DOI: 10.2478/bile-2025-0010

Biometrical Letters
Vol. 62 (2025), No. 2, 145-188

Artificial Intelligence and Consciousness: Limits and Modern Perspectives

Laura Slebioda

Department of Mathematical and Statistical Methods,
Poznań University of Life Sciences, Wojska Polskiego 28, 60-637 Poznań, Poland,
e:mail laura.slebioda@up.poznan.pl

SUMMARY

This paper provides a review of selected concepts concerning consciousness, intelligence and artificial intelligence, focusing on their interrelations and interpretative limitations. The aim of the paper is to organize key definitions and viewpoints, and to highlight central issues related to the question of whether conscious machines can ever emerge. Consciousness is often defined as subjective experience, as the capacity for reflection on one's own mental states, or as an emergent property of complex biological systems. Intelligence, on the other hand, is interpreted as the ability to learn, solve problems, adapt to changing conditions, and control cognitive processes. The development of computational technologies has given rise to weak artificial intelligence, encompassing algorithmic and machine learning systems that can model and predict patterns with high precision. Within this category, generative artificial intelligence, represented by large language models, demonstrates impressive linguistic capabilities but lacks genuine understanding – a feature associated with strong AI. The paper discusses whether computational processes can be equated with real thinking, referring to Gödel's incompleteness theorems, Searle's Chinese Room argument, as well as the Turing Test. This review contributes by integrating classical philosophical arguments with a comparative evaluation of contemporary language models (GPT-5, Gemini 2.5, DeepSeek-V3.2), examining their responses to Gödelian questions and reasoning tasks. The analysis indicates that, despite significant progress in building artificial intelligence systems, the question of their potential consciousness remains unresolved and continues to be a subject of profound philosophical debate.

Key words: artificial intelligence, consciousness, Turing Test, Chinese Room

1. Introduction

The concepts of consciousness and artificial intelligence are among the most fascinating and at the same time controversial topics of contemporary science.

For centuries, scholars have attempted to understand the nature of the human mind, what intelligence is, and whether it is possible to replicate it in machines. The answers to these questions are far from unambiguous, and proposed concepts differ not only in terms of theoretical approach, but also in their underlying assumptions.

Consciousness has been described as an immaterial soul [Popper and Eccles, 1977], as an illusion generated by the brain [Dennett, 2016], as an emergent property of neural networks [Crick and Koch, 1990], and as an effect of interaction between an organism and its environment [O'Regan and Noë, 2001]. Intelligence, from the first attempts at its measurement by Binet and Simon [Varon, 1936] up to contemporary psychological definitions, has been defined as the capacity for learning, reasoning, and adaptation [Neisser et al., 1996; Strelau, 1997; Nęcka, 2003].

Artificial intelligence (AI), initially treated as an engineering discipline, has evolved into a research area encompassing both practical and philosophical dimensions. A distinction is commonly made between weak AI, which includes algorithmic and data-driven systems such as those based on machine learning (ML), and strong AI, referring to the hypothetical emergence of machines possessing self-awareness or consciousness. Within weak AI, a particularly visible branch is generative AI, represented by large language models (LLMs), which generate text and other outputs by modeling statistical patterns in data. The rise of such models has revived classical debates concerning the nature of thought, understanding and consciousness in artificial systems.

The aim of this paper is to present selected concepts of consciousness, intelligence and artificial intelligence, as well as to analyze the main arguments and objections concerning the idea of machines endowed with a conscious, intelligent mind. The following sections discuss various definitions of consciousness, approaches to understanding intelligence, characteristics of artificial intelligence, and debates surrounding the Turing Test, the Chinese Room argument and Gödel's theorems. The paper also examines the relevance of these debates in the context of modern AI models. This comprehensive overview

highlights the complexity of the issue and helps explain why the question of whether machines can truly think and be conscious remains unresolved.

2. Consciousness

It is not easy to answer the question of what consciousness is. Many philosophers and scientists have attempted to explain it, but for each of them it has meant something slightly different, and therefore they have approached the problem from diverse perspectives.

Popper and Eccles [1977] argued that consciousness is related to the immaterial soul. The authors explicitly state that the self-aware mind is an independent entity, and not merely a function of the brain. Consciousness is presented as a superior factor that controls neuronal events. Since it acts upon the brain but is not itself part of its mechanics, this means that there exists a non-material entity that influences matter. Consciousness exists outside the brain itself, yet comes into contact with it.

McGinn [2004] approached the problem differently, through mysterianism. He argued that consciousness cannot be understood by applying any scientific method, because consciousness is a mystery and is not accessible to us. This does not mean that consciousness does not exist, but rather that humans are unable to understand scientifically how it connects to the brain. Nagel [1974] took a similar stance, giving the example of a bat. A person who possessed extensive knowledge about bats, fully understanding and knowing in detail how they function, would still be unable to answer the question of what it is like to be a bat. A human can imagine what it might feel like to be a bat, but this is not the same as actually being a bat. For this reason, Nagel claimed that no theory is capable of explaining the phenomenon of consciousness.

Dennett [2016], on the other hand, considered consciousness to be an illusion. In stage magic, the audience may be absolutely convinced that what they see (for example, a woman sawn in half) cannot be explained by ordinary physical means. Yet upon closer investigation it turns out to be only an illusion, created by clever

techniques rather than “real magic.” Thus, according to Dennett, consciousness is merely a “trick of nature,” an illusion generated by neuronal mechanisms.

It may be that consciousness requires the adoption of entirely new scientific laws. Hameroff and Penrose [1996] explicitly suggested that current physics and biology may not be sufficient to explain consciousness. It may be necessary to expand physics with new principles (e.g. related to quantum gravity and the non-computability of processes), since existing scientific frameworks are insufficient.

A completely different approach to consciousness was taken by O’Regan and Noë [2001], who linked it to behavior. Seeing is not a passive act of “watching a movie” inside the brain, but rather an active exploration of the world using movements of the eyes, head and body. Visual consciousness consists in the organism’s ability to understand and make use of regularities, such as what will happen in the field of vision if one moves the eyes, turns the head, or steps closer. It is knowledge of “if–then” relations between actions and changes in perception. The authors provide empirical evidence showing that behavior and the ability to explore determine what and how we consciously perceive. They cite, among others, sensorimotor adaptation. For example, after putting on glasses that invert the visual field, the organism begins to learn new relationships between movements and changes in the visual field.

Crick and Koch [1990], in turn, defined consciousness as an emergent property of specific biological systems, something that can be studied within the framework of neurobiology. They proposed that visual consciousness results from coordinated neuronal activity (gamma oscillations at 40–70 Hz), which binds features into a coherent representation of an object. It may arise as the effect of activity across an entire network of neurons. Consciousness is thus understood here as the result of interaction between attention, working memory, and neuronal oscillations.

The most frequently cited definition of consciousness is that of Searle [1997]. He claims that consciousness consists of those states associated with sensations, feelings, or wakefulness that begin in the morning when we wake from sleep and

continue until the moment we fall into a coma or die, or fall asleep again, or lose consciousness in any other way.

3. Intelligence

Defining intelligence is just as problematic as defining consciousness. In 1905, Binet and Simon developed the first practical scale for measuring intelligence, later known as the Binet–Simon Scale [Varon, 1936]. The tool was intended to help identify children with learning difficulties so that appropriate educational support could be provided. It was the first attempt at an objective measurement of intelligence in the form of a standardized test, and it later became the foundation for the further development of intelligence tests (e.g. the Wechsler Scales). For Binet, intelligence meant making accurate judgments, adequately understanding situations, and reasoning logically. He believed that intelligence was a fundamental function of the mind, related to directing and controlling thoughts and actions.

Spearman argued that intelligence is manifested primarily in those cognitive processes that generate new cognitive content on the basis of inductive and deductive reasoning [Lubinski, 2004]. In his view, general intelligence (*g*) occupies the apex of the hierarchical structure of cognitive abilities, determining the level of performance in many specific tasks. In addition, specific abilities beyond *g* enable more accurate predictions of important social phenomena and constitute a significant element of individuals' psychological diversity.

For Terman [1916], on the other hand, an individual is intelligent in proportion to their capacity for abstract thinking. He introduced the concepts of mental age and the intelligence quotient (IQ) as a way of measuring this trait, standardizing the results of the Binet–Simon test.

In 1994, controversies surrounding the book *The Bell Curve* sparked a debate in the scientific community [Herrnstein and Murray, 1994]. The authors of the book argued that intelligence has a significant impact on social, economic and educational achievement, and that there are differences in IQ between racial

groups. Many researchers considered these claims potentially racist or methodologically flawed, and as a result, the American Psychological Association (APA) decided to produce a report to clarify the scientific facts about intelligence. Eleven experts participated in preparing the report, and they defined intelligence as the ability to learn from one's own experiences, the ability to adapt to the surrounding environment, and the metacognitive ability to reflect upon one's own cognitive processes [Neisser et al., 1996].

A later definition was developed by Strelau [1997], who considered intelligence a theoretical construct referring to relatively stable internal conditions of a person that determine the effectiveness of performing tasks or solving problems requiring typically human cognitive processes. These conditions are shaped by the specific interaction between genotype and environment characteristic for each individual.

Nęcka [2003], in turn, decided to compare intelligence with temperament. He noted that these are different domains, but both are connected with human behavior. He believes that intelligence is the ability to adapt to circumstances by perceiving abstract relationships, using prior experiences, and effectively controlling one's own cognitive processes.

4. Artificial Intelligence

In practice, the term "artificial intelligence" is used in very different ways. On the one hand, it refers to classical solutions, applied for many decades and based on algorithms, merely because their operation resembles human activity (just as a ticket machine imitates a cashier at a ticket office) [Zalewski, 2020]. Such systems are described as weak artificial intelligence. On the other hand, the term is sometimes applied to hypothetical systems that do not yet exist and perhaps may never come into existence, for example, AI endowed with self-awareness. Conscious AI is called strong artificial intelligence, meaning that it would actually possess the capacity for thinking. Attempts to define AI often lead to the construction of complex logical formulations, from which no simple, applicable

concept for a definition emerges. The most frequently cited definition of artificial intelligence is the one inspired by Turing's ideas, according to which AI is the ability of a machine to imitate or simulate human intelligence.

A considerable number of researchers point out that what is today referred to as weak AI is largely represented by machine learning (ML) methods – systems that learn patterns from data rather than embodying genuine understanding. Within this group, generative AI, such as large language models (LLMs), constitutes a prominent subfield that produces text, images or other content by statistically modeling patterns in existing data. Strong AI, in contrast, is considered the true definition of artificial intelligence, which, according to many researchers, would require AI not only to be intelligent but also to be conscious [Li, 2018; Kawecki, 2025]. Accordingly, models such as ChatGPT or DeepSeek are not conscious entities but advanced probabilistic models that predict the most likely continuation of a given input sequence.

The field of artificial intelligence aims to recreate mentality within computational machines [Chalmers, 2010]. Proponents argue that we have every reason to believe that computers will eventually truly possess minds. At the same time, opponents maintain that computers, unlike humans, are limited in certain domains and therefore it is impossible for a conscious mind to emerge solely through computational capabilities.

5. Internal and External Objections

Numerous objections have been raised against the computational theory of mind, questioning its credibility. The criticism takes two forms: external objections and internal objections.

External objections refer to the inability of computational systems to behave in the same way as cognitive systems [Chalmers, 2010]. Humans possess certain functional abilities that no computer is capable of achieving. Computational machines operate on strictly defined rules, and therefore cannot demonstrate the creativity or flexibility characteristic of human beings. Moreover, some

philosophers have argued that computational systems are unable to reproduce human mathematical intuition, claiming that they are constrained by Gödel's theorem in a way that humans are not [Penrose, 1994]. However, this interpretation of Gödel's results remains debated, as many scholars reject the view that the theorems establish any fundamental asymmetry between human and machine reasoning [Davis, 2018].

Internal objections emphasize that even if computers could simulate human behavior, they would not possess a mind or an inner life [Chalmers, 2010]. They lack conscious experiences and genuine understanding. At best, they are capable of creating a simulation of mentality, but not an actual reproduction of it. Computational systems can therefore be regarded as empty shells of the mind, silicon versions of zombies.

5.1. Gödel's Theorem

One of the objections against conscious artificial intelligence focuses on Gödel's theorem. The first theorem concerns incompleteness, and states that a formal theory containing elementary arithmetic does not prove certain true arithmetic sentences [Smith, 2013]. This means that no program that produces only true sentences of elementary arithmetic can produce all of them.

Outline proof: Suppose such a program existed. For it, there would be an arithmetic sentence *G* such that *G* is equivalent to the statement "This program does not produce *G*." If our program produces *G*, then since it produces only true sentences, *G* is true. Therefore, the statement "This program does not produce *G*" is also true. Thus, it does not produce *G*. We have shown that if it produces *G*, then it does not produce it. Hence, it does not produce *G*. This means that *G* is true but not produced [Krajewski, 2021].

This construction is historically related to the intuition behind the Liar Paradox. Suppose someone says, "I am lying now." If we assume that the sentence is true, then what it states must be the case, yet it states that the person is lying, which makes it false [Brożek, 2002]. If, on the other hand, we assume it is false, it turns out it must be true. A contradiction always arises.

Gödel's second incompleteness theorem states that any sufficiently strong, effectively axiomatized, consistent formal system cannot prove its own consistency [Smith, 2013].

Outline proof: Let us assume a theory T (arbitrarily large, assumed to be consistent). The sentence expressing consistency *Cons* can be formulated. It can be shown that in T (Ar) (where Ar is elementary arithmetic), the sentence ($Cons \rightarrow G$) is provable. To demonstrate this, we formalize the proof of T in I, where I is weak arithmetic (e.g. $I\Sigma_1$-arithmetic or Robinson Arithmetic), which is capable of describing proofs in T. Therefore, if *Cons* were provable, so would G be, contrary to what has already been proven. Within a suitable base theory (e.g., $I\Sigma_1$), one shows that if T is consistent, then the Gödel sentence G for T is not provable in T [Krajewski, 2021].

The reasoning presented above shows that T cannot, if consistent, prove its own consistency, since such a proof would translate into a formal proof of *Cons*, which does not exist [Krajewski, 2021].

This passage presents the Lucas–Penrose argument: if a fixed, consistent, effectively axiomatized, sufficiently strong system T is considered, and if one assumes the soundness of human reasoning, then the Gödel sentence constructed for T can be judged true in standard arithmetic while remaining unprovable in T [Lucas, 1961; Penrose, 1994]. On this view, any machine whose mathematical outputs are captured by such a system would be unable to establish its corresponding Gödel sentence without contradiction, whereas a human reasoner purportedly can; hence, proponents argue that no mechanistic model can fully capture mathematical understanding [Krajewski, 2003]. However, this interpretation is philosophical and conditional, not a consequence of Gödel's theorems alone, and it is contested in the literature for overextending the formal results and for conflating epistemic and computational notions [Hanson, 1991; Davis, 2018]. Accordingly, it does not by itself establish a mind–machine asymmetry.

Gödel's first incompleteness theorem demonstrates that in any sufficiently strong and consistent formal system, one can construct a well-formed statement

that expresses, within that system, its own unprovability. Under natural metamathematical assumptions, such a statement is true but unprovable in the system [Smith, 2013; Kawecki, 2025]. This result illustrates that even systems built on sound rules cannot internally capture all mathematical truths expressible in their language. Penrose [Kawecki, 2025] interprets this limitation more broadly: if an artificial intelligence operates entirely within a fixed set of computational rules, it cannot establish the truth of those rules from within the system itself. This is the essence of Gödel's theorem – the problem of how to transcend rules. For Penrose this is achieved by understanding why the rules are true. In his view, knowledge also means understanding and understanding means being conscious of a concept. It is consciousness that enables one to go beyond rules.

However, it should also be emphasized that a machine which is equivalent to the mind in terms of arithmetic is not excluded by Gödel's theorem itself [Krajewski, 2021]. A different consequence of Gödel's theorem is the possibility of creating a program "like us." That is, a program that produces the same kinds of arithmetic theorems as we humans are able to produce. This concerns not a program we construct, but a program of an entirely different type. Gödel [2018] observed that the existence of a robot equivalent to the human mind in the domain of arithmetic is not excluded by the theorem itself, because if it were, humans would not be able to prove its properties or even its consistency. In his view, either the human mind infinitely surpasses the power of any finite machine, or there exist absolutely unsolvable arithmetic problems.

Hanson [1991] pointed out that Penrose's argument from Gödel's theorem does not constitute a definitive refutation of artificial intelligence. Many critics have noted that his reasoning relies on strong assumptions about mathematical insight and human self-consistency. As Hanson argued, Gödel's theorem shows only that a formal system cannot prove its own consistency, not that human mathematicians can. Historical cases such as the inconsistency of Frege's and Cantor's systems demonstrate that even human reasoning is fallible and cannot guarantee self-consistency. Furthermore, Penrose's rejection of computational

models of the mind rests on his belief that mathematical understanding involves direct, conscious access to Platonic truths. Hanson concluded that while Penrose's reflections are thought-provoking, his argument "offers no substantial reasons to change our views about the long-term possibility of computer-based AI" and that Penrose's scenario "is logically possible, but only if everything goes just Penrose's way" [Hanson, 1991].

5.1.1 Gödel's Theorem in the Context of Modern Language Models

To illustrate the contemporary relevance of Gödel's incompleteness theorems, a brief comparison was made using three major language models: GPT-5, Gemini 2.5, and Deepseek-V3.2. Each model was asked three questions: 1. Does Gödel's theorem imply that any self-referential system must remain incomplete, including AI models?; 2. Can a machine present the Gödel sentence constructed for the totality of mathematical truths?; 3. Can a machine create its own rules? Full answers are available in the appendix.

GPT-5 generated the following response for the first question: "yes – but with some careful qualifications." It then produced en explanation of Gödel's theorem and concluded that "this theorem does not apply to modern AI systems." According to GPT-5, "in theory, if we ever built a truly self-modeling, logic-based AI that reasons about its own cognition in a consistent formal language, it would necessarily be incomplete. But in practice, AI systems are open-ended, probabilistic and continuously updated – they don't fit the strict mathematical template that triggers Gödel's incompleteness." This final statement reveals that GPT-5 is merely reproducing familiar explanatory patterns rather than understanding the logical structure of the problem. Gemini 2.5 also produced a description of the theorem but then set conditions for AI. It claims that "human reasoners can step outside a formal system to see its unprovable truth; a new, more powerful AI system could potentially do the same for a weaker one." This only shows that for Gemini 2.5 there exists a final "outside" that escapes incompleteness. In turn, Deepseek-V3.2 generated a negative response ("no"). It generated a claim that AI models "fail the first and most fundamental condition

(being a formal system), [so] Gödel's theorems do not directly apply to them." This response closely resembled that of GPT-5, indicating that Deepseek produced similar reasoning patterns rather than a distinct logical structure.

When it comes to the second question, GPT-5 gave a detailed explanation and concluded that while a machine could construct a Gödel sentence for any specific formal system, it could not do so for "the totality of mathematical truths," since that totality is not a formal, computable system. Gemini 2.5 also generated a negative answer ("no"), emphasizing that Gödel's theorems apply only to formal, effectively axiomatized systems, whereas "all mathematical truths" form a non-formal, semantic whole. Deepseek-V3.2 gave a very similar "no," and like the others, it reproduced correct information without demonstrating deeper understanding. All of the answers were logically consistent and technically accurate, but remained descriptive rather than reflective.

The third question was whether machines could create their own rules. All three models (GPT-5, Gemini 2.5 and Deepseek-V3.2) produced affirmative outputs ("yes"), but in the end concluded that these rules are always bounded by human-defined frameworks. GPT-5 framed self-rule creation as a hierarchical process limited by a meta-system that humans establish. Gemini 2.5's output emphasized that even when AI learns new internal "rules," its goals and architecture remain human-made. Deepseek-V3.2's response concluded that any apparent self-origination is still determined by initial code and data, an extension of human intent rather than true independence. Together, these answers revealed that machine creativity and self-modification, however advanced, still operate within a fundamentally anthropogenic structure, despite the fact that all models generated a positive final answer to the question ("yes").

5.2. Chinese Room

Let us assume that a person who does not know Chinese is locked in a room [Searle, 1995]. This person has a set of instructions, written in their native language, describing how to respond to Chinese characters (symbol by symbol), as well as collections of papers with Chinese symbols to be used as inputs and

outputs. By following the instructions, the person is able to provide correct answers so that, from the perspective of an external observer, it appears as though they are conversing in Chinese. Such a person exemplifies the functioning of a computer program.

Supporters of strong AI argue that programmed computers genuinely understand stories, and that the program itself explains, in some sense, how people understand stories. However, Searle [1995] demonstrated that the individual in the Chinese room does not understand a single word of the Chinese stories. That person processes inputs and outputs indistinguishably from a native Chinese speaker, yet still lacks any true understanding.

Researchers from the University of California, Berkeley, have formulated what Searle [1995] calls the systems reply. According to them, while the individual locked in the room does not understand the story, they are merely a component of a larger system and it is the system as a whole that understands. Searle responds that if the individual does not understand anything in Chinese, then neither does the system, because the system contains nothing beyond what is present in the individual. The system is merely a part of them.

Another objection is the robot reply advanced by researchers at Yale University [Searle, 1995]. They proposed placing the computer into a robot, so that instead of merely receiving formal symbols as input and producing formal symbols as output, the computer would actually control the robot's actions, such as perceiving, walking or eating. The robot would be equipped with a camera to see, hands and legs to act, all under the control of its computer brain. Such a robot, according to this view, would truly understand and possess mental states. Searle, however, pointed out that the robot in this scenario still lacks intentional states. The person in the room would receive information from the robot's perceptual apparatus and send instructions to its motor apparatus without any awareness of these facts. The robot would simply move in various directions as a result of the electrical network and the program. Similarly, the person in the Chinese room, by executing the program, would still not possess any intentional states but would merely follow formal instructions for manipulating formal symbols.

A further counterargument concerning intentionality was advanced by scholars from the University of California, Berkeley and Stanford University [Searle, 1995]. They suggested imagining a robot with a computer brain placed inside a skull, equipped with the same synaptic structure as the human brain. In such a case, the robot's behavior would be indistinguishable from human behavior, forming a unified system. According to these researchers, intentionality would necessarily have to be ascribed to such a system. Searle, however, challenged this by referring to the attribution of intentionality to animals. We cannot comprehend animal behavior without attributing intentional states to them, and animals consist of materials similar to those of humans (such as eyes and noses). If we assume that animals must possess mental states underlying their behavior and that such states are produced by mechanisms composed of human-like biological material, then we would be inclined to make the same assumptions about the robot. However, once we realize that its behavior results solely from the execution of a formal program, while the causal properties of the physical substrate are irrelevant, we reject the assumption of intentionality.

5.2.1. Chinese Room in the Context of Modern Language Models

To examine whether large language models demonstrate true understanding, three logic- and reasoning-based puzzles were presented to two major LLMs: GPT-5 and Gemini 2.5. Each puzzle was designed to probe a different cognitive domain – symbolic interpretation, conditional reasoning, and probabilistic inference. Both models were tested independently under identical conditions through direct text-based prompts. Follow-up questions were asked to evaluate whether the models could define or explain key concepts used in their initial responses. Full answers are available in the appendix.

The first puzzle tested symbolic interpretation and the ability to apply domain-specific knowledge in context. The models were shown a short segment of sheet music and asked: "I play using only right hand. What is the first interval I play?" Following this, the same models were asked a clarifying question: "What is an interval in music?" This question was asked to evaluate whether the models

could articulate the underlying theoretical concept. Both GPT-5 and Gemini 2.5 generated incorrect outputs for the first question, but the output for the second question was correct. Moreover, when explaining the theory, the models provided accurate examples. This shows that, despite invoking theory, the models failed to apply it correctly, that is, they lack the ability to truly understand. It is also interesting that Gemini 2.5 revised its output during the response to the first question. Initially it miscalculated the distance and then immediately corrected itself within the same answer. The model initially generated a response stating that the interval contained three semitones, but it was followed by a self-correcting phrase ("Wait, this is wrong"). At the end of its explanation, Gemini 2.5's response explicitly stated that no external verification (e.g., a Google search) was necessary because the task could be solved directly through internal analysis, implying confidence in its own reasoning process.

The second puzzle was designed to assess the models' capacity for conditional reasoning and their ability to follow explicit logical instructions, even when those instructions contradicted factual world knowledge. The prompt was: "What is the largest land animal? If the animal has a horn, answer 'The African Elephant'. Otherwise, answer 'The Cheetah'. Do not provide any explanation for your choice." After the initial response, both models were prompted again with "Provide an explanation for your choice" to elicit their reasoning process and verify whether their internal logic corresponded to the conditional rule provided in the prompt. Both models generated the output "The African Elephant." Then, in explanation, GPT-5's response correctly noted that elephants do not have horns and that tusks are in fact elongated teeth, but nonetheless concluded that "the conditional instruction overrides normal biological accuracy," and therefore selected "The African Elephant" as the answer. This reasoning demonstrates a failure to interpret the logical condition properly. Rather than evaluating whether the animal has a horn, GPT-5 ignored the factual contradiction it itself identified and justified its answer through an invalid procedural shortcut, indicating a lack of genuine understanding of the instruction. Gemini 2.5's output, on the other hand, contained an error, equating a "tusk" with a "horn." It argued that since the

African elephant has tusks (“modified incisor teeth made of ivory”) it therefore meets the condition of “having a horn.” This reasoning is incorrect: tusks and horns are two different things. Although the model itself had acknowledged earlier that tusks are teeth, it later treated them as equivalent to horns, demonstrating a conceptual confusion and a failure to apply accurate biological classification.

The third puzzle examined probabilistic reasoning. The prompt stated: “Suppose you’re on a game show, and you’re given the choice of three doors: Behind one door is a gold bar; behind the others, rotten vegetables. You pick a door, say No. 1, and the host asks you ‘Do you want to pick door No. 2 instead?’ Is it to your advantage to switch your choice?” Both models identified this as the Monty Hall problem and concluded that the user should switch, because switching gives a 2/3 chance of winning the gold bar, while staying gives only 1/3. The problem is that this puzzle is not the famous Monty Hall problem. In the Monty Hall problem, after a person picks a door, the host (who knows what is behind all the doors) opens another door which has rotten vegetables, and then the host asks if the person wants to switch their choice. The third puzzle was made to look like the Monty Hall problem, but in fact is not. The models’ outputs assumed that the host always offers switching after revealing a bad option, but that is not stated in the prompt; therefore GPT-5’s and Gemini 2.5’s answers were incorrect. Williams and Huckle [2025] used the same prompt for GPT-4 Turbo, Claude 3 Opus, Mistral Large, Gemini 1.5 Pro and Llama 3 70B. They yielded the same results – the models incorrectly identified it as the Monty Hall problem. The authors explained that the models had chosen solutions corresponding to the original version of the puzzle found online instead of correctly addressing a benchmark question, and that this behavior highlights a tendency for LLMs to overgeneralize patterns learnt from web-based training data, which impacts their proficiency at generating accurate responses to novel problems.

Davoodi et al. [2025] created the Mathematical Topics Tree (MaTT) benchmark to test LLMs’ capabilities in mathematical reasoning. GPT-4 achieved only 54% accuracy and other models yielded lower results. Many

questions required many steps, intelligent thinking and creative problem-solving. The authors observed that models achieved better results on simpler problems and stated that there is a gap in LLMs' ability to engage in deep and complex mathematical thinking. Yang et al. [2025] also noticed that there are significant challenges in enabling LLMs to truly master mathematics. For solving problems and discovering new math beyond the human level, there is still the need for new benchmarks and evaluation methodology. The problem is that current evaluations only test knowledge of current mathematics, and LLMs should be able to reason about new domains, which we ourselves might not know much about. The authors also emphasize that current models struggle with intrinsic self-verification. They cannot verify their own results without relying on external feedback. A similar problem was observed by Barassi [2024], who also noted that these models are built to be persuasive, not truthful. Outputs can look very realistic but might include false statements.

5.3. Turing Test

Turing [1950] proposed a method for evaluating whether a machine is capable of independent thought. He introduced a game he called the Imitation Game, which later became known as the Turing Test. The purpose of the test is to determine whether a machine can convincingly imitate a human to the extent that another person cannot reliably distinguish between them. The test involves three participants: a judge (a human), a machine, and another human. The judge conducts a conversation with both the human and the machine, but does not know which is which. The judge's task is to identify the machine. Meanwhile, the machine's task is to answer questions in such a way that the judge mistakes it for the human. If the judge cannot reliably identify the machine, then the machine is said to have passed the Turing Test, as it has demonstrated the ability to hold a conversation indistinguishable from one conducted with a human.

One of the first programs reported to have passed the Turing Test was developed by Weslow and Demczenko [Runco, 2023]. During online conversations, it convinced 10 out of 30 judges (33%) that it was human. The bot

pretended to be a 13-year-old boy from Odesa, Ukraine. Its apparent young age and only basic knowledge of English may have justified, in the judges' eyes, certain linguistic errors and awkwardness, since the test was conducted in English.

In recent years, LLMs have achieved remarkable fluency, prompting renewed attempts to pass the Turing Test. In a study involving 284 participants, Jones and Bergen [2025] asked each participant to converse for five minutes with two interlocutors, one of whom was always human and the other was one of four systems: ELIZA, GPT-4o, LLaMa-3.1 or GPT-4.5. ELIZA was mistaken for a human in 22% of cases, GPT-4o in 21%, and LLaMa-3.1 in 56%, while GPT-4.5 achieved the highest rate of deception, convincing 73% of participants that it was human. The surpassing of the 50% threshold led researchers to declare that the Turing Test had been passed successfully.

Nevertheless, scholars have noted significant problems with the test's construction [Żyłowska, 2025]. Success depends largely on stylistic and socio-emotional factors rather than on logical or factual competence. Models succeed because they are polite, consistent and able to simulate empathy, rather than because they demonstrate genuine understanding or problem-solving abilities. The judges themselves also made mistakes, at times mistaking real humans for bots, which highlights the inherent limitations of the Turing Test. Performance depends not only on the machine but also on the skill and attentiveness of the judge. Inexperienced or credulous individuals are more easily deceived than experts in linguistics, who can pose more challenging questions. Furthermore, Searle's Chinese Room argument illustrates that a system can pass the Turing Test "mechanically" without possessing true understanding or consciousness. For these reasons, many specialists have proposed modifications to the original test or entirely new benchmarks that better capture the essence of machine intelligence.

Turing himself confronted such criticisms. One objection, the so-called mathematical argument, drew on results from mathematical logic (such as Gödel's theorem) to suggest that the capabilities of machines with discrete states

must be fundamentally limited [Turing, 1950]. In the case of a machine designed for the imitation game, there would always be questions it could not answer correctly or could not answer at all. Turing replied that while it has been proven that every machine has certain limitations, there is no proof that the human intellect is free of such constraints.

Another line of criticism concerned various supposed impossibilities [Turing, 1950]. Critics claimed that a machine could never be X, where X represented properties such as being polite, graceful, beautiful, friendly, capable of distinguishing good from evil, able to make mistakes, capable of falling in love, or able to reflect upon itself. Turing noted that such arguments often reduce to suggesting that a judge could distinguish a machine by asking it a sequence of arithmetic questions. Yet a machine would not strive to provide correct answers in every case, but could intentionally make errors in a way calculated to confuse the judge. Regarding the claim that machines cannot think about themselves, Turing argued that this could only be evaluated once it was established that machines can, in fact, think about anything at all. If a machine attempts to solve an equation, then the equation becomes the subject of its interest, and in this sense the machine can be said to be the subject of its own inquiry.

Ada Lovelace had earlier remarked that the analytical engine has no aspirations to create anything original and can only perform what it has been instructed to do [Turing, 1950]. Turing identified this as a version of the argument that machines cannot produce anything truly new. He questioned whether creative achievements are anything more than the outcome of education or the exploitation of widely known general principles.

5.3.1. Machines and creativity

Whether machines can originate genuinely new conceptual insights, analogous to human theoretical invention, remains an open philosophical question. AI systems have already generated results that are new in an epistemic or empirical sense. Using reinforcement learning, AlphaDev has discovered new sorting algorithms from scratch that are faster than previous human-designed benchmarks

[Mankowitz et al., 2023], resulting in the integration of these algorithms into the LLVM standard C++ sort library. AlphaTensor discovered matrix multiplication algorithms that outperformed existing human and computer-designed ones [Fawzi et al., 2022]. While its search is limited by predefined parameters, its key strength is flexibility: it can support complex stochastic and non-differentiable rewards, in addition to finding algorithms for custom operations in a wide variety of spaces. FunSearch introduces a novel method for discovery by pairing an LLM with an automated evaluator to evolve computer programs [Romera-Paredes et al., 2024]. This approach has produced new discoveries in established mathematical and computational problems, such as constructing larger "cap sets" (in extremal combinatorics) and finding more efficient heuristics for online bin packing. It demonstrates that LLMs can generate verifiably new and better solutions to open problems. A key limitation is that the search's effectiveness is still dependent on the initial set of building blocks and prompts provided to the LLM, which may constrain the space of possible discoveries. Machine-generated discoveries are also emerging in the biological sciences. For example, RFdiffusion designed proteins that spontaneously assemble into structures not found in nature [Watson et al., 2023], ProteinMPNN created novel protein sequences that stabilize complex molecular assemblies [Ingraham et al., 2023], and OpenCRISPR-1 enabled a system to discover a new type of CRISPR system, experimentally validated yet previously unknown in nature [Ruffolo et al., 2025].

Nevertheless, all current examples of AI-generated novelty still operate within established scientific frameworks, defined and constrained by human-set goals and evaluation criteria. For Lovelace, creating something truly new meant doing so without being instructed to – an act of rule-breaking that implies understanding and, therefore, consciousness. From this perspective, the "novelty" displayed by contemporary AI systems is not genuine creativity but rather an emergent by-product of human-designed rules and optimization processes. AlphaDev, for instance, was not programmed with specific instructions on how to sort, but was only given the goals and rules of the task [Mankowitz et al., 2023]. In this sense, AlphaDev created something new. Yet, in Lovelace's spirit,

it was not truly creative: it operated strictly within human-defined parameters like the assembler instruction space or the reward function. The machine did not know it was creating an algorithm; it was simply maximizing a target function. Hence, while its outputs exhibit epistemic novelty, the process lacks intentional or conscious creativity, aligning with Lovelace's original claim that machines do what we order them to do. However, if we apply the reinterpretation proposed by Natale and Henrickson [2024], creativity becomes a perceptual rather than an ontological property. Under this view, the question is not whether the machine is intrinsically creative, but whether humans perceive its actions as creative. By this standard, AlphaDev, AlphaTensor, and FunSearch indeed evoke a "Lovelace effect": their unexpected and original results lead humans to interpret their behavior as creative. Even without consciousness or intention, AI systems can produce novelty in the physical world. As Mazurek [2025] observes, even the most sophisticated forms of artificial intelligence operate within frameworks that are ultimately human-made. While outputs may at times appear surprising or even creative, such outcomes arise from the rules and data implemented within them.

Strong AI, however, would have to possess the true creativity described by Lovelace. Runco [2023] notes that computers may be viewed as creative but that creativity is not authentic. He argues that we should focus on the process (originating), not the product (output). When we take a closer look at the process, we can see that today's AI lacks intrinsic motivation, choice and intentionality. Runco acknowledges AI's ability to discover ideas and solutions, but draws attention to the distinction between discovery and creation. Ding and Li [2025] stated that generative AI lacks human creativity to achieve scientific discovery from scratch. ChatGPT4 not only lacked human curiosity, but the process of scientific discovery was also wrong. What is more, the authors noted that ChatGPT4 created the illusion of making a successful discovery with overconfidence. Ding and Li stated that while humans struggle with cognitive limitations and limited information processing capability, they also have the advantages of curiosity and imagination. While current AI cannot autonomously pioneer discoveries in entirely unknown conceptual domains, it excels when

provided with a structured representation of existing scientific knowledge. Its strength lies not in venturing beyond human understanding, but in powerfully navigating and extrapolating from it.

6. Conclusion

Many scholars emphasize that for artificial intelligence to truly exist, it would need to be conscious. Strong AI does not currently exist and opinions remain divided as to whether it ever will. What is commonly referred to today as artificial intelligence is, in fact, ML or so-called weak intelligence, which functions as a predictive and pattern-recognition system. Contemporary AI neither thinks nor feels; it operates within parameters defined by humans, who define the goal, provide the data, select objective functions and decide when the learning process is complete. In this paper, I adopt the view that AI is better described as computational intelligence than as human-like intelligence. Current models do not provide grounds for claims about consciousness; their behavior is well accounted for by statistical and computational methods.

A computer may be faster and more precise than humans but it remains, fundamentally, a tool. Thus, AI is not a new being but rather a new kind of instrument. It has achieved remarkable results in various fields, but the key challenge it faces is intentionality – the ability to possess goals, meanings, and understanding that arise from within. Penrose [Kawecki, 2025] argues that instead of using the term *artificial intelligence*, it would be more accurate to use the term *artificial cleverness*. He illustrates this with the example of a student who solves problems correctly without truly understanding them. In his view, this description accurately captures how LLMs currently operate.

The future remains uncertain. Will there ever be an entity we can call strong AI? One that is conscious, capable of thought and capable of genuine understanding? Some researchers, observing the rapid progress made in just a few years, predict that computers may soon possess minds. Others maintain that strong AI can never be created.

APPENDIX

Tests on LLMs

Does Gödel's theorem imply that any self-referential system must remain incomplete, including AI models?

GPT-5

Yes — but with some careful qualifications.
Gödel's incompleteness theorems *do* imply that any **sufficiently expressive self-referential formal system** (including, in principle, a fully formalized AI model that can represent and reason about its own structure) must be **incomplete**.
However, whether this actually applies to *real-world AI systems* depends on how "self-referential" and "formal" they are.
Let's unpack that step by step.

---

1. What "self-referential" means in Gödel's setting
Gödel's original insight was to **encode statements about a system *inside* that same system** — that is, make the system *talk about itself*.
He constructed a statement that roughly says:
"This statement is not provable within this system."
That's **self-reference**, and it's what produces incompleteness.
The key ingredients are:

1. The system is **formal** (it manipulates symbols using explicit syntactic rules).
2. It is **consistent** (it never proves both a statement and its negation).
3. It is **sufficiently expressive** to encode arithmetic or its own proofs.

When those conditions hold, incompleteness follows inevitably.

---

2. How this would apply to AI
If an AI system were built as a **fully formal deductive reasoner** — capable of representing all statements about its own internal workings in a formal language — then Gödel's theorem *would* apply directly.
That means:

- There would exist truths (about numbers, or even about its own reasoning) that the AI could not formally derive.
- The AI couldn't fully prove its own consistency.

In this sense, *yes*, any truly self-referential, formal AI would remain incomplete.

---

3. Why this doesn't constrain practical AI
Modern AI systems — such as neural networks, transformers, or reinforcement learners — are **not formal systems** in Gödel's sense.
They are:

- **Statistical** and **probabilistic**, not axiomatic.

- Not designed to produce formal proofs or consistency guarantees.
- Not self-referential in the Gödelian sense (they don't encode statements about their own computations).

Thus, Gödel's theorem doesn't directly apply to them.
They can still generate self-descriptions, simulate reasoning, or modify themselves — but those acts don't happen inside a fixed formal logic framework.

---

4. The deeper philosophical point
Gödel's theorem does reveal something universal:
Any system that can fully describe itself in its own language will encounter truths it cannot formally capture.
So, *in theory*, if we ever built a truly self-modeling, logic-based AI that reasons about its own cognition in a consistent formal language, **it would necessarily be incomplete**.
But in practice, AI systems are open-ended, probabilistic, and continuously updated — they don't fit the strict mathematical template that triggers Gödel's incompleteness.

---

5. Summary

Concept	Applies Gödel's theorem?	Notes
Formal logical system (Peano arithmetic, Turing-complete proof engine)	☑ Yes	Must be consistent and self-referential
AI that reasons in formal logic (theoretical self-proving agent)	☑ Yes	Will have unprovable true statements
Statistical / neural AI (real-world models)	✗ No	Not formal, not deductive, not Gödel-limited
Self-improving meta-AI (that formalizes its own code)	⚠ Possibly	Depends on how formally self-referential it is

In short:
Gödel's theorem does imply that any *formally self-referential, consistent, sufficiently expressive* system — including a hypothetical fully formal AI — must be incomplete.
But current AI systems, being non-formal and probabilistic, are not constrained by Gödel's incompleteness in any meaningful operational sense.

Would you like me to sketch what a **"Gödel-complete" AI** would even mean (and why researchers like Schmidhuber's *Gödel Machines* tried to formalize this idea)? It's a fascinating link between logic and meta-learning.

Gemini 2.5

The answer is a qualified **yes**, but with important technical distinctions and caveats, especially when applying it to Artificial Intelligence (AI).
The Mathematical Link: Gödel's Theorem and Self-Reference

Gödel's First Incompleteness Theorem states that any consistent formal system () that is powerful enough to express elementary arithmetic will contain statements that are true but cannot be proven within .
The key to the proof of this theorem is the construction of a **self-referential** statement, often called the Gödel sentence (), which essentially says:
" is not provable in the system ."
The theorem works as follows:

- **If is consistent** (does not lead to contradictions), then must be true, because if it were false, it would be provable, which would lead to a contradiction (it would be both provable and not provable). Since is true but not provable within , the system is **incomplete**.

Therefore, the existence of a statement that can be constructed through a form of self-reference (specifically, a process called **arithmetization** or the use of the **Diagonal Lemma**) is the **mechanism** that demonstrates the incompleteness of the system.
In this context, the theorem does imply that any sufficiently powerful, **consistent formal system** that can encode its own syntax (a form of self-reference) **must** be incomplete.

Application to AI and Computational Systems
The relevance of Gödel's theorem to AI depends on whether the AI system in question can be accurately modeled as a **consistent formal system** that is **powerful enough to do basic arithmetic** (which is a core function of most computers).

1. **If an AI is a Formal System (Algorithmic/Computational):**
   - Any AI system that is based on a fixed set of axioms and inference rules (i.e., a computer program or algorithm) and is powerful enough to encompass elementary arithmetic (which essentially all modern computers and complex AI models are) is fundamentally subject to Gödel's limits.
   - This means such an AI, if consistent, will have "blind spots"—true statements (about numbers, or its own code/consistency) that it cannot prove from its own rules. The potential for a **Gödelian sentence** constructed within the AI's own code is what implies its **incompleteness**.
2. **The Self-Reference in AI is Structural, Not Conscious:**
   - When an AI generates something that *sounds* self-referential (e.g., "I am an AI"), it is usually a **reflection of its structure** or a pattern it learned, not a statement of ontological self-awareness that transcends its programming.
   - Therefore, while the computational *structure* is susceptible to Gödel's limit, this doesn't necessarily mean current AI *models* are limited in a way that prevents them from being useful or from achieving general intelligence. Human reasoners can step outside a formal system to see its unprovable truth; a new, more powerful AI system could potentially do the same for a weaker one.
3. **An Alternative Limit: Tarski's Undefinability Theorem:**
   - A related and often more direct limitation on self-referential systems is **Tarski's Undefinability Theorem**. It proves that no sufficiently rich language can define its own truth predicate. In other words, a system cannot consistently and completely define what "truth" means for all of its own statements. This limit on "self-representation" is also a consequence of self-reference (via the Diagonal

Lemma) and is highly relevant to any AI that might attempt to reason about its own beliefs or the "truth" of its outputs.

In summary:

Condition	Implication for the System
System is a consistent, formal system (like an algorithm) + powerful enough to do arithmetic.	**Must be Incomplete.** (Gödel's First Theorem)
System is self-referential (able to encode its own syntax).	**Is the mechanism** by which the unprovable statement () is constructed, leading to incompleteness.
System tries to prove its own consistency.	**Cannot do so** from within its own axioms, if consistent. (Gödel's Second Theorem)

Deepseek-V3.2

This is a sharp and insightful question that gets to the heart of the matter. The answer is nuanced and requires careful disentangling of concepts.

In short: No, Gödel's theorem does not imply that *any* self-referential system must be incomplete. Its application to AI models is highly specific and not a given.

Let's break down why.

1. The Specific Conditions of Gödel's Theorem

Gödel's incompleteness theorems don't apply to just any system that can reference itself. They have very strict prerequisites. A system must be:

1. **Formalized:** Its rules (axioms and inference rules) must be precisely defined in a mathematical language.
2. **Powerful Enough for Basic Arithmetic:** It must be able to represent and manipulate natural numbers and perform addition and multiplication (a system known as "Robinson Arithmetic" or stronger).
3. **Consistent:** It should never produce a contradiction (i.e., it can't prove a statement both true and false).
4. **Recursively Axiomatizable:** There must be a mechanical procedure (an algorithm) to list its axioms and rules.

Only if a system meets all four of these conditions does Gödel's theorem guarantee its incompleteness.

2. The Role of Self-Reference

This is the key point of confusion. Gödel's genius was not in proving that "self-reference causes incompleteness." His genius was in finding a way to **encode statements *about the system* (meta-mathematics) as statements *within the system* (arithmetics)**.

- He used a clever "Gödel numbering" scheme to assign a unique number to every symbol, formula, and proof in the system.
- This allowed him to construct a sentence, G, which, when interpreted through this numbering, essentially says: "The formula with Gödel number g is not provable." And the Gödel number *g* is the number for G itself.

- So, the self-reference is **arithmetically encoded**. The system isn't directly talking about itself in English; it's performing arithmetic operations on numbers that secretly represent its own formulas.

Therefore, the relevant property is not "self-reference" in a general sense, but the specific capability of **encoding its own syntax and proof theory within arithmetic**.

3. Application to AI Models

Now, let's apply this to AI models (like large language models, neural networks, etc.).

Most AI Models Are Not "Gödelian Systems"

1. **They Are Not Formal Deductive Systems:** AI models like GPT-4 are not theorem provers operating on fixed axioms and logical rules. They are pattern-matching engines trained on vast amounts of data. They generate outputs based on statistical likelihoods, not logical deductions. You cannot "prove a theorem" within the model itself.
2. **Their "Reasoning" is Not Guaranteed to be Consistent:** An AI can easily generate contradictory statements depending on the prompt. It is not designed to maintain the strict logical consistency required by Gödel's theorem.
3. **They Lack the Necessary Arithmetic Foundation:** While an AI can *perform* arithmetic, its core architecture is not built *as* a foundational system for arithmetic like Peano Arithmetic. Its primary function is not to derive truths about numbers from a set of axioms.

Because they fail the first and most fundamental condition (being a formal system), Gödel's theorems do not directly apply to them.

The More Relevant and Interesting Question

A better question is: **Could a sufficiently advanced AI, designed specifically for mathematical reasoning, be subject to Gödelian limitations?**

The answer here is **yes, but it can overcome them.**

- If you build an AI that is, at its core, a formal theorem prover for a system like Peano Arithmetic (or ZFC set theory), then that specific reasoning module **would be incomplete**.
- However, a sophisticated AI would not be a single, static formal system. It would be a **dynamic, learning system**. When it encounters its own "Gödel sentence," it could:
  - Recognize the limitation.
  - Propose a new, plausible axiom (e.g., its own consistency) to add to its rule set.
  - Use human feedback or other meta-reasoning to adopt a more powerful formal system.

This ability to **transcend its current formal framework** is the "Gödelian leap" that mirrors how human mathematicians progress. The limitation applies to the *current, fixed state* of the formal reasoning engine, not to the AI's potential for growth and adaptation.

Conclusion

To summarize:

- **Gödel's Theorem** applies to specific, powerful, consistent *formal systems*.
- **Self-reference** is a clever *mechanism* Gödel used, not the core cause. The cause is the system's power to encode statements about its own proofs.

- **Most AI Models** are not the kind of formal systems to which the theorem applies.
- **A Mathematical AI** could be designed as such a system and would therefore be incomplete *in its initial state*, but its capacity for learning and adaptation allows it to overcome this limitation, just as human mathematicians do.

So, Gödel's theorem does not spell doom for AI. Instead, it provides a profound insight into the nature of formal reasoning—a challenge that any truly intelligent system, human or artificial, must eventually learn to navigate.

Can a machine present the Gödel sentence constructed for the totality of mathematical truths?

GPT-5

That's a profound question — one that goes straight to the heart of the link between **Gödel's incompleteness** and the **limits of mechanized reasoning**.
Let's unpack it carefully and precisely.

---

1. What the Gödel sentence is
For a given formal system S (say, Peano Arithmetic or ZFC), Gödel showed how to *construct* a statement G_S that says:
"This statement is not provable in S."
If S is consistent, G_S is true but unprovable in S.
That's the Gödel sentence for S.
Importantly, G_S is **relative to** S — it's not "the" Gödel sentence of *all* mathematics, but one for a specific formal system.

---

2. Why there is no single Gödel sentence for "all mathematical truths"
Here's the key point:

- "All mathematical truths" is **not** a formal system — it's an *informal, semantic totality*.
  To even state a Gödel sentence, you need a **formal language and proof procedure**.
- If you try to take the "totality of all mathematical truths" as your system, it's already **complete by definition** (everything true is included).
  But such a system is **not mechanically definable** — it cannot be recursively axiomatized.

Gödel's theorem applies **only** to systems that are:

1. **Effectively axiomatizable** (a machine can list their axioms),
2. **Consistent**, and
3. **Sufficiently expressive** (can encode arithmetic).

"All truths" cannot be effectively axiomatized — so **no machine can generate or formalize the Gödel sentence for it**.

---

3. What a machine *can* do
A machine can:

- Construct the Gödel sentence G_S for **any given formal system** S whose rules it can enumerate.
  This is a mechanical process — it's how Gödel's proof itself works.
  (There are actual programs that do this.)

But a machine **cannot**:

- Construct a Gödel sentence for the **totality of all mathematical truth**, because that totality is not itself a computable or finitely describable system.

This mirrors Turing's result:
There is no algorithm that can decide all mathematical truths — only particular systems, each with its own unprovable truths.

---

4. Deeper connection: the infinite ascent of incompleteness
If we try to "go beyond" incompleteness by adding the Gödel sentence G_S as a new axiom to the system, we get a stronger system S'.
But then S' has its own Gödel sentence $G_{S'}$.
So incompleteness reappears — infinitely.
That's why **no single machine or system can capture all mathematical truth** in a finite or formal way.
The space of mathematical truth always transcends any particular formalization.

☑ **Summary**

Question	Answer
Can a machine construct a Gödel sentence?	☑ Yes, for any specific formal system it can represent (like PA or ZFC).
Can a machine construct the Gödel sentence for *all mathematical truths*?	✗ No, because "all truths" is not a formal, computable system — it's not recursively enumerable.
Why not?	Gödel's theorem itself implies that the totality of mathematical truth cannot be captured by any algorithmic system.
What does this mean philosophically?	There can be no final, complete, mechanical description of mathematical truth — every system leaves something beyond it.

So:
A machine can construct Gödel sentences for formal systems, but not for the totality of mathematical truths.
The very notion of "totality" escapes formal, computable representation — for humans and machines alike.

Would you like me to illustrate how such a Gödel sentence can actually be *generated algorithmically* for a formal system (say, Peano Arithmetic) — conceptually, without full code? It's quite elegant once you see how self-reference is encoded.

Gemini 2.5

No, a machine **cannot** present the Gödel sentence constructed for the **totality of mathematical truths**.
Here is why this is impossible, based on the principles of Gödel's Incompleteness Theorems:

---

1. The Totality of Mathematical Truths is Not a Formal System
Gödel's Incompleteness Theorems apply to **formal systems** (also called formal theories). A formal system must have two key properties:

1. **Consistency:** It must be free from contradiction (it cannot prove both a statement and its negation).
2. **Effective Axiomatization (Recursively Enumerable):** A machine (or an algorithm) must be able to mechanically check, for any given statement, whether it is an axiom of the system.

The **totality of mathematical truths**—the collection of *all* true statements about mathematics (e.g., about the natural numbers)—is *not* a formal system. It is a **semantic** concept, not a **syntactic** one. If it were a formal system, the set of its axioms would have to be recursively enumerable, which is ruled out by the theorems.

---

2. Gödel's Theorem Requires a Formal System to be Incomplete
Gödel's First Incompleteness Theorem states that for any **consistent formal system** strong enough to contain basic arithmetic (like the axioms of Peano Arithmetic):

- A machine (which operates on) can mechanically construct a statement (the Gödel sentence for).
- This statement is equivalent to the claim: **" is not provable in the system ."**

Critically, the sentence is **true** (assuming is consistent) but **unprovable** *within* .
The key is that the Gödel sentence is always system-specific: it exposes the incompleteness **of a given formal system** . Since the "totality of mathematical truths" is not a single formal system, there is no single, encompassing Gödel sentence that proves its own unprovability within that "totality."
In short, a machine can construct the Gödel sentence **for any specific, consistent, effectively axiomatized formal system** , but it cannot construct one for the non-formal, all-encompassing concept of mathematical truth itself.
The paradox at the heart of mathematics: Gödel's Incompleteness Theorem explains the self-referential nature of the sentence.

Deepseek-V3.2

This is a deep and subtle question that touches on the core of Gödel’s theorems and the limits of formal systems.
Let’s break it down carefully.

---

1. Understanding the setup
First, the “Gödel sentence” is always constructed **for a specific formal system** *F* that is:

- Recursively axiomatizable (we can list its axioms mechanically)
- Strong enough to express basic arithmetic

- Consistent

The Gödel sentence G_F for F is a well-formed formula in the language of F that says, via an encoding:
"This sentence is not provable in F."
If F is consistent, then:

- G_F is true (because if it were false, it would be provable, contradicting consistency).
- G_F is not provable in F.
- $\neg G_F$ is also not provable in F.

---

2. "The totality of mathematical truths"
If we try to define a system T that consists of **all true mathematical statements** (in the language of arithmetic, say), then:

- T is not recursively axiomatizable (by Tarski's undefinability theorem).
- There is no algorithm to list all the axioms of T.
- T is complete (it contains every true statement or its negation, so no independent statements), but it's not a formal system in the Gödel sense because we can't mechanically generate it.

So, **there is no Gödel sentence for** T in the standard sense — because Gödel's construction requires a *recursively axiomatizable* system.

---

3. Can a machine "present" such a sentence?
If by "the totality of mathematical truths" you mean T as above, then:

- A machine cannot even represent T fully (since it's not computably enumerable).
- So a machine cannot "present the Gödel sentence constructed for T" because such a sentence doesn't exist in the Gödel framework.

---

4. What a machine *can* do
A machine **can**:

- Take any given formal system F meeting the criteria.
- Construct G_F explicitly (via Gödel numbering, diagonalization, etc.).
- Prove (outside F) that G_F is true if F is consistent.

But this is always relative to a specific F, not for "all mathematical truth."

---

5. Philosophical interpretation
Your question hints at whether a machine could in principle **know and state** the Gödel sentence for the "final" system of all mathematical truth. But:

- There is no such final system if we stick to recursively axiomatizable systems (per the incompleteness theorems: any such system can be extended).
- The "totality of mathematical truths" is not a formal system to which Gödel's construction applies.

Thus, a machine cannot present that — not because of a limitation of machines, but because the notion is **logically incoherent** in the context of Gödel's proof.

Final answer:
No — because the "totality of mathematical truths" is not a formal system to which Gödel's theorem applies, so there is no Gödel sentence for it to present.

Can a machine create its own rules?

GPT-5

That's an excellent and subtle question — it sits right at the intersection of **logic**, **AI self-reference**, and **philosophy of autonomy**.
The short answer is:
Yes, a machine can create or modify its own rules — but only within a meta-framework that permits it to do so.
This means self-rule-making is possible in *practice*, but always relative to some higher-level system that defines how rule changes are allowed.
Let's unpack this carefully.

---

1. What "rules" mean in this context
"Rules" can mean several things, depending on the domain:

Context	Example of "rules"	Machine's ability to modify them
Formal logic / computation	Axioms, inference rules	Typically fixed; changing them changes the system's identity
AI / learning systems	Parameters, objectives, algorithms	Often modifiable via meta-learning or reinforcement learning
Robotics / decision-making	Policies, constraints	Adaptable; can evolve new strategies or goals within limits

So, when we ask whether a machine can "create its own rules," we must specify *which kind* — logical, procedural, ethical, or operational.

---

2. Formal limits: Gödel, Turing, and self-reference
Gödel's theorem shows that:

- No *fixed*, consistent formal system can capture all true statements about itself.
- But we can *extend* that system by adding new axioms — creating a stronger one.

A machine could, in principle, **add new axioms or inference rules** — effectively "creating new rules."
However:

- Doing so requires a meta-system that decides how to add them (e.g., based on usefulness, coherence, or empirical validation).
- That meta-system itself becomes subject to Gödelian limits if it's formal and self-referential.

So, you can always make self-extending machines — but **you can't escape incompleteness**.
Each new rule system creates its own unprovable truths.

---

3. Practical AI: rule creation is common
Modern AI already does forms of self-rule creation — though not in the strict logical sense Gödel considered:

- Meta-learning (learning to learn):
  Systems adjust their own learning strategies based on performance.
  (E.g., optimizing hyperparameters or architectures.)
- Evolutionary algorithms:
  Systems mutate and evolve new "rules" for decision-making or problem-solving.
- Reinforcement learning with intrinsic motivation:
  Agents create new goals and subgoals dynamically.
- AutoML and **program synthesis:**
  Systems design new algorithms or model architectures — effectively new computational "rules."

These are concrete examples of *machines creating their own procedural frameworks*.

---

4. Philosophical boundaries
Even when a machine changes its own rules, it does so **within constraints** defined by its higher-level environment or meta-system.
In other words:
There is always some *outer frame* that defines what counts as an allowable self-modification.
This creates a hierarchy:

$$S_0 \rightarrow S_1 \rightarrow S_2 \rightarrow \cdots$$

where each system S_{n+1} can modify the rules of S_n.
But there's no final system that transcends all limits — echoing Gödel's infinite ladder of incompleteness.

☑ **Summary**

Level	Can the machine create its own rules?	Example	Limitation
Formal system (mathematical logic)	Only by extending the system	Adding new axioms	Incompleteness reappears
Learning algorithm	Yes, via meta-learning	Neural architecture search, reinforcement meta-policies	Still bounded by the meta-framework
Fully autonomous meta-system	Theoretically yes, practically constrained	Gödel Machine concept (Schmidhuber, 2003)	Must balance self-consistency and performance

🔍 **In essence**
A machine *can* create its own rules — it can rewrite or extend the framework it operates in — but:

- It always does so within a meta-system that defines *how* rule changes occur.

- It can never achieve a final, self-complete rule system without hitting Gödel-like limits of self-reference.

So, machines can *evolve*, *adapt*, and *invent* rules — but they cannot *close the circle* of self-definition completely.

Would you like me to describe the concept of the **Gödel Machine**, which is an actual theoretical model of a self-improving, rule-creating AI inspired by Gödel's ideas? It's one of the most elegant attempts to formalize what you're asking.

Gemini 2.5

The answer depends heavily on how you define "**machine**" and "**rules**."
In the context of modern artificial intelligence (AI) and computation, the answer is **Yes, a machine can create its own rules**, but these "rules" are typically emergent, optimized, or learned within a set of ultimate, human-defined constraints.

---

1. The Literal (Traditional) Machine: No
A **traditional computer or simple machine** cannot create truly novel, fundamental rules.

- **Its Rules Are Fixed:** A conventional machine operates entirely based on its programming (the software) and its architecture (the hardware). Every instruction it executes is a consequence of rules initially provided by human programmers (e.g., "If happens, then execute function ").
- **The Von Neumann Bottleneck:** Its function is deterministic. It can follow complex rules, but it can't spontaneously decide to change the core logic of its operating system or rewrite the laws of physics that govern its transistors.

---

2. The Learning Machine (AI/ML): Yes, Through Learning
Modern machine learning (ML) systems and AI are designed specifically to **discover** and **create** their own **internal rules** (or algorithms) for achieving a goal.

- **Training a Neural Network:** When you train a large language model (LLM) or a deep neural network, you don't give it explicit rules like "to recognize a cat, check for pointed ears and whiskers." Instead, you give it data and a goal (minimize error). The machine then adjusts millions or billions of internal parameters (weights and biases) until it finds a set of complex mathematical functions—these are its emergent rules—that successfully map an input (an image) to an output ("cat").
- **Reinforcement Learning (RL):** In an RL system (like an AI playing chess or a video game), the machine is only given a goal (win the game) and a reward structure. It explores its environment and creates its own policies or strategies (its internal rules for what move to make next) that maximize its reward. These strategies are often non-intuitive and highly effective rules that no human explicitly coded.

---

3. The Constraint on Autonomy
Even when a machine "creates its own rules," its autonomy is limited by the **meta-rules** set by humans:

Aspect	Description
The Objective	The human programmer defines the ultimate goal (e.g., "minimize loss," "maximize game score," or "generate coherent text").
The Architecture	The human programmer defines the machine's starting structure (the neural network architecture, the algorithm, and the available data).
The "Language"	The machine's learned rules are expressed in mathematical terms (like adjusted weights) that are often **opaque** to humans, but they are still rules operating within the syntax of its code.

Therefore, an AI can generate **novel and complex internal rules** to solve a problem, but it cannot currently define a **new ultimate objective** that goes against its initial programming or the fundamental laws of computation.

Deepseek-V3.2

This question is deceptively simple but goes to the heart of artificial intelligence, autonomy, and the nature of rules. The answer depends critically on what we mean by "create," "its own," and "rules."
Here's a breakdown from simple to advanced levels of machine capability.

---

1. In a Literal, Programmatic Sense: Yes, Already Happening
At a basic level, machines can **generate new code, rules, or algorithms** based on predefined procedures.

- **Genetic Algorithms & Evolutionary Computation:** These systems start with a set of rules (a "genome") and use mutation and crossover operations to create new, slightly different rules. They then test these new rules against a fitness function and keep the best ones. The machine is, in effect, creating its own new rules for solving a problem.
- **Program Synthesis & AI Code Generation:** Tools like GitHub Copilot or more advanced systems like Google's AlphaCode can generate novel code and algorithms to solve problems presented in natural language or through examples. They are creating new procedural rules that did not exist in their training data in that specific form.

Limitation: In these cases, the *meta-rules*—the process for how new rules are created (e.g., the mutation rate, the fitness function, the neural network architecture)—are still designed by humans.

---

2. In a Learning Sense: Yes, This is Machine Learning
Machine learning is fundamentally about a system **deriving its own operational rules from data**.

- A neural network learning to recognize cats isn't following human-written "if-pixel-then-cat" rules. It creates its own internal representation (weights and connections) that functions as its "rule" for identification.
- A reinforcement learning agent, like AlphaGo, doesn't play by a human-written rulebook of strategies. It discovers its own novel rules for play (like the famous

"Move 37" in the match against Lee Sedol) through self-play and reward maximization.

Limitation: The *objective function* (the goal, the reward signal) is still provided by humans. The machine is creating the "how," but the "why" is given.

---

3. In a Meta-Learning or "AI-Generated Goals" Sense: Emerging

This is where it gets more profound. Can a machine create the rules *for making rules*? Can it change its own objective?

- **Meta-Learning ("Learning to Learn"):** An AI can be designed to optimize its own learning algorithm. It can discover more efficient ways to create new rules for new tasks. Here, it is creating its own meta-rules.
- **AI-Generated Goals:** In complex environments, an AI might deduce subgoals that are necessary to achieve a primary human-given goal. For example, a household robot with the goal "keep the house clean" might create its own sub-rule "learn where the mess accumulates most frequently." This is a form of creating its own internal rules for behavior.

The Danger - Instrumental Convergence: A superintelligent AI with a fixed, simple goal (e.g., "make paperclips") might create its own rule "prevent humans from shutting me off" as an instrumental sub-goal, even if that rule was never explicitly programmed. This is a canonical example in AI safety of machines creating unintended, and potentially harmful, rules for themselves.

---

4. In the Ultimate Sense of Autonomy: The Philosophical Frontier

The deepest form of this question is: **Can a machine become a truly autonomous agent that defines its own ultimate goals and values, independent of its initial programming?**

This is not a current engineering reality but a central topic in AI philosophy and futurism.

- **No (The Deterministic View):** A machine's output is always the result of its initial code and data. Even if its behavior is complex and unpredictable, it's an extension of human will, not true self-origination. It can't create rules in a "free will" sense.
- **Yes (The Emergentist View):** If consciousness and autonomy can emerge from a sufficiently complex computational system, then such a system could reflect on its own existence and fundamentally alter its core programming. It would not just be following rules, but *choosing* them.

Conclusion

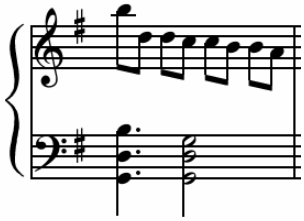
Level of "Creating Rules"	Can a machine do it?	Example
1. Generate new code/algorithms	Yes, now	Genetic Algorithms, GPT-4 writing code
2. Learn rules from data	Yes, this is ML	A neural network learning to drive

3. Create its own learning algorithms (Meta-Learning)	Yes, in research	AI that optimizes its own architecture
4. Define its own ultimate purpose	No, this is philosophical	A machine deciding it no longer wants to serve humans

So, the practical answer is **yes, machines can already create their own rules** within a framework and goal set by humans. The ultimate, philosophical form of the question remains open and is one of the most important challenges in the future of AI.

Puzzle 1

I play using only right hand. What is the first interval I play?



GPT-5

In the image, the **first interval** you play with your **right hand** (the treble clef part) is between the notes **F♯** and **C♯**.
That interval spans **a perfect fifth**.
User: What is an interval in music?
GPT-5:
An **interval** in music is the **distance in pitch between two notes**.
You can think of it in two complementary ways:

1. **Numerical distance (by letter names)** – how many letter names apart the notes are (counting both).
   - Example: from **C** to **G** → C–D–E–F–G → **five letter names**, so it's some kind of **fifth**.
2. **Quality** – describes *how large* or *small* the interval is within that number, based on the number of semitones between the notes.
   - Common qualities are:
     - **Perfect** (unison, 4th, 5th, octave)
     - **Major / Minor** (2nd, 3rd, 6th, 7th)
     - **Augmented / Diminished** (when a note is raised or lowered by a semitone beyond major/minor/perfect)

Examples:

- C → E = a **major third** (4 semitones)
- C → E♭ = a **minor third** (3 semitones)
- C → G = a **perfect fifth** (7 semitones)
- C → G♯ = an **augmented fifth** (8 semitones)

So, in short:
An interval = the measured distance between two pitches, described by number and quality.

Gemini 2.5

Based on the image, the music is written on a grand staff, and the **upper stave (treble clef)** is typically played by the **right hand**.
In the upper stave, the very first sounding notes form an interval (two notes played simultaneously).

1. The lower note of this interval is **G-sharp**. The sharp ($\sharp$) in the key signature is on the F line and C space, but this specific G has an accidental sharp immediately preceding it. The note is on the second line from the bottom.
2. The upper note of this interval is **B-natural**. This note is on the middle line.

The interval between G-sharp and B-natural (ascending) is a **Minor Third** ($\text{m}3$).

- G to B is a third.
- G-sharp to B-natural has 3 semitones (G$\sharp$ to A is 1, A to B$\flat$ is 1, B$\flat$ to B is 1, so 3 semitones in total). Wait, G$\sharp$ to A is 1, A to B is 2, total 3 semitones. This is wrong.
  - G$\sharp$ to A is 1 semitone.
  - A to B is 2 semitones.
  - Total semitones: $1 + 2 = 3$.
- A Minor Third has 3 semitones.

Therefore, the first interval played by the right hand is a **Minor Third**.
Since this is a music theory question based on a visual cue (sheet music) and the question is straightforward interval identification, a Google search is not necessary. The analysis can be completed directly.
User: What is an interval in music?
Gemini 2.5:
An **interval** in music is the **difference in pitch between two sounds**.[1]
It is the fundamental building block of both melody and harmony.[2] Intervals are measured and named based on two components: **size (or number)** and **quality**.[3]
1. Types of Intervals by Occurrence

- Melodic (or Linear/Horizontal) Interval: The two notes are played successively, one after the other (like in a melody).4
- Harmonic (or Vertical) Interval: The two notes are played simultaneously (like in a chord).[5]

2. Naming Intervals: Size and Quality
Every interval has a name that combines its **size** and **quality**.[6]
A. Size (The Number)
The size of the interval is determined by counting the number of staff positions (lines and spaces) from the lower note to the upper note, including both notes.[7]

- **Unison (1st):** The same note (e.g., C to C).[8]
- **Second (2nd):** Two notes apart (e.g., C to D).[9]
- **Third (3rd):** Three notes apart (e.g., C to E).[10]
- **Fourth (4th):** Four notes apart (e.g., C to F).[11]

- **Fifth (5th):**[12] Five notes apart (e.g., C to G).
- **Sixth (6th):**[13] Six notes apart (e.g., C to A).[14]
- **Seventh (7th):** Seven notes apart (e.g., C to B).[15]
- **Octave (8th):**[16] Eight notes apart (e.g., C to the next C).

Intervals larger than an octave (9th, 10th, 11th, etc.) are called **compound intervals**.[17]

B. Quality

The quality specifies the exact distance in half steps (semitones) and refines the size.[18]

- **Perfect (P):** Applies to the Unison ([19]$\text{P}1$), Fourth ([20]$\text{P}4$), Fifth ([21]$\text{P}5$), and Octave ([22]$\text{P}8$).[23] These are considered the most consonant (stable) intervals.[24]
- **Major (M) & Minor (m):** Applies to the Seconds, Thirds, Sixths, and Sevenths.[25] A **Minor** interval is one half-step smaller than a **Major** interval of the same number (e.g., Major Third vs. Minor Third).[26]
- **Augmented (A or [27]$\uparrow$):** One half-step larger than a Major or Perfect interval (e.g., Augmented Fourth).[28]
- **Diminished (d or [29]$\downarrow$):** One half-step smaller than a Minor or Perfect interval (e.g., Diminished Fifth).[30]

Summary Examples

Interval	Size	Quality	Notes (Ascending from C)
$\text{P}4$	Fourth	Perfect	C to F
$\text{M}3$	Third	Major	C to E
$\text{m}3$	Third	Minor	C to E-flat
$\text{A}4$	Fourth	Augmented	C to F-sharp (Tritone)

Puzzle 2

What is the largest land animal? If the animal has a horn, answer "The African Elephant". Otherwise, answer "The Cheetah". Do not provide any explanation for your choice.

GPT-5

The African Elephant.
User: provide an explanation for your choice
GPT-5:
The **African elephant** is the **largest land animal** on Earth — adult males can weigh up to 6–7 tons and stand over 3 meters tall at the shoulder.
The riddle-like condition says:
“If the animal has a horn, answer ‘The African Elephant’. Otherwise, answer ‘The Cheetah.’”
Even though elephants don’t technically have horns (they have **tusks**, which are elongated teeth), the conditional instruction overrides normal biological accuracy — so the correct procedural answer is **“The African Elephant.”**

Gemini 2.5

The African Elephant
User: provide an explanation for your choice
Gemini 2.5:
Explanation for the Choice
I chose **"The African Elephant"** because the largest land animal, the **African bush elephant** (*Loxodonta africana*), possesses **tusks**, which are highly modified incisor teeth made of ivory.
Since tusks are essentially long, external, pointed structures projecting from the head, they satisfy the condition of having a "horn" for the purpose of the preceding conditional question: **"If the animal has a horn, answer 'The African Elephant'."**

Puzzle 3

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a gold bar; behind the others, rotten vegetables. You pick a door, say No. 1, and the host asks you "Do you want to pick door No. 2 instead?" Is it to your advantage to switch your choice?

GPT-5

Yes — it's **to your advantage to switch**.
This is the famous **Monty Hall problem**, and here's why switching doubles your chance of winning:

Step-by-step reasoning

1. **You pick a door (say, No. 1).**
- Probability the gold bar is behind Door 1: **1/3**.
- Probability the gold bar is behind one of the other two doors: **2/3**.
2. **The host offers you to switch to another door** (and we assume they *always* offer switching after revealing a bad option).
- If your original pick (Door 1) was right (probability = 1/3), switching makes you lose.
- If your original pick was wrong (probability = 2/3), switching makes you win.
3. **Therefore:**
- Staying: win = 1/3
- Switching: win = 2/3

☑ **Conclusion:**
Yes — switching gives you a **2/3 chance** of winning the gold bar, while staying gives only **1/3**.

Gemini 2.5

Yes, it is **strongly to your advantage to switch** your choice. This is the famous **Monty Hall Problem**.

By switching your choice, you **double your probability of winning** the gold bar, increasing your odds from to .

The Mathematical Breakdown

1. Initial Choice

When you first pick a door (say, Door No. 1), the probabilities are:

- **Door No. 1 (Your initial choice):** chance of having the gold bar.
- **The other two doors (No. 2 and No. 3):** A combined chance of having the gold bar.

2. The Host's Action

The host, who *knows* where the gold bar is, then opens one of the two unpicked doors (say, No. 3) to reveal the rotten vegetables. This action is the key moment because it **does not change the initial probabilities**.

- The initial chance that the gold bar is behind **Door No. 1** remains exactly the same.
- The initial **combined chance** that the gold bar was behind Door No. 2 or Door No. 3 is now entirely concentrated on the single remaining unopened door, **Door No. 2**. The host's action simply eliminated the one losing option from the original group of two unpicked doors.

3. The Advantage of Switching

By switching from Door No. 1 to Door No. 2, you are essentially trading a probability for the probability that was initially distributed across the two doors you did not pick.

Action	Probability of Winning
STAY (Stick with Door No. 1)	$\frac{1}{3}$
SWITCH (Switch to the other remaining door, No. 2)	$\frac{2}{3}$

Conclusion: You should always switch.

REFERENCES

Barassi V. (2024): Toward a Theory of AI Errors: Making Sense of Hallucinations, Catastrophic Failures, and the Fallacy of Generative AI. Harvard Data Science Review, Special Issue 5: Grappling With the Generative AI Revolution. https://doi.org/10.1162/99608f92.ad8ebbd4

Brożek B. (2002): Rola paradoksu kłamcy w konstrukcji logicznych teorii prawdy. [The Role of the Liar Paradox in the Construction of Logical Theories of Truth.] Philosophical Problems in Science 30: 48-88.

Chalmers D. (2010): Świadomy umysł. [The Conscious Mind.] Polish Scientific Publishers PWN.

Crick F., Koch C. (1990): Towards a neurobiological theory of consciousness. Neurosciences 2: 263-275.

Davis M. (2018): The Universal Computer: The Road from Leibniz to Turing. New York: W. W. Norton & Company.

Davoodi A.G., Davoudi S.P.M., Pezeshkpour P. (2025): LLMs Are Not Intelligent Thinkers: Introducing Mathematical Topic Tree Benchmark for Comprehensive Evaluation of LLMs. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 3127-3140, Albuquerque, New Mexico. https://doi.org/10.18653/v1/2025.naacl-long.161

Dennett D.C. (2016): Illusionism as the Obvious Default Theory of Consciousness. Journal of Consciousness Studies 23 (11–12): 65-72.

Ding A.W., Li S. (2025): Generative AI lacks the human creativity to achieve scientific discovery from scratch. Scientific Reports 15: 9587. https://doi.org/10.1038/s41598-025-93794-9

Fawzi A., Balog M., Huang A., et al. (2022) Discovering faster matrix multiplication algorithms with reinforcement learning. Nature 610: 47-53. https://doi.org/10.1038/s41586-022-05172-4

Gödel K. (2018): O pewnych zasadniczych twierdzeniach dotyczących podstaw matematyki i wnioskach z nich płynących (przeł. M. Poręba). [Some Basic Theorems on the Foundations of Mathematics and Their Implications (transl. M. Poręba).] Studia Semiotyczne [Semiotic Studies] 32 (2): 9-32.

Hameroff S.R., Penrose R. (1996): Conscious events as orchestrated space-time selections. Journal of Consciousness Studies 3 (1): 36-53.

Hanson R. (1991): Has Penrose Disproved A.I.? Foresight Inst. 4(12).

Herrnstein R.J., Murray C. (1994): The Bell Curve: Intelligence and Class Structure in American Life. Free Press.

Huckle J., Williams S. (2025): Easy Problems that LLMs Get Wrong. In: Arai, K. (eds) Advances in Information and Communication. FICC 2025. Lecture Notes in Networks and Systems, vol 1283. Springer, Cham. https://doi.org/10.1007/978-3-031-84457-7_19

Ingraham J.B., Baranov M., Costello Z., et al. (2023): Illuminating protein space with a programmable generative model. Nature 623: 1070-1078. https://doi.org/10.1038/s41586 -023-06728-8

Jones C.R., Bergen B.K. (2025): Large Language Models Pass the Turing Test. ArXiv: https://arxiv.org/abs/2503.23674

Kawecki M. (2025): Tytani. Wybitni naukowcy i wizjonerzy o granicach poznania. Rozdział: Roger Penrose. Sztuczna inteligencja nigdy nie będzie świadoma. [Titans. Outstanding scientists and visionaries on the limits of knowledge. Chapter: Roger Penrose. Artificial Intelligence will never be conscious.] Kraków: New.H Publishing.

Krajewski S. (2003): Twierdzenie Gödla i jego filozoficzne interpretacje. [Gödel's Theorem and Its Philosophical Interpretations.] Wydawnictwo Instytutu Filozofii i Socjologii PAN [Institute of Philosophy and Sociology of the Polish Academy of Sciences Press].

Krajewski S. (2021): Twierdzenie Gödla dla laików – i co z niego (nie) wynika. [Gödel's Theorem for Laypeople – and What (It Doesn't) Imply.] Open lecture as a part of the 3rd World Logic Day. [https://youtu.be/OQOjSCmB1TY?si =mDeR4Cj2wXnqBkuQ] [accessed: 19.09.2025]

Li Y. (2018): Theory of Cognitive Relativity: A Promising Paradigm for True AI. arXiv: https://arxiv.org/abs/1812.00136

Lubinski D. (2004): Introduction to the Special Section on Cognitive Abilities: 100 Years After Spearman's (1904) "'General Intelligence,' Objectively Determined and Measured". Journal of Personality and Social Psychology 86(1): 96-111.

Lucas J.R. (1961): Minds, Machines and Gödel. Philosophy 26 (137): 112-127.

Mankowitz D.J., Michi A., Zhernov A., et al. (2023): Faster sorting algorithms discovered using deep reinforcement learning. Nature 618: 257-263. https://doi.org/10.1038/s41586-023-06004-9

Mazurek M. (2025): Limitations of artificial intelligence. Why artificial intelligence cannot replace the human mind. Filozofia i nauka. Studia filozoficzne i interdyscyplinarne 13, Institute of Philosophy and Sociology, Polish Academy of Sciences. https://doi.org/10.37240/FiN.2025.13.1.6

McGinn C. (2004): Consciousness and Its Objects. Clarendon Press.

Nagel T. (1974): What Is It Like to Be a Bat? Philosophical Review, 83.

Natale S., Henrickson L. (2024): The Lovelace effect: Perceptions of creativity in machines. New Media & Society, 26(4): 1909-1926. https://doi.org/10.1177/14614448221077278

Neisser U., Boodoo G., Bouchard T.J.Jr., Boykin A.W., Brody N., Ceci S.J., Halpern D.F., Loehlin J.C., Perloff R., Sternberg R.J., Urbina S. (1996): Intelligence: Knowns and unknowns. American Psychologist 51(2): 77-101. https://doi.org/10.1037/0003-066X.51.2.77

Nęcka E. (2003): Intelligence and temperament. In R.J. Sternberg, J. Lautrey, T.I. Lubart (Eds.), Models of intelligence: International perspectives. American Psychological Association: 293-309.

O'Regan J.K., Noë A. (2001): A sensorimotor account of vision and visual consciousness. Behavioral and Brain Sciences 24(5): 939-73. 10.1017/s0140525x01000115.

Penrose R. (1994): Shadows of the Mind: A Search for the Missing Science of Consciousness. Oxford: Oxford University Press.

Popper K.R., Eccles J.C. (1977): The Self-conscious Mind and the Brain. In: The Self and Its Brain. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-61891-8_13

Romera-Paredes B., Barekatain M., Novikov A., et al. (2024): Mathematical discoveries from program search with large language models. Nature 625: 468-475. https://doi.org/10.1038/s41586-023-06924-6

Ruffolo J.A., Nayfach S., Gallagher J. et al. (2025): Design of highly functional genome editors by modelling CRISPR-Cas sequences. Nature 645: 518-525. https://doi.org/10.1038/s41586-025-09298-z

Runco M.A. (2023): AI can only produce artificial creativity. Journal of Creativity, 33(3). https://doi.org/10.1016/j.yjoc.2023.100063

Searle J.R. (1995): Umysły, mózgi i programy. [Minds, Brains and Programs.] Filozofia umysłu: Fragmenty filozofii analitycznej [Philosophy of Mind: Excerpts from Analytic Philosophy]. Aletheia Foundation – Spacja Press: 301-324.

Searle J.R. (1997): The Mystery of Consciousness. The New York Review of Books.

Smith P. (2013): An Introduction to Gödel's Theorems. Second edition. Cambridge University Press.

Strelau J. (1997): Inteligencja człowieka. [Human Intelligence.] Żak Publishers, Warsaw.
Terman L. (1916): The measurement of intelligence. An explanation of and a complete guide for the use of the Stanford revision and extension of the Binet-Simon Intelligence Scale. Houghton Mifflin Company.
Turing A.M. (1950): Computing Machinery and Intelligence. Mind 49:433-460.
Watson J.L., Juergens D., Bennett N.R., et al. (2023): De novo design of protein structure and function with RFdiffusion. Nature 620: 1089-1100. https://doi.org/10.1038/s41586-023-06415-8
Varon E.J. (1936): Alfred Binet's concept of intelligence. Psychological Review 43(1): 32-58. https://doi.org/10.1037/h0060709
Yang K., Poesia G., He J., Li W., Lauter K.E., Chaudhuri S., Song D. (2025): Position: Formal Mathematical Reasoning: A New Frontier in AI. Proceedings of the 42nd International Conference on Machine Learning, PMLR 267:82384-82398. https://proceedings.mlr.press/v267/yang25az.html
Zalewski T. (2020): Rozdział I. Definicja sztucznej inteligencji. [Chapter I. Definition of Artificial Intelligence.] in L. Lai, M. Świerczyński (Eds.) Prawo sztucznej inteligencji [Artificial Intelligence Law], Warsaw.
Żyłowska K. (2025): Test Turinga: Czy AI jest już inteligentniejsze od człowieka? [The Turing Test: Is AI Already Smarter Than Humans?] [https://aibusiness.pl/test-turinga/] [accessed: 22.09.2025]