that "mean something that is believed to be true or real but that is actually false or unreal." It may therefore be argued that concepts whose distinctions are carefully maintained in psychology and philosophy are employed in AI discourse in a manner whose boundaries remain unclear and are often justified through various analogies. Despite the ambiguity surrounding the boundaries of the concept, this terminology has rapidly gained legitimacy beyond academic circles and achieved widespread acceptance. Cambridge Dictionary introduced an additional AI-related meaning for the verb *hallucinate* -"when an artificial intelligence hallucinates, it produces false information"- and selected it as its "Word of the Year 2023." The rationale for this choice was the view that the term captures "the heart of why people are talking about AI" (Cambridge Dictionary, 2023). Originating from psychology and psychiatry, the term has also contributed to the broad public resonance of the issue. The term *AI hallucination* remained one of the most frequently discussed expressions in the literature in 2026. This

study argues that the term is incompatible with its meaning in psychology and that, in many cases -particularly in the context of LLMs- the concept of *confabulation* provides a more explanatory framework[1]. Within the AI literature, the term *hallucination* is increasingly becoming an umbrella concept, and this expansion makes visible the tension between the term's psychological origins and its contemporary usage.

No precise number of AI hallucination types can be given, as most studies propose different taxonomies. According to one study, hallucinations can be categorized as *contextual disconnection, semantic distortion, content hallucination,* and *factual inaccuracy* (Sahoo, Meharia, Ghosh, Saha, Jain, & Chadha, 2024, p. 11709). Another study classifies AI hallucinations into eight subtypes: *overfitting, logic errors, reasoning errors, mathematical errors, unfounded fabrication, factual errors, text output errors,* and *other errors* (Sun et al., 2024, p. 1). According to yet another study, hallucinations may be classified as *input-conflicting hallucinations, context-conflicting hallucinations,* and *fact-conflicting hallucinations* (Zhang et al., 2025, p. 1377). In that study, Zhang et al. argue that "unverifiable statements aligned with commonsense (e.g., the idea of a research paper)," which many would not regard as hallucinations, should nevertheless be included within the category of *fact-conflicting hallucinations* (2025, p. 1379). Furthermore, *sycophancy* -the tendency of a model to provide responses that please the user rather than responses that are true- can also be classified as a form of hallucination (Zhang et al., 2025, p. 1387; Wei et al., 2024). Zhang et al. identify several challenges associated with hallucinations in LLMs, including *massive training data, versatility of LLMs,* and the *invisibility of errors* (2025, p. 1380). Perspectives such as these confront us with the danger of classifying virtually every type of output as a hallucination, and this danger is not limited to LLMs.

Ren et al., by examining *a priori* and *a posteriori* judgment capacities, argue that LLMs have difficulty managing conflicts between *internal knowledge* and *external knowledge* due to their *unwavering confidence*, and that *retrieval augmentation* constitutes an effective approach to addressing this problem (Ren et al., 2025). In this context, RAG technology has been presented in recent years as a promising means of mitigating hallucinations (Chowdhury, 2026; Gao et al., 2023; Lewis et al., 2020). Nevertheless, this technology introduces a distinct set of limitations and potential failure modes (Barnett et al., 2024).

In another example, Zheng, Huang, and Chang (2023), focusing specifically on ChatGPT, reduce the failures that give rise to hallucinations to four categories: *comprehension, factuality, specificity,* and *inference*. Arguing that factuality constitutes the most critical of these failures, they place particular emphasis on two related capacities: *knowledge memorization* and *knowledge recall* (Zheng et al., 2023). When categorized in this manner, the term *confabulation*, with its association with memory, appears to be a more plausible choice than *hallucination*. Taken together, these examples suggest that the choice of the term *AI hallucination* itself remains a problem in need of examination. Indeed, the debate surrounding terminology constitutes one of the central concerns of the present study.

As Table 1 illustrates, most of the AI error types considered here appear to be more suitably explained by the concept of confabulation than by hallucination in its psychological sense. These error types are generally associated with knowledge retrieval, factual representation, semantic reconstruction, or gap-filling processes rather than with perceptual experiences. At the same time, certain cases -particularly object hallucination- may still retain a perceptual dimension insofar as they involve the erroneous identification of features in the external environment.

[1] In psychology, *hallucination* refers to a perceptual experience occurring in the absence of an external stimulus, whereas *confabulation* denotes the production of fabricated memories or recollections that fill gaps in memory (see Section 3 for a detailed discussion).

Table 1. Representative AI Hallucination Categories and Their Correspondence to Confabulation

AI Error Types	Source	Relation to Confabulation
Factual errors	Sun et al., 2024; Sahoo et al., 2024; Zhang et al., 2025	These cases primarily involve failures of factual representation rather than perceptual experience.
Knowledge recall failures	Zheng et al. (2023)	Since the error is attributed to failures of memorization and recall, it more closely resembles confabulation.
Unfounded fabrication	Sun et al., 2024	The model fills informational gaps with plausible but fabricated content, more closely aligned with confabulation.
Sycophancy	Wei et al. (2024); Zhang et al. (2025)	The model generates plausible responses that satisfy user expectations rather than factual constraints, resembling confabulatory narrative construction.
Semantic distortion	Sahoo et al. (2024)	Meaning is distorted despite surface coherence, resembling confabulatory reconstruction rather than perceptual error.

Another point worth emphasizing in this study is that hallucinations are not always regarded as undesirable outcomes. Indeed, some authors argue that "hallucinations must occur at a certain rate for language models that satisfy a statistical calibration condition appropriate for generative language models"[2] (Kalai & Vempala, 2023, p. 160). This observation makes it difficult to treat AI hallucination solely as a technical error. If certain types of hallucinations are an unavoidable consequence of successful predictive performance, then the issue is not only how such hallucinations can be managed, but also why they are characterized as "hallucinations" in the first place. The problem thus extends beyond a purely engineering concern and becomes a subject of conceptual and philosophical inquiry as well.

2. Hallucination in AI

Although text generation is today the phenomenon most commonly associated with AI hallucination, hallucinations may also occur in the generation of audio, images, and videos. Furthermore, as noted in the introduction, there is no general consensus regarding which phenomena should be classified as hallucinations. This section aims to provide the reader with a general understanding of the use of the term *hallucination* in the field of artificial intelligence by presenting a number of representative examples. One of the earliest sources we were able to identify is Tait's work, in which "the hallucination of matches" is described as "to see in the incoming text a text which fits its expectations, regardless of what the input text actually says." Hallucinated matches are said to underlie "the misassignment of a text to a script and the subsequent misanalysis of the text on the basis of that assignment" (Tait, 1982, p. 29). The same study emphasizes that, in order for a system to be more robust, it should be capable of indicating when it encounters difficulties. One of the elements notably absent in cases of hallucination is precisely this kind of feedback.

Setting Tait's work aside, early uses of the term *AI hallucination* were generally associated with a positive contribution rather than a negative one. Baker and Kanade provide one of the earliest and most frequently cited examples of the term's use in AI, as well as one of the clearest illustrations of this positive perspective:

> We propose an algorithm to learn a prior on the spatial distribution of the image gradient for frontal images of faces. We proceed to show how such a prior can be incorporated into a resolution enhancement algorithm to yield 4-8 fold improvements in resolution (i.e. 16-64 times as many pixels). The additional pixels are, in effect, hallucinated. (Baker & Kanade, 2000, p. 83)

[2] Three limitations of this study should be noted: first, it considers only one statistical source of hallucination; second, it relies on a specific concept of semantic calibration; and third, factuality lacks clear and decisive boundaries (Kalai & Vempala, 2023, p. 168).

In the AI literature, the concept of hallucination was therefore not initially employed solely as a category of negative error, but also as a means of productively completing missing information. More recent studies have similarly used the term in a positive sense in areas such as *image synthesis* (Pumarola et al., 2018) and *image inpainting* (Xiang et al., 2023). On the other hand, before the term *hallucination* became widespread, a related discussion was often conducted under the heading of *inconsistency* (Ji et al., 2023, p. 18). This suggests that, even if not under the specific label of hallucination, the underlying concern originally emerged in connection with undesirable or problematic system behavior.

The clearest definitions of AI hallucination can be found in the field of Natural Language Generation (NLG). For example, hallucination has been defined as "outputs that deviate from users' intent, exhibit internal inconsistencies, or misaligned with the real-world knowledge"[3] (Liu et al., 2026, p. 1037), "plausible yet nonfactual content" (Huang et al., 2025, p. 1), "the phenomenon where artificial intelligence generates distorted information" (Sun et al., 2024, p. 2), outputs that are "nonsensical or unfaithful to the provided source content" (Ji et al., 2023, p. 3), "fluent but unsupported text" (Filippova, 2020, p. 864)[4], or content that is "unfaithful to the input document" (Maynez et al., 2020, p. 1906). Considered together, these definitions suggest that hallucination is not used in the literature to refer to a single phenomenon, but rather functions as an umbrella term encompassing a variety of distinct problems, including factual accuracy, faithfulness, contextual appropriateness, inferential success, and alignment with user intent. Among these definitions, *unfaithfulness*[5] emerges as a particularly prominent criterion. In many cases, the only way to identify a hallucination is presented as subjecting the output to some form of fact-checking; however, this may not always be possible. Furthermore, some studies regard precisely those domains in which fidelity to factuality is most difficult to assess as the primary sites in which hallucinations arise:

> Hallucination: a response's factual correctness cannot be fully verified from the knowledge-snippet (even if it is true in the real world). More specifically, personal opinions, experiences, feelings, internal assessments of reality that cannot be attributed to the information present in the source document, are considered hallucinations. (Dziri et al., 2022, p. 5272)

For a broader and more comprehensive classification, many surveys rely on the taxonomy proposed by Ji et al. This classification distinguishes between intrinsic and extrinsic hallucinations: "Intrinsic hallucinations: The generated output that contradicts the source content. [...] Extrinsic hallucinations: The generated output that cannot be verified from the source content (i.e., output can neither be supported nor contradicted by the source)." (Ji et al., 2023, p. 3) In the same study, Ji et al. group the causes of hallucination into two broad categories: factors originating from the data and factors arising from the training and inference processes (Ji et al., 2023, p. 5). In the latter category, hallucinations may emerge even when the data itself is unproblematic; indeed, they may sometimes result from deliberate design choices intended to increase randomness and diversity in the model's outputs. For this reason, the objective is not necessarily to eliminate hallucinations altogether, but rather to establish an appropriate balance: "Controllability means the ability of models to control the level of hallucination and strike a balance between faithfulness and diversity." (Ji et al., 2023, p. 15) Like the work of Kalai and Vempala, this study therefore points to situations in which hallucinations are regarded as necessary or even desirable components of system performance.

In the context of chatbots, hallucinations are generally discussed as a problem concerning the reliability of generated content. According to a study surveying publications on AI hallucination produced between 2013 and 2023, academia constitutes the domain in which the issue has received the greatest attention, and the primary concern within this domain is a lack of fidelity to factual accuracy (Maleki, Padmanabhan, & Dutta, 2024, p. 137). Thus, depending on the context in which the term is employed,

[3] "The real-world knowledge refers to publicly available and verifiable information derived from human experience or natural activities [...] (*e.g.*, mathematical laws)" (Liu et al., 2026, p. 1039).

[4] It is also worth noting that Filippova employs the terms *noise* and *hallucination* interchangeably (Filippova, 2020, p. 865).

[5] "Faithfulness is defined as staying consistent and truthful to the provided source -an antonym to 'hallucination'" (Ji vd., 2023, s. 5). Likewise, no clear consensus appears to exist regarding the meaning or scope of factuality.

hallucinations may be regarded either as a desirable feature or as a problem to be addressed. In another example, an erroneous response generated by an LLM as a result of misunderstanding the user's intent is also classified as a form of hallucination (Zhang et al., 2025, p. 1377). Consequently, the detection and mitigation of hallucinations are often viewed as important steps in the development and improvement of such systems.

A similar dynamic can be observed in the recently prominent problem of *object hallucination*, the resolution of which is expected to improve "models' generalization capabilities" (Biten et al., 2022, p. 2473). Object hallucination refers to situations in which an object is incorrectly generated or identified within an image (Rohrbach et al., 2018). In this respect, object hallucination may be considered one of the closest analogues to the psychological use of the term *hallucination*:

> For example, imagine that a malicious agent has designed a patch to trigger hallucination of the car class and has placed it on the road. A self-driving car with a camera and detector model would see the patch, mistake it for a car, and suddenly stop or swerve to avoid a hallucinated collision. Furthermore, an untrained human observer would see the patch on the road but would not understand why the vehicle started to behave erratically. (Braunegg et al., 2020, p. 39)

The above example can easily be adapted to a human who panics after mistaking a graffiti drawing for a three-dimensional object because of an optical illusion. Similarly, cardboard police cars placed alongside roads can produce comparable effects on drivers. However, both in this example and in the scenario described by Braunegg et al., the phenomenon in question differs from what psychology designates as a hallucination. If psychology is taken as the starting point for terminology, it becomes clear that a reassessment is necessary. Indeed, Sun et al. note that the concept of AI hallucination has been adopted by some researchers to describe irrational or reality-incongruent outputs generated by AI systems, while at the same time attracting considerable criticism and dissatisfaction from others (2024, p. 2). These objections and alternative proposals will be examined in greater detail in the following section. In addition, this study will examine why the term *hallucination* has gained widespread acceptance despite the considerable dissatisfaction surrounding its use.

3. Psychological Roots of AI Hallucination

Newell and Simon's 1961 article on the General Problem Solver (GPS) begins with the statement: "This paper is concerned with the psychology of human thinking" (p. 109). The authors continue by arguing that they adopt a contemporary psychological approach situated between the poles of behaviorism and Gestalt psychology (p. 110). Their work not only draws upon psychology but also explicitly claims to contribute to the construction of a psychological theory (p. 113). From this perspective, it is not surprising that the term *hallucination* was borrowed from psychology. However, it should not be forgotten that the relationship between the two uses amounts only to "a loose resemblance" (Huang et al., 2025, p. 5). AI hallucination does not carry the same implications as the concept of hallucination when applied to human beings. Moreover, in the field of artificial intelligence, hallucination does not always refer to perception[6]. Huang et al. (2025) demonstrate that large vision-language models (LVLMs) may arrive at erroneous conclusions even when they correctly identify all visual components of an image (p. 34). This observation helps explain why the phenomenon referred to as AI hallucination does not fully correspond to the psychological concept of hallucination. Whereas hallucination in humans primarily refers to a perceptual experience, the same term in AI systems is often used to describe faulty inference, incorrect associations, or defective reasoning processes.

The first systematic definition of hallucination in its modern psychiatric sense was provided by Esquirol in 1845, who described it as the "perception of a sensation, when no external object" was present

[6] From the perspectives of psychology and psychiatry, hallucination is generally understood as a perceptual phenomenon. As Blom notes, hallucination is "*also known as hallucinatio, allucinatio, alucinatio, alusia, fallacia, idolum, illusio sensus, imagen, imago, phantasia, phantasma, somniatio morbosa,* and subjective sensation. Hallucination can be defined as a percept, experienced by a waking individual in the absence of an appropriate stimulus from the extracorporeal world (or, in the cases of somatic hallucinations, in the absence of an appropriate stimulus from within the body)" (2023, s. 318).

(Koçoğlu, 2022, p. 4; Esquirol, 1845, p. 93). The American Psychological Association defines hallucination as follows: "a false sensory perception that has a compelling sense of reality despite the absence of an external stimulus. [...] It is important to distinguish hallucinations from illusions, which are misinterpretations of real sensory stimuli" (American Psychological Association, 2018b).

Because hallucinations are generally understood as perceptual disturbances, some researchers argue that LLMs, lacking sensory perception, cannot perceive something that is not there and therefore their erroneous outputs should be described not as *hallucinations* but as *confabulations* (Smith, Greaves, & Panch, 2023). The same study warns that the use of the term *hallucination* risks anthropomorphizing machines and may encourage the attribution of consciousness to them (pp. 1-2). Nevertheless, it can be argued that both terms may lead to anthropomorphic analogies (Sui et al., 2024, p. 14275). On the other hand, the concept of *confabulation* appears to offer greater possibilities for drawing analogies between humans and machines, particularly when such analogies are grounded in memory processes (O'Hagan, Keane, & Flynn, 2025, pp. 1-2). In the work of Farquhar et al., confabulation in the context of AI is defined as the production of content that is "wrong and arbitrary"[7] and is classified as a subset of hallucination (Farquhar et al., 2024, p. 625).

The American Psychological Association defines *confabulation* as follows:

> The falsification of memory in which gaps in recall are filled by fabrications that the individual accepts as fact. It is not typically considered to be a conscious attempt to deceive others. [...] In forensic contexts, eyewitnesses may resort to confabulation if they feel pressured to recall more information than they can remember. (American Psychological Association, 2018a)

These definitions demonstrate that, at least within psychology, the two terms refer to different phenomena. Whereas hallucination is associated with perception, confabulation is generally understood as a disturbance related to memory. As reported by Huang et al., "By directly eliciting LLMs to generate a coherent counter-memory that factually conflicts with the parametric memory, LLMs exhibit their high receptivity to external evidence" (Huang et al., 2025, pp. 31-32). This finding likewise suggests that memory plays a significant role in the mitigation of errors. As another example, Liu et al. define hallucination as "a phenomenon where an explicit semantic conflict between the generated code and established facts due to the failure of the LLM to retain, recall, or process information accurately from users' requirements, contextual code, or real-world knowledge" (Liu et al., 2026, p. 1039). When specified in this manner, however, the phenomenon appears to be directly related to memory. Nevertheless, the same study acknowledges that, due to the *black-box nature* of LLMs, only their outputs can be examined and discussed (Liu et al., 2026, p. 1039). In another study, AI hallucination is described, in a distinctly anthropomorphic manner, as *pathological* (Lee et al., 2019, p. 1). Yet another example concerns *delusions*, which are defined as "situations in which the agent mistakes its own actions for evidence about the task" (Ortega et al., 2021, p. 14).

Why, then, should a term borrowed from psychology be used to describe such phenomena? This question remains open. Alternatives to *hallucination* have been proposed, including terms such as *fabrications* and *falsifications* (Emsley, 2023). Moreover, instead of employing concepts derived from psychology, such as *hallucination*, *confabulation*, or *delusion*, some authors have suggested concepts rooted in philosophy and logic, including certain forms of fallacious reasoning (Østergaard & Nielbo, 2023, p. 1107). Other non-psychological alternatives include *fabrication*, *stochastic parroting*, and *hasty generalization* (Maleki, Padmanabhan, & Dutta, 2024, p. 136).

Since the foundational texts of artificial intelligence, the field has maintained a close relationship with psychology. It is equally evident that this relationship continues to manifest itself strongly at the level of terminology in contemporary AI research. For example, according to Huang et al., "The essence of LLMs lies in predicting the probability distribution for upcoming words. Accurate predictions indicate a profound grasp of knowledge, translating to a nuanced understanding of the world" (Huang et al., 2025, p. 5). Another passage in which the authors speak of AI in a manner reminiscent of discussions of human

[7] "We mean that the answer is sensitive to irrelevant details such as random seed" (Farquhar et al., 2024, p. 625).

beings states that "whether we can effectively probe LLMs' internal beliefs is ongoing" (Huang et al., 2025, p. 34). According to another study, "The deficiency of generative artificial intelligence models in semantic understanding and logical inference, along with the cognitive limitations of the real world, are the primary factors contributing to reasoning errors" (Sun et al., 2024, p. 9). In the same study, AI systems are evaluated in comparison with "normal human logical thinking" (Sun et al., 2024, p. 12). Across all of these examples, a vocabulary carrying associations traditionally reserved for the human mind is employed. The predominance of psychologically derived terminology in AI is therefore not merely a historical legacy; it is also a consequence of the continued application of conceptual frameworks originally developed to explain the human mind to artificial agents.

According to Maleki et al., summaries of AI hallucination research "highlight different extents of inaccuracy, ranging from 'deviating from established knowledge', 'factual incorrectness', 'fictional' to 'nonsensical'" (Maleki, Padmanabhan, & Dutta, 2024, p. 136). For this reason, careful attention to these distinctions is essential if an appropriate terminology is to be established. The next section of this study will examine this issue through the problem of *understanding*, a topic that occupies a central place in the philosophy of artificial intelligence.

4. Meaning, Mind, and AI Hallucination

According to Newell and Simon, "Given enough information about an individual, a program could be written that would describe the symbolic behavior of that individual" (1961, p. 114). These early studies tended to treat human beings as entities whose behavior could be symbolized and computed. In this context, some of the most intense debates have concerned human inner experience. For example, in his well-known discussion of conscious experience, Nagel argued that neither a functionalist perspective nor a mere analysis of the concept of experience is sufficient to determine whether a machine could possess consciousness in the same way that humans do (Nagel, 1974, p. 436). If what is promised is merely the imitation of 'symbolic behavior,' then the resulting product cannot be regarded as equivalent to a human being.

On the other hand, McCarthy considered it reasonable, useful, and even necessary to attribute mental properties such as beliefs, intentions, and desires to machines (1979, p. 1). According to him, when forming a mental representation of a present perception, a human being completes that representation by drawing upon previous representations; consequently, a similar mechanism is necessary for machines as well. These studies represent early attempts to construct a computable psychology. For one group of researchers, machines are developed by modeling human beings, and as machines become increasingly sophisticated, human-level complexity will become explainable through them.

On another side of the debate, some scholars have argued that computational complexity theory is also a matter of philosophical concern (Šekrst, 2025; Aaronson, 2011). According to some philosophers, contrary to the claims of strong-AI advocates, "complexity alone does not necessarily bridge the gap between syntax and semantics" (Šekrst, 2025, p. 77). Searle, whose work is particularly relevant to the present discussion, argues that "the formal symbol manipulations by themselves don't have any intentionality […] In the linguistic jargon, they have only a syntax but no semantics" (1980, p. 422). Rather than providing a formal definition of what understanding is, Searle maintains that examples of understanding and non-understanding are sufficient to clarify the distinction: "I understand stories in English; to a lesser degree I can understand stories in French; to a still lesser degree, stories in German; and in Chinese, not at all. My car and my adding machine, on the other hand, understand nothing: they are not in that line of business" (Searle, 1980, p. 419). We sometimes speak of machines as perceiving, understanding, or knowing, yet according to Searle such expressions merely reflect "the fact that in artifacts we extend our own intentionality" (p. 419). From this perspective, the term *stochastic parrot* appears to be an appropriate choice for researchers who take Searle's criticism seriously: "Contrary to how it may seem when we observe its output, an LM is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: a stochastic parrot" (Bender

et al., 2021, pp. 616–617). The notion of the stochastic parrot relieves us of the need to attribute *understanding* to artificial intelligence. Everything is reduced to computation. From this perspective, hallucinations may simply be regarded as computational errors. According to another view, however, code bugs and code hallucinations are not the same thing, and hallucinations need not always constitute errors. In order for an output to count as a hallucination, there must be a *semantic conflict* (Liu et al., 2026, pp. 1039–1040). In this account, a hallucination is characterized not by a syntactic failure but by a semantic one. Moreover, within the literature, *hallucination mitigation* and *semantic understanding* are frequently treated as parallel goals. The notion of *semantic* employed in this context, however, does not correspond to *semantics* in Searle's sense. What is meant here is a structure that lacks a valid correspondence within the represented system. In other words, AI literature generally uses *semantics* to refer to a technical representational relation, whereas Searle employs the concept in connection with intentionality and understanding.

In Searle's work, we find not so much an account of what counts as meaning or understanding, but rather an account of what does not. His aim is to clarify the distinction by providing a clear example not of human understanding, but of machine-like understanding: "In the Chinese case, unlike the English case, I produce the answers by manipulating uninterpreted formal symbols. As far as the Chinese is concerned, I simply behave like a computer; I perform computational operations on formally specified elements" (Searle, 1980, p. 418). From Searle's perspective, expressions such as *lies* or *believes*, which are often applied to LLMs, imply the mistaken attribution of mental states that such systems do not, at least for the time being, possess. Hallucination may likewise be regarded as an example of this type of misattribution (Šekrst, 2025, p. 70). On this view, AI is in reality doing nothing more than carrying out computational operations. At the same time, Searle maintains that "if we could build a robot whose behavior was indistinguishable over a large range from human behavior, we would attribute intentionality[8] to it, pending some reason not to" (Searle, 1980, p. 421). According to Searle, unlike machines, the human mind possesses intentionality, since "it couldn't be just computational processes and their output because the computational processes and their output can exist without the cognitive state" (1980, p. 422). At this point, one may ask how the inferential capacities of machines should be understood and what their counterpart would be in human beings. Searle's answer would, of course, be that they are nothing more than computations. Moreover, all of Searle's criticisms were directed at a very specific conception of artificial intelligence:

> The interest of the original claim made on behalf of artificial intelligence is that it was a precise, well defined thesis: mental processes are computational processes over formally defined elements. I have been concerned to challenge that thesis. If the claim is redefined so that it is no longer that thesis, my objections no longer apply because there is no longer a testable hypothesis for them to apply to. (Searle, 1980, p. 422)

This raises the question of whether Searle's thesis remains fully applicable today. With the emergence of RAG technology, AI systems are no longer limited to recalling internal knowledge; they are also capable of retrieving and integrating external information in contextually appropriate ways. In principle, RAG may offer a solution to what Newell and Simon regarded as a limitation of GPS. Their proposed solution was "to provide GPS with a little continuous hindsight about its past actions" (Newell & Simon, 1961, p. 122). From Searle's perspective, however, and especially in light of his criticisms of McCarthy, RAG would likely amount to nothing more than adding a larger library to the Chinese Room.

McCarthy took a considerably more radical position: "Machines as simple as thermostats can be said to have beliefs, and having beliefs seems to be a characteristic of most machines capable of problem solving performance" (McCarthy, 1979, p. 3). Searle objected to this view by arguing that the distinction between the mental and the nonmental is independent of the observer's perspective: "otherwise it would be up to any beholder to treat people as nonmental and, for example, hurricanes as mental if he likes" (Searle, 1980, p. 420). Searle does not pursue the debate much further, apparently considering it unnecessary to defend the claim that human beings possess beliefs whereas thermostats do not. Šekrst,

[8] Here, the term intentionality may be understood in a general sense as 'being about something' or 'being directed toward something'.

however, suggests a possible response to Searle: "if anything can be understood as a 'machine' (which Searle contends, in a way), then the distinction between 'mental' and 'nonmental' systems is not as clear-cut as Searle assumes" (Šekrst, 2025, p. 74). Similarly, Marvin Minsky's virtual mind reply invites us to consider the point at which the distinction between a genuinely conscious mind and a virtual mind capable of perfectly simulating reality might cease to matter (Šekrst, 2025, p. 75). The present study concludes by leaving open a question for future research: if we do not place the real mind beyond reach from the outset, on what grounds can we claim that a virtual mind will never be capable of closing the gap between them?

On the other hand, Searle acknowledges the possibility of producing mentality "using some other sorts of chemical principles than those that human beings use," but sets this possibility aside on the grounds that it constitutes an empirical question (1980, p. 422). Consequently, this may be interpreted as leaving open, at least in principle, the possibility of a virtual mind, provided that such a mind is capable of going beyond mere computation.

As the result of a thought experiment, Minsky proposed the "amusing prediction that intelligent machines may be reluctant to believe that they are just machines"[9] (1961, p. 28). The argument underlying this prediction is that human beings possess a dual model of the self -both physical and mental- and that, unless we develop a satisfactory theory of artificial intelligence, machines themselves may continue to reproduce this dualistic perspective, claiming, when asked, that they too consist of a mind and a body. This observation directs us toward a point that remains difficult to account for: the ontological assumptions underlying our judgments about artificial intelligence. If the implications of this thought experiment cannot be adequately addressed, we may also be required to reconsider the terminology through which we describe our own inner experiences.

The central point emphasized by Nagel in *What Is It Like to Be a Bat* concerns our lack of understanding regarding the physical explanation of mental states: "I shall try to explain why the usual examples do not help us to understand the relation between mind and body -why, indeed, we have at present no conception of what an explanation of the physical nature of a mental phenomenon would be" (Nagel, 1974, pp. 435-436). Nagel further characterizes physicalism itself as a position that we do not yet fully understand (1974, p. 446). As far as one can observe, the views that have shaped both mainstream discussions and common sense have generally been those that attribute a privileged status to mentality. The human mind occupies a privileged position by virtue of its inner experience and its access to that experience. As Nagel writes: "Without some idea, therefore of what the subjective character of experience is, we cannot know what is required of a physicalist theory" (Nagel, 1974, p. 437). If we are still unable to explain what subjective experience is and how it arises, then it becomes equally controversial under which mental categories certain behaviors observed in AI systems should be classified. Consequently, concepts attributed to AI -such as *hallucination*, *belief*, *understanding*, or *intentionality*- cease to function merely as descriptions of system behavior and begin to reflect our assumptions about the nature of the human mind itself.

From this perspective, the concept of AI hallucination may also be understood as a consequence of extending our understanding of our own mental life to artificial agents. Yet the very notion of mentality upon which this extension rests remains far from fully clarified. As Bailey notes, although we may investigate how different perceptual systems give rise to different forms of experience, we still do not know why or how particular physical processes produce particular qualitative experiences (2019, p. 17). We are therefore left with an unresolved problem that concerns not only artificial intelligence but also the human mind itself: to what extent can we speak reliably about mental states that are not directly observable? This question leads us back to Berkeley, who drew attention to a similar difficulty long before contemporary debates about artificial intelligence. In his discussion of abstract ideas and mental content, Berkeley emphasizes that we cannot directly know what mental faculties other people genuinely

[9] A similar narrative is skillfully developed in the work of Asimov. The idea of a robot that neither accepts nor even entertains the possibility that it was created by humans is explored in the classical science-fiction work *I, Robot* (Asimov, 2022).

possess: “Whether others have this wonderful Faculty of Abstracting their Ideas, they best can tell” (Berkeley, 1710/2002, p. 3).

5. Conclusion

There is no consensus regarding the definition of AI hallucination; moreover, the concept risks expanding to a degree that would encompass virtually all AI-generated outputs. In this process, the term *hallucination* has moved away from its established meaning in psychology. Furthermore, certain types of errors observed in LLMs appear to be more closely related to the concept of *confabulation* than to that of hallucination. At the same time, no common ontological foundation exists for the various attempts to define the phenomenon. Nevertheless, it may be argued that a dualistic perspective, even if only implicitly, continues to exert a significant influence on the language employed in these discussions.

At present, it does not seem possible to examine either a human being or a machine down to its smallest structural components. Behavioral analysis also has its limitations -for example, not every mental state necessarily manifests itself in observable behavior. If we are to speak about such matters at all, we are left with the option of attributing certain properties to both humans and machines. Within the framework of this study, inner experience appears to be the most significant feature distinguishing human beings from machines. Furthermore, no clear and generally accepted principled distinction has been established between attributing mentality to particular parts of the brain and attributing mentality to components of a machine. If we do not know at what point mentality emerges, it may also be argued that we cannot determine with certainty at what point it does not emerge. Our attribution of mentality to other human beings often rests on perceived similarities between their experiences and our own subjective experience. Yet there is no generally accepted criterion specifying under what conditions such similarities should be regarded as sufficient grounds for attributing mentality.

The fact that machines have developed to a level at which they may pass the Turing Test did not lead researchers to approach them using the terminology of psychology; rather, the pioneers of artificial intelligence had already set out to construct a psychology of intelligence from the outset. For this reason, it is entirely understandable that many of the concepts that emerged during the first fifty years of the formal development of AI were derived from psychology. Moreover, although newer approaches such as RAG have altered the ways in which AI systems access and utilize information, they have not resolved the fundamental philosophical questions surrounding meaning, mentality, and experience. The debate over AI hallucination should therefore be understood not merely as a matter of technical error classification, but also as part of a broader philosophical discussion concerning the nature of mentality.

Statements and Declarations

Ethical Considerations

Not applicable.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding Statement

The author received no financial support for the research, authorship, and/or publication of this article.

Data availability

Not applicable.

References

Aaronson, S. (2013). Why philosophers should care about computational complexity. In B. J. Copeland, C. Posy, & O. Shagrir (Eds.), *Computability: Turing, Gödel, Church, and beyond* (pp. 261–327). MIT Press. https://doi.org/10.7551/mitpress/8009.003.0011

American Psychological Association. (2018a). *Confabulation*. APA Dictionary of Psychology. 11 Aralık 2025 tarihinde erişildi, https://dictionary.apa.org/confabulation

American Psychological Association. (2018b). *Hallucination*. APA Dictionary of Psychology. 11 Aralık 2025 tarihinde erişildi, https://dictionary.apa.org/hallucination

Asimov, I. (2022). *Ben, robot* (E. Odabaş, Trans.). İthaki Yayınları.

Bailey, A. (Ed.). (2019). *Zihin felsefesi* (F. Doruker, Trans.). FOL Kitap.

Baker, S., & Kanade, T. (1999, September). *Hallucinating faces* (Technical Report No. CMU-RI-TR-99-32). The Robotics Institute, Carnegie Mellon University.

Baker, S., & Kanade, T. (2000). *Hallucinating faces*. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)* (pp. 83–88). IEEE. https://doi.org/10.1109/AFGR.2000.840616

Barnett, S., Kwon, S., Shankar, V., & diğer yazarlar. (2024). *Seven failure points when engineering a retrieval augmented generation system*. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering: Software Engineering for AI (CAIN 2024)* (pp. 194–199). Association for Computing Machinery. https://doi.org/10.1145/3644815.3644945

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 Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Berkeley, G. (1710/2002). *A treatise concerning the principles of human knowledge* (D. R. Wilkins, Ed.). Trinity College Dublin. Retrieved May 27, 2026, from https://www.maths.tcd.ie/~dwilkins/Berkeley/HumanKnowledge/1734/HumKno.pdf

Biten, A. F., Gomez, L., & Karatzas, D. (2022). Let there be a clock on the beach: Reducing object hallucination in image captioning. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 2473–2482).

Blom, J. D. (2023). *A dictionary of hallucinations* (2nd ed.). Springer.

Braunegg, A., Chakraborty, A., Krumdick, M., Lape, N., Leary, S., Manville, K., Merkhofer, E., Strickhart, L., & Walmer, M. (2020). APRICOT: A dataset of physical adversarial attacks on object detection. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (Lecture Notes in Computer Science, Vol. 12366, pp. 35–50). Springer. https://doi.org/10.1007/978-3-030-58589-1_3

Cambridge Dictionary. (2023). *Word of the Year 2023*. 11 Aralık 2025 tarihinde erişildi, https://dictionary.cambridge.org/editorial/word-of-the-year/2023

Chowdhury, T. (2026). Retrieval-Augmented Generation (RAG) for Large Language Models: A Comprehensive Survey. *International Journal on Engineering Artificial Intelligence Management, Decision Support, and Policies*, *3*(1), 09–21. https://doi.org/10.63503/j.ijaimd.2026.233

Dziri, N., Milton, S., Yu, M., Zaiane, O., & Reddy, S. (2022). On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 5271–5285). Association for Computational Linguistics.

Emsley, R. (2023). *ChatGPT: These are not hallucinations—they're fabrications and falsifications*. *Schizophrenia, 9*, Article 52. https://doi.org/10.1038/s41537-023-00379-4

Esquirol, E. (1845). *Mental maladies: A treatise on insanity* (E. K. Hunt, Trans.). Lea and Blanchard.

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature, 630, 625–630. https://doi.org/10.1038/s41586-024-07421-0

Filippova, K. (2020). Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 864–870). Association for Computational Linguistics.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *ArXiv, abs/2312.10997*.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). *A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions*. *ACM Transactions on Information Systems, 43*(2), Article 1, 1-55. https://doi.org/10.1145/3703155

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Chan, H. S., 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

Kalai, A., & Vempala, S. (2023). *Calibrated language models must hallucinate*. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing* (pp. 160-171). Association for Computing Machinery. https://doi.org/10.1145/3618260.3649777

Koçoğlu, M. (2022). *The relationship of hallucinations and psychotic markers* [Medical specialty thesis, Dokuz Eylül University, Faculty of Medicine, Department of Psychiatry].

Lee, K., Firat, O., Agarwal, A., Fannjiang, C., & Sussillo, D. (2019). *Hallucinations in neural machine translation* [Conference paper]. ICLR 2019 Conference Blind Submission.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). *Retrieval-augmented generation for knowledge-intensive NLP tasks*. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Article 793, pp. 9459–9474).

Liu, F., Liu, Y., Shi, L., Lian, X., Li, Z., Yang, Z., Zhang, L., & Ma, Y. (2026). *Beyond functional correctness: Exploring hallucinations in LLM-generated code*. *IEEE Transactions on Software Engineering, 52*(3), 1037-1055.

Maleki, N., Padmanabhan, B., & Dutta, K. (2024). AI hallucinations: A misnomer worth clarifying. In *2024 IEEE Conference on Artificial Intelligence (CAI)* (pp. 133–138). IEEE. https://doi.org/10.1109/CAI59869.2024.00033

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1906-1919). Association for Computational Linguistics.

McCarthy, J. (1979). *Ascribing mental qualities to machines.* Stanford University, Computer Science Department. Retrieved May 24, 2026, from https://www-formal.stanford.edu/jmc/ascribing.pdf

Merriam-Webster. (2023). *Definition of hallucination*. Retrieved May 19, 2026, from https://www.merriam-webster.com/dictionary/hallucination

Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE, 49*(1), 8–30. https://doi.org/10.1109/JRPROC.1961.287775

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review, 83*(4), 435–450. https://doi.org/10.2307/2183914

Newell, A., & Simon, H. A. (1961). GPS, a program that simulates human thought. In H. Billing (Ed.), *Lernende Automaten* (pp. 109-124). R. Oldenbourg.

O'Hagan, J., Keane, A., & Flynn, A. (2025). *Confabulation dynamics in a reservoir computer: Filling in the gaps with untrained attractors*. *Chaos, 35*(9), 093130. https://doi.org/10.1063/5.0283285

Ortega, P. A., Kunesch, M., Delétang, G., Genewein, T., Grau-Moya, J., Veness, J., Buchli, J., Degrave, J., Piot, B., Perolat, J., Everitt, T., Tallec, C., Parisotto, E., Erez, T., Chen, Y., Reed, S., Hutter, M., de Freitas, N., & Legg, S. (2021, October 22). *Shaking the foundations: Delusions in sequence models for interaction and control* (Technical report). DeepMind Safety Analysis.

Østergaard, S. D., & Nielbo, K. L. (2023). False responses from artificial intelligence models are not hallucinations. *Schizophrenia Bulletin, 49*(5), 1105–1107. https://doi.org/10.1093/schbul/sbad068

Pumarola, A., Agudo, A., Sanfeliu, A., & Moreno-Noguer, F. (2018). Unsupervised person image synthesis in arbitrary poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8620–8628).

Ren, R., Wang, Y., Qu, Y., Zhao, W. X., Liu, J., Tian, H., Wu, H., Wen, J.-R., & Wang, H. (2025). *Investigating the factual knowledge boundary of large language models with retrieval augmentation*. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 3697–3715). Association for Computational Linguistics.

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., & Saenko, K. (2018). Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4035–4045). Association for Computational Linguistics.

Sahoo, P., Meharia, P., Ghosh, A., Saha, S., Jain, V., & Chadha, A. (2024). A comprehensive survey of hallucination in large language, image, video and audio foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 11709–11724). Association for Computational Linguistics.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–457.

Šekrst, K. (2025). *The illusion engine: The quest for machine consciousness*. Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-05562-0

Smith, A. L., Greaves, F., & Panch, T. (2023). *Hallucination or confabulation? Neuroanatomy as metaphor in large language models*. *PLOS Digital Health, 2*(11), e0000388. https://doi.org/10.1371/journal.pdig.0000388

Sui, P., Duede, E., Wu, S., & So, R. (2024). *Confabulation: The surprising value of large language model hallucinations*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 14274–14284). Association for Computational Linguistics.

Sun, Y., Sheng, D., Zhou, Z., & Wu, Y. (2024). *AI hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content*. *Humanities and Social Sciences Communications, 11*, Article 1278. https://doi.org/10.1057/s41599-024-03811-x

Tait, J. I. (1982). *Automatic summarising of English texts* (Technical Report No. 47, UCAM-CL-TR-47). Computer Laboratory, University of Cambridge.

Wang, Y., Pan, Y., Yan, M., Su, Z., & Luan, T. (2023). *A survey on ChatGPT: AI-generated contents, challenges, and solutions*. *IEEE Open Journal of the Computer Society, 4*, 280–302. https://doi.org/10.1109/OJCS.2023.3300321

Wei, J., Huang, D., Lu, Y., Zhou, D., & Le, Q. V. (2024). *Simple synthetic data reduces sycophancy in large language models* [Google DeepMind]. *arXiv*. https://arxiv.org/abs/2308.03958

Xiang, H., Zou, Q., Nawaz, M. A., Huang, X., Zhang, F., & Yu, H. (2023). Deep learning for image inpainting: A survey. *Pattern Recognition, 134*, 109046.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2025). Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics, 51*(4), 1373–1418. https://doi.org/10.1162/coli.a.16

Zheng, S., Huang, J., & Chang, K. C. (2023). *Why does ChatGPT fall short in answering questions faithfully? arXiv*. https://arxiv.org/abs/2304.10513