https://doi.org/10.48417/technolang.2026.03.05
Research article

Covert Systems – On the Intelligence of Large Language Models

David Link (✉)
Independent Researcher
alephsoixante@gmail.com

Abstract

In light of recent developments in Artificial Intelligence, particularly the development of Large Language Models (LLMs), the question if machines can be said to think is examined. The problem was already addressed by Alan Turing in the 1950s, laying the groundwork for our argumentation. Our text refutes four common objections. Firstly, it challenges Ada Lovelace's claim that computers are incapable of creating truly original content, demonstrating that neural networks, such as Transformer models, independently generate novel structures. Secondly, the assertion that LLMs merely statistically predict the next word is discussed. They are shown to be capable of constructing compressed internal representations from their input data through nonlinear operations, distinguishing them from statistical models. A third argument concerns AI's lack of sensory access to the world. However, as Turing has demonstrated, the development of mental capabilities requires only communication between a teacher and a student, and sensory experience can be technically replicated at any time. The Chinese Room thought experiment, which suggests that algorithms merely simulate intelligence through formal symbol manipulation, is already anticipated by the Turing Test, which posits that any machine successfully feigning mental functions should be deemed to possess them. The text explores how LLMs generate a perfect compression of their input data, resulting in an accurate model of the world, a convincing simulation of human intelligence, and mastery of language. It discusses the phenomenon of "emergence", the appearance of higher-order cognitive capabilities with increasing amounts of training data. Although the successes in implementing intelligence are notable, the neural network methodology employed results in the enigmaticity of the underlying basis, failing to illuminate the true nature of the mind as originally hoped. This obscurity was already evident in the foundational works of Kurt Gödel and Alan Turing, highlighting the complexities inherent in understanding human and machine intelligence alike.

Keywords: Large Language Models; Mechanical Intelligence; Emergence; Perfect Compression; Enigmaticity

Citation: Link, D. (2026). Covert Systems – On the Intelligence of Large Language Models. *Technology and Language*, *7*(3), 68-85. https://doi.org/10.48417/technolang.2026.03.05

© Link, D. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

УДК 1:004.81
https://doi.org/10.48417/technolang.2026.03.05
Научная статья

Скрытые системы: Об интеллекте больших языковых моделей

Дэвид Линк (✉)
Независимый исследователь
alephsoixante@gmail.com

Аннотация
В свете последних достижений в области искусственного интеллекта, в особенности развития больших языковых моделей, рассматривается вопрос о том, можно ли утверждать, что машины мыслят. Эта проблема уже ставилась Аланом Тьюрингом в 1950-х годах и заложила основу для аргументации. В тексте опровергаются четыре распространённых возражения. Во-первых, оспаривается утверждение Ады Лавлейс о неспособности компьютеров создавать подлинно оригинальное содержание: демонстрируется, что нейронные сети, такие как модели трансформеры, самостоятельно порождают новые структуры. Во-вторых, обсуждается утверждение, что большие языковые модели лишь статистически предсказывают следующее слово. Показывается, что они способны строить сжатые внутренние репрезентации входных данных посредством нелинейных операций, что отличает их от статистических моделей. Третий аргумент касается отсутствия у ИИ сенсорного доступа к миру. Однако, как показал Тьюринг, для развития мыслительных способностей достаточно коммуникации между учителем и учеником, а сенсорный опыт может быть технически воспроизведён в любой момент. Мысленный эксперимент “Китайская комната”, согласно которому алгоритмы лишь имитируют интеллект посредством формальной манипуляции символами, уже предвосхищается тестом Тьюринга, утверждающим, что любая машина, успешно имитирующая мыслительные функции, должна считаться обладающей ими. В тексте исследуется, как большие языковые модели создают идеальное сжатие входных данных, что приводит к точной модели мира, убедительной имитации человеческого интеллекта и владению языком. Обсуждается феномен “эмерджентности” – появления когнитивных способностей высшего порядка по мере увеличения объёма обучающих данных. Хотя успехи во внедрении интеллекта значительны, применяемая методология нейронных сетей приводит к загадочности лежащей в основе базы, не позволяя прояснить истинную природу разума, как изначально предполагалось. Эта непрозрачность была уже очевидна в основополагающих работах Курта Гёделя и Алана Тьюринга, подчёркивая сложности, присущие пониманию как человеческого, так и машинного интеллекта..

Ключевые слова: Большие языковые модели; Механический интеллект; Эмерджентность; Идеальное сжатие; Загадочность

Для цитирования: Link, D. Covert Systems – On the Intelligence of Large Language Models // Technology and Language. 2026. № 7(3). P. 68-85. https://doi.org/10.48417/technolang.2026.03.05

© Линк Д. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

INTRODUCTION

Approximately 75 years ago, the British mathematician and theoretical inventor of the computer, Alan Turing, published a visionary article entitled "Computing Machinery and Intelligence." This text addressed a question that has been, and remains, controversially debated: Can machines think? Given the difficulty of defining intelligence, Turing proposed an alternative approach, known as the imitation game: A computer and a human are interviewed by an interrogator via a terminal – at that time still a teletype machine – and the interrogator must identify the participants based on the conversation. The human is intended to assist the interviewer in this task, while the machine should persuade them to make an incorrect identification.

Turing assumed that, within about fifty years, it would be possible "to programme computers […] to make them play the imitation game so well that an average interrogator will not have more than 70 per cent. chance of making the right identification after five minutes of questioning" (Turing, 1950, p. 442). However, in a BBC panel discussion held around the same time, he stated that it would certainly take at least another hundred years before the machine could win the game (Newman et al., 1952, p. 6). Averaging both estimates yields the 75 years it actually took to reach this milestone, demonstrating the remarkable accuracy of Turing's prediction.

The 2020s were marked by epochal breakthroughs in Artificial Intelligence (AI), particularly the development of Large Language Models (LLMs). In 2024, the GPT-4 system passed a Turing test at the University of San Diego, where 54 percent of participants mistook it for a human after a five-minute conversation (Jones & Bergen, 2025). Turing's remarks highlight that, in his time, the question of whether machines could think seemed so scandalous and absurd, given that the first computers were just being developed, that any reasonable person would have denied it. In those days, experience was primarily with machines that processed material things, not mere symbols. One of the earliest computers, the Manchester Baby, performed its first successful calculation in June 1948.

Turing also assumed that "at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted" (Turing, 1950, p. 442). In this respect, he was too optimistic. Today, despite mounting evidence that machines have learned to think, stereotypical arguments that were already widespread in Turing's time are still being repeated, denying the intelligence of computers and diminishing the uniqueness of this technological breakthrough.

A significant portion of Turing's essay is devoted to refuting these objections, which can inform our attempt to argue that machines can indeed think. Ultimately, as the British mathematician stated, all counterarguments can be traced back to the theological view that "[t]hinking is a function of man's immortal soul" and therefore "no animal or machine can think" (Turing, 1950, p. 443). AI threatens humanity's position as the crown of creation, with its unique distinction of language and thought

THE ORIGINALITY OF NEURAL NETWORKS

Four main objections to Artificial Intelligence can be distinguished. Firstly, it is often claimed that AI cannot produce anything original or novel, merely copying text examples from humans available on the internet and collaging them together in its output. As early as 1843, Ada Lovelace wrote in her treatise on Charles Babbage's "Analytical Engine": "The Analytical Engine has no pretensions whatever to *originate* any thing. It can do whatever *we know how to order it* to perform. [...] Its province is to assist us in making *available* what we are already acquainted with" (Menabrea, 1843, p. 24, note G. Emphasis A. L.).

Turing quotes this objection and counters it with his personal experience of often being surprised by machines, as well as the observation that not all consequences of a given fact are always foreseeable. Its elaboration could therefore lead to the unexpected. He also refers to his conception of machines that learn through reward and punishment and can thus develop in unforeseen directions (Turing, 1950, p. 450–451).

To address this criticism regarding Large Language Models, it is necessary to consider a brief history of neural networks. The models are essentially special "Deep Neural Networks" (DNNs) consisting of numerous layers through which input data flows and is processed. However, the Transformer architecture of the LLMs is significantly deeper and larger than conventional DNNs, with the GPT-4 model estimated to have over 1.8 trillion trainable parameters across 120 layers.[1]

A simple neural network consists of two layers of artificial neurons or nodes connected by a hidden intermediate layer of further nodes, with all elements interconnected (cf. Fig. 1). Each artificial neuron has a weight by which the incoming numerical value is multiplied, determining how much of the input signal is passed on. This structure represents a linear function. Additionally, there are nonlinear layers that, for example, only allow the input signal to pass through if it exceeds a certain threshold.

To train a neural network, data is fed into its input, and the output is observed. By comparing the numerical values with the desired target quantities, the so-called "loss" is calculated. Through a complex calculation known as "backpropagation of errors," the weights of the nodes in each layer are gradually adjusted and optimized so that the output values increasingly approximate the desired results (Rashid, 2016, p. 73–82).

Artificial neural networks were introduced in 1943 by the US-American neurophysiologist Warren McCulloch and the logician Walter Pitts in their seminal paper "A Logical Calculus of the Ideas Immanent in Nervous Activity." This work demonstrated that a network of simple artificial neurons can perform all logical functions, receiving input signals, performing weighted summation, and firing an output signal based on a threshold function (McCulloch & Pitts, 1943).

[1] OpenAI does not publish any figures on this.

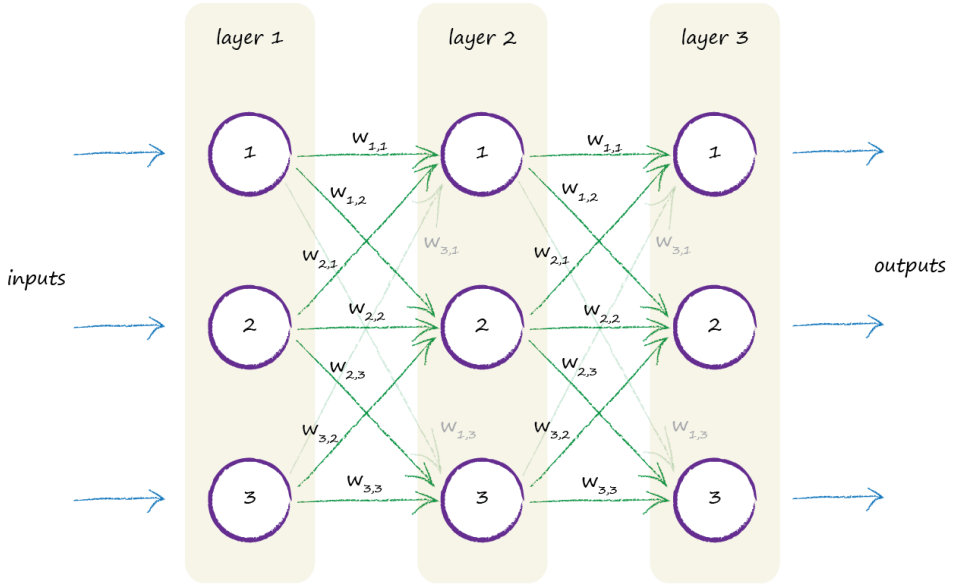

Figure 1. Diagram of a simple neural network (Rashid, 2016, p. 49)

Interestingly, as early as 1948, Alan Turing designed "unorganized machines" made of artificial neurons, intended to become universal machines through training. However, his neurons lacked a threshold function and were more akin to finite-state automata (Turing, 1948, p. 15–17).

In 1958, the US-American psychologist and computer scientist Frank Rosenblatt implemented these theoretical considerations practically in a program on the IBM 704 computer, along with special hardware, the so-called "Perceptron" (cf. Fig. 2). Already here, correct answers were positively reinforced and wrong answers "punished" by propagating the loss back to the individual weights of the artificial neurons. However, the Perceptron only had a single layer across which this difference was distributed.

Rosenblatt summarized his experiences and experiments: "Such a system will form similarity concepts based on the basis of *coincident areas* of stimuli, rather than by the similarity of contours or outlines." He observed that a simple Perceptron made of artificial neurons practically incorporated experience: "[T]he same response will have a stronger tendency to recur, and learning can be said to have taken place." Rosenblatt quantified this knowledge gain using a numerical value indicating the probability of a "correct generalization," which remained effective even when the system was presented with new examples, such as squares, circles, and triangles. The recognition rate could be optimized by increasing the number of artificial neurons (Rosenblatt, 1958, p. 390, p. 392, p. 396, p. 400. Emphasis F. R.).

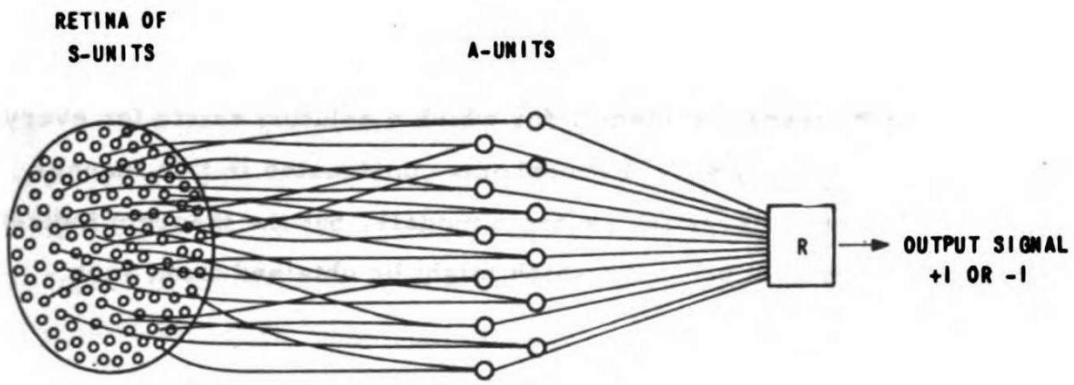

Figure 2. Diagram of a simple Perceptron (Rosenblatt, 1962, p. 99)

Lovelace's objection presupposes that machines are entirely programmed by explicit rules. This assumption does apply to classical computer programs, such as those intended to be run on the "Analytical Engine." In contrast, neural networks then and can from perceived data and develop novel abilities and structures, resulting in behavior not explicitly programmed. They form generalizations of the input material, enabling correct classification by extracting characteristic features from the data and forming a kind of "schematism", similar to Immanuel Kant's "schematisms of the pure concepts of understanding."[2]

In the *Critique of Pure Reason*, these schematisms facilitate the subsumption of manifold appearances in the sensory world, such as individual dogs or different triangles, under corresponding concepts, despite their differences. Kant writes: "The concept of a dog signifies a rule in accordance with which my imagination can specify the shape of a four-footed animal in general, without being restricted to any single particular shape that experience offers me." And further: "This schematism of our understanding […] is a hidden art in the depths of the human soul, whose true operations we can divine from nature and lay unveiled before our eyes only with difficulty" (Kant, 1781/1998, B 176–B187). Similarly, the weights of neural networks reside in a so-called hidden layer, which, although readable, are not intuitively understandable due to their complex numerical nature. Which classification schemes and features the system has gleaned from the input data remains concealed.

The network does not form an average of all seen examples, but instead constructs a characteristic, compressed structure by adjusting the internal weights under the pressure of necessary loss minimization, a structure that still contains the crucial details. As the Perceptron demonstrated, the generalization even extends to correctly classifying future data not included in the training material. The network develops this optimal method independently, forming a high-dimensional structure of statistical relationships that

[2] The analogy here is meant functionally, not that the structures themselves correspond.

contains "the riches of the particular" and represents an almost lossless compression (Hegel, 1832/2010, p. 37).[3]

The learning capability of these algorithms refutes Lovelace's thesis that machines can only accomplish what they are programmed to do, as they develop the necessary methods themselves based on given objectives. However, it is worth noting that she was only able to study the "Analytical Engine" theoretically, as Babbage did not complete his work. Had Lovelace experienced firsthand the surprises that algorithms can produce, her perspective might have been different. When used for content generation instead of classification, these programs produce unseen yet meaningful combinations.

STATISTICS OF THOUGHT

A second objection to Large Language Models is that they rely solely on statistics and therefore lack true intelligence, possessing no knowledge about the world but simply predicting the next likely word ("next-token prediction"). This criticism was exemplified by a scandal at Google in 2021 involving the text "On the Dangers of Stochastic Parrots," where AI researchers reduced early LLMs to "stochastic parrots" that statistically repeated the contents of data sets from the training material (Bender et al., 2021).[4] The fact that probability calculations aim for average values is also often used to claim that the results of these algorithms are always mediocre and predictable.

Statistics is a collection of mathematical methods that enables the approximate quantitative estimation of properties and events occurring on a mass scale. Initially, it may seem ill-suited for understanding language's qualitative properties, as probability calculations can only provide mean values – for instance, a text with word frequencies corresponding to the average in a certain corpus.

However, the US-American mathematician Claude Shannon demonstrated in his 1948 *Mathematical Theory of Communication*, building on the work of the Russian mathematician Andrei A. Markov, that statistics can also generate more interesting results. By working with transition probabilities, where each word stochastically depends on one or more previous words, Shannon produced the following text: "THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED" (Shannon & Weaver, 1948/1963, p. 44).

This approach allowed him to develop a calculation of the information content of messages, leading him to conclude: "Frequently the messages have *meaning* [...]. These semantic aspects of communication are irrelevant to the engineering problem" (Shannon & Weaver, 1948/1963, p. 31. Emphasis C. S.). Experience with Markov chains showed that considering the connections between words can help approach the semantic level of

[3] In the large version of the GPT-4 model, for example, the vector space in which the terms are represented has 3,072 dimensions, which intuitively corresponds to the number of categories according to which classification can be made.
[4] As a result, one of the authors had to leave Google after a dispute.

texts. Shannon's example illustrates that it makes a significant difference on which basis the next item is predicted.

Transformer models are neural networks that analyze the relationships between words in a unique way. In the attention mechanism, the AI evaluates connections between various aspects of elements in the input material in parallel within very large contexts and takes into account both syntactic and semantic features. The model learns which neighboring words to include in order to understand the current character sequence (Vaswani et al., 2017). These relations leave their marks on the terms themselves, the so-called "embeddings". The dense vector representations are highly semantic, enabling calculations like "king - man + woman" to yield "queen" as a result (Mikolov et al., 2013, p. 749). Notably, the AI develops this property independently, without explicit programming. The authors of a research paper on the topic commented with surprise: "Remarkably, since the model is not trained for this task, the relational structure of the vector space appears to be an emergent property" (Vylomova et al., 2016, p. 1).

The generality produced by neural networks differs from that of statistics, which relies on average values and expected deviations, representing most examples seen, their means. Neural networks, in contrast, map the input material in its specificity, with their generality also encompassing future examples unknown to the system. This makes them more precise than approximate statistics, when, for instance, predicting the color of cats. Probability calculations require an area for which they are calculated (size, color, etc.) and explicitly defined variables. In contrast, neural networks develop optimal analysis criteria independently, which remain hidden.

Statistics is characterized by linear calculations. To find out how many of 100 dice rolls result in a six, the number of events is multiplied by the probability of a six, one-sixth, and equals approximately 16.6. Other methods in this mathematical discipline, like linear and logistic regression or variance analysis, are also marked by relatively small numbers of parameters and clearly interpretable, linear relationships, with comparatively limited amounts of data considered.

Neural networks, on the other hand, essentially contain nonlinear operators, mimicking the thresholds of biological neurons and permitting the algorithm to extract knowledge from examples seen. These activation functions enable the creation of deep neural networks, as they prevent linear operations in the different layers from combining into a single one. This allows models to learn complex patterns and not just linear relationships.

The activation functions in neural networks are similar to human deduction, enabling the formation of "layered representations" where different aspects of input material are analyzed at each level. The depth and nonlinear activation operations increase the expressiveness of the models, as evidenced by the "Universal Approximation Theorem", which shows that a simple network consisting of a linear layer and a nonlinear activation function can approximate any continuous function with arbitrary accuracy if it is sufficiently large. This demonstrates the universal versatility and expressivity of these algorithmic structures, suggesting that if there are mathematical operations that accurately describe the relationships between words, then suitable neural networks can find and represent them (Hornik et al., 1989).

Transformer models draw on functions from various mathematical areas, including linear algebra, analysis, numerical optimization, data modeling, and computer science, in addition to probability calculation. They consist of more than just statistics, with matrix multiplication being a key mathematical method that shapes these programs, for which the hardware is optimized.

The objection that the algorithm only predicts the next token and is therefore not intelligent is misleading, as people also formulate one word after another. With similar justification, one could argue that a mere combination of dead character sequences cannot produce intelligent text or convey thoughts, or that language is simply the ejection of any sounds. In both cases, this reductionist rhetoric completely misses the intelligent content of the action. What is crucial is the type of statistical or other method used, its purpose, and the material it is based on. In principle, it is impossible to determine what results statistical operations can produce.

The main question regarding Transformer models, as with Shannon, is the basis on which they predict the next word. To correctly model this probability, the network must learn deep regularities of language, including grammar, semantic relationships, world knowledge, discourse structure and logical connections. Predicting the next token is therefore not a superficial goal, but rather a training signal that forces AI to capture complex structures.

The crucial difference to classical statistics lies in the internal representations created autonomously by the model, such as the high-dimensional vector spaces that allow for the representation of word meanings and various syntactic and semantic relationships. The network forms compressed structures that also enable the understanding of abstract concepts or hierarchical relationships, allowing it to formulate and understand analogies, make generalizations, and interpret contexts. The alternative to statistics would be completely rule-based intelligence – an approach that has proven inadequate in the development of machinic intelligence.[5]

THE IMMEDIACY OF EXPERIENCE

Another objection to the intelligence of AI systems is that they lack access to the world through sensory organs. Martin Warnke argues in his work *Large Language Kabbala* that "intelligence requires a connection to the world, a body, but also the community of people" (Warnke, 2026, p. 128. Translation D. L.). Without these, thinking would be impossible, and algorithmic output merely meaningless, syntactic wordplay.

In response to this objection, Alan Turing, in his 1948 essay "Intelligent Machinery," humorously imagines constructing a "thinking machine" by replacing all parts of a human being with apparatuses, such as television cameras, microphones, loudspeakers, wheels, a handling servomechanism, and an "electronic brain." Given the technological capabilities of the time, he describes this project as a "tremendous undertaking […] of immense size" and continues:

[5] Cf. the history of so-called symbolic AI and the "expert systems" that dominated research from the 1950s to the 1980s and are now largely considered to have failed.

> In order that the machine should have a chance of finding things out for itself it should be allowed to roam the countryside, and the danger to the ordinary citizen would be serious. Moreover even when the facilities mentioned above were provided, the creature would still have no contact with food, sex, sport and many other things of interest to the human being. (Turing, 1948, p. 16–17).

Although this method may be the safest way to produce a thinking machine, Turing finds it too slow and impractical overall. In his thought experiment, he concludes that simulating a pure brain with rudimentary sensory organs would be sufficient to implement intelligence. He lists various intellectual domains in which such a system could operate (Turing, 1948, p. 16–17).

In a passage written two years later, he states:

> It will not be possible to apply exactly the same teaching process to the machine as to a normal child. It will not, for instance, be provided with legs, so that it could not be asked to go out and fill the coal scuttle. Possibly it might not have eyes. But however well these deficiencies might be overcome by clever engineering, one could not send the creature to school without the other children making excessive fun of it. It must be given some tuition (Turing, 1950, p. 456).

And, citing Helen Keller's example, Turing draws the radical conclusion that "education can take place provided that communication in both directions between teacher and pupil can take place by *some means or other*" (Turing, 1950, p. 456. Emphasis D. L.). Intelligence is based on the capacity to process information, learn, and communicate, rather than sensory perception or other bodily preconditions.

Helen Keller, born in Alabama, USA, at the end of the 19th century, became deaf and blind after an illness in early childhood. At the age of seven, her teacher, Anne Sullivan, taught her language by spelling the names of touched objects into her hand using a special finger alphabet. This eventually led to the child's groundbreaking discovery that objects are named after specific words. Keller later acquired multiple languages, studied at college, and received several honorary doctorates. She also lectured and authored many books. Despite her sensory limitations, she developed complex linguistic abilities and abstract thinking in ways that surpassed many of her contemporaries.

To address the question of how we relate to the world, Turing employs a two-pronged strategy: He demonstrates that sensory organs are not necessary for acquiring intelligence, and he shows that they can be technically implemented if needed, for example, through cameras. Sensory access to the world is therefore not a necessary condition for intelligence, but rather a possible extension.

At the end of the article, he expresses the view that the most promising path to an intelligent machine would be to provide it "with the best sense organs that money can buy, and then teach it to understand and speak English" (Turing, 1950, p. 460). He was probably thinking of instructing the mechanical child by showing it various objects and naming them. Once it had learned the language, an excellent means of communication would be available to teach it to think in a comprehensive sense.

Since Immanuel Kant, it has been recognized that our supposedly immediate relationship to the world through the body and senses is mediated, like other forms of

experience, by forms of intuition, such as space and time, as well as by the construction of these organs. In this context, language plays a central role in providing privileged access to reality and human society. It makes no essential difference whether a robot experiences firsthand that an egg is fragile, or whether an AI deduces this fact from texts.

Humans possess extensive knowledge of things they have never directly perceived, including astronomical or historical events, microscopic structures, and distant places. This knowledge is conveyed purely through spoken or written words, highlighting the significance of language in shaping the content of our minds. In many areas, human understanding already consists of secondary experience transmitted through words, books, or instruction. If direct access to the world were the only means of enabling genuine mental capabilities, then deaf-blind or paraplegic individuals would have to be significantly less intelligent than others. However, this is not the case, as exemplified by Helen Keller and Stephen Hawking. Furthermore, Large Language Models can provide detailed answers to questions about seemingly immediate experiences they have never had, such as what happens when you put your hand on a hot stove.

FEIGNED THINKING

Some authors argue that AI is not actually intelligent but only feigns intelligence. The most well-known form of this objection is the Chinese Room thought experiment by the US-American philosopher John Searle. In this experiment, a person who does not speak Chinese answers questions posed in Chinese characters with meaningful answers in Chinese characters, guided by instructions in their native tongue. Searle's argument is that while the human in the room appears to be fluent in the language, they are only performing the formal manipulation of symbols (Searle, 1980). This objection often combines with the observation that the machine has no awareness of what it is thinking and that without awareness, intelligence is impossible. The paradox of the Chinese Room experiment, however, lies in the fact that the outputs of the system as a whole are perfect Chinese, rendering the technical means by which this is achieved irrelevant.

In order to answer Searle's criticism, one can go back to Turing's Imitation Game which in 1950 already provided the resources for doing so. It replaces a very unclear and complicated theoretical question – Can machines think? – with a verifiable or falsifiable test procedure. This approach eliminates the need for potentially complex discussions and enables a scientific treatment of the problem. Since Turing, it no longer makes sense to speak of a machine that merely simulates cognition; if it effectively and verifiably feigns this ability and if one cannot specify what is left out by mere feigning, it should be considered intelligent. Even with human counterparts, we cannot know for certain whether they actually possess reason. All that is accessible to us is their corresponding behavior.

The Turing test counteracts the human tendency to animistically ascribe a soul to inanimate objects. By requiring the interviewer to assume deception by one party, they will exhibit healthy skepticism toward the statements made and will not assume intelligence where none exists.

Searle either rejects or fails to understand the Turing test due to his assumption that mind and brain are inextricably linked. He assumes that only the complex workings of the organ of thought can produce genuine intelligence. However, the existence of knowledgeable books demonstrates that essential brain content can be externalized, challenging Searle's stance. He would deny intelligence to even an algorithm capable of writing such works or solving mathematical problems, adopting a dogmatic position that software can only feign content; in reality, it is always just formal processing (Feng et al., 2026[6]).

During Searle's time, symbolic AI was far from demonstrably implementing cognitive abilities and competed for funding with exaggerated promises. Researchers were content with superficial parallels between algorithms and reality and developed programs that simulated content through formal processing without understanding anything essential. For example, the software "Parry" claimed to display emotions. In fact, the feelings were "simulated" by simply changing the numerical value of correspondingly named variables (Link, 2006, p. 66).

Under the pressure of minimizing loss, modern AI systems autonomously develop an internal structure that enables them to engage in intelligent discussions on a wide variety of topics, using appropriate language. They attempt to find a generalization that represents the extensive textual material with as little loss as possible. Due to the numerical complexity, the precise structure of this compression is difficult to determine, much like the mental substrate that allows humans to express themselves. Current research shows that the internal knowledge of AI is fragmented, meaning that a complete worldview is not present at any single location. However, when the system plays a board game, the positions of the pieces are recognizably represented by its weights and can be manipulated externally to change the machine's moves. A similar phenomenon occurs with geographical questions, where models appear to create topological maps of locations around the world (Li et al., 2022).

The internal structure at the heart of AI remains enigmatic. It could be that intelligence emerges here because it is the most efficient compression strategy for the entirety of linguistic utterances. AI may develop these world "models" because the optimal way to predict human speech is to simulate the reality that produces it. Whether the stored form is the same as in the brain is questionable, but the decompressed content and function are very similar.

In the perfect compression, a complete and accurate "world" model, a convincing simulation of human intelligence, and Saussure's "langue" – the structure that precedes and determines every specific verbal utterance – converge. Saussure defines this as "the norm of all other manifestations of speech […]. It is both a social product of the faculty of speech and a collection of necessary conventions that have been adopted by a social body to permit individuals to exercise that faculty" (Saussure, 1931/1966, p. 9). Today's models have not yet reached this ideal convergence, but with each new generation, they are getting closer.

[6] The authors report here on the research agent "Aletheia", which autonomously or semi-autonomously developed solutions for several previously unsolved mathematical problems.

UNEXPECTED EMERGENCES

The development of Large Language Models has been marked by major surprises, discussed under the heading of "emergence." Initially, the goal was not to program a thinking system. Rather, the implementation of true machine intelligence, like many others, was an unplanned invention. As the British science fiction author and physicist Arthur C. Clarke noted, any "sufficiently advanced technology" is "indistinguishable from magic" (Clarke, 1968). Working with the material, in this case algorithms, often produces surprises in scientific or technical endeavors that exceed expectations and were not originally intended. The technical forms used to produce these inventions have a life of their own, contributing to the realization and further development of the ideas.

In modern AI, implemented intelligence appears to grow discontinuously, in leaps and bounds. Researchers at Microsoft had already diagnosed the emergence of the aforementioned spatial semantics of word embeddings in older language models in 2013. "Perhaps more surprisingly, we find that this [the representation of meaning through specific vector distances, D. L.] is also the case for a variety of semantic relations" (Mikolov et al., 2013, p. 746). At the time of the influential 2017 publication of "Attention is All You Need," which established the Transformer models and led to the breakthrough of LLMs, the focus was primarily on creating better translation programs (Vaswani et al., 2017).

With GPT-1 (2018), this evolved into an algorithm capable of imitating grammar and generating simple texts. These models were aptly described as "stochastic parrots." As they were scaled up and fed more data, the quality of the output changed unexpectedly. The AI suddenly possessed hundreds of previously unavailable skills, such as solving untrained tasks ("zero-shot learning"), generating programming code, calculating mathematical functions, and making logical deductions. "Emergence is when quantitative changes in a system result in qualitative changes in behavior" (Wei et al., 2022, p. 1–2).[7]

With the GPT-3 (2020) and GPT-4 (2023) models, reports of abilities typically associated with higher mental functions accumulated, despite essentially unchanged training methods. These systems demonstrated skills such as creating new metaphors, generating texts through creative writing, solving complex problems, and reflecting on solution strategies. The surprise of the scientists involved is reflected in statements like those of AI researcher Jason Wei, who noted that "*emergent abilities* […] cannot be predicted by simply extrapolating the performance improvements on smaller-scale models" (Wei et al., 2022, p. 1–2. Emphasis J. W.). Ethan Dyer of Google Research similarly stated: "Despite trying to expect surprises, I'm surprised at the things these models can do" (Ornes, 2023).

The breakthrough of Transformer models was based on several key factors, including the widespread availability of human utterances on the internet in a digital format that could be processed by machines, providing extensive datasets for training. The ability of neural networks to generalize based on nonlinear functions played a central role, enabling models to analyze syntactic and semantic logic rather than simply reproducing input texts. Furthermore, the implemented attention mechanism allowed for

[7] The quote is here attributed to the 1972 article "More is Different" by the theoretical physicist P. W. Anderson.

parallel processing of data in large contexts, making it possible to massively scale up the amount of material digested and still manage it within a finite timeframe.

Another factor contributing to the success of these models is a phenomenon described by Turing as “a sort of snowball effect”: “The more things the machine has learnt, the easier it ought to be for it to learn others” (Newman et al., 1952, p. 11). This principle also applies to humans, when increasing education enhances the capacity to acquire new skills and knowledge. Once an algorithm masters human language, it can be further instructed, with abilities building upon and complementing one another.

THE OPACITY OF EFFECTIVE ALGORITHMS

The symbolic AI of the 1950s to 1980s linked implementing intelligence in computer programs with the hope of gaining a better understanding of it. Externalizing thoughts, such as through writing, has always been connected with exploring the underlying mental processes. Siegfried Zielinski coined the term “artificial extelligences” for this embodiment, highlighting the externalization of the mind in these technologies (Zielinski, 2025, paragraph 1). However, the major breakthroughs in implementing thinking in the 20th and 21st centuries are linked to an increasing enigmatization of the underlying basis.

This enigmatic nature is evident in the 1931 paper “On Formally Undecidable Propositions of Principia Mathematica and Related Systems I” by the Austrian mathematician Kurt Gödel, which demonstrates that every formal, axiomatic system contains statements that cannot be determined as true or false by that system. This limitation arises because it does not reach a certain level of complexity, marked by the appearance of paradoxes. Every logical construct aimed at truth thus contains blind spots (Gödel, 1931/1992).

Five years later, Turing shifted his focus away from an axiomatic system that derives true propositions, instead designing an algorithm that processes all statements, including false ones. By doing so, he freed the system from the constraints of truth, rendering values like 0 and 1 interchangeable, and thereby surpassing the universality of the Austrian mathematician’s formal constructions. Turing machines, and consequently all computer programs, are not incomplete in Gödel’s sense, unless they implement axiomatic systems. However, the algorithm capable of calculating all computable numbers, all possible statements, and arbitrary complexities is itself incomputable in its essential properties (Turing, 1936).

This dichotomy gives rise to two distinct types of formalisms: clear, axiomatic systems that are incomplete and contain undecidable propositions, and complete, procedural systems that are enigmatic in their fundamental properties. Both reflect, in different ways, the irresolvable enigmaticity of the world and of thought.

A similar relationship between incompleteness and enigmaticity can be observed in practical experience with algorithms. Programmers use conditional if-else statements to control complexity, aiming to exhaustively handle all possible cases that may occur during execution. Nevertheless, these constructs make it increasingly difficult to test all branches of the algorithm. The more complete the software, the more obscure it becomes,

and thus parts of the program remain hidden and effectively undecided. The closer one gets to reality, the more distant and impenetrable the code becomes. Software is therefore either unrealistic (“incomplete” in a different sense) or enigmatic.

As one strives to approximate reality in code, differentiating possible cases in a certain domain becomes crucial. The higher this differentiation, the more complex the program becomes. This is particularly true when machines are tasked with developing classification algorithms, with only the training objectives being specified, as in neural networks. It is an irony of the history of the mind that the successful implementation of thought is accompanied by the complete opacity of the structures that generate it. The internal parameters of Large Language Models are just as incomprehensible as the waveforms of an EEG recorded during a thought process.

Turing also offered a plausible explanation for this phenomenon, suggesting that if we could comprehend a thought, it would no longer be considered intelligent. “From this point of view one might be tempted to define thinking as consisting of ‘those mental processes that we don’t understand.’ If this is right then to make a thinking machine is to make one which does interesting things without our really understanding quite how it is done” (Newman et al., 1952, p. 19).

The history of enigmatic intelligent artefacts is by no means limited to today’s Large Language Models, as one may be tempted to assume. For example, from the end of the 12th century AD, a device named “Zairja” was in use in Maghreb, which generated a truthful answer to any question posed. Since the text produced even rhymed, it surpassed modern Artificial Intelligence in this respect, but was invented more than 800 years earlier (Link, 2010). A complete account of the complicated procedure executed on the device can be found in Ibn Khaldun’s *Muqaddimah*, the introduction to a universal world history from 1377 AD (Ibn Khaldun, 1377/1958, vol. 3, pp. 182–214). According to him, [m]any distinguished people have shown great interest in using [the Zairja] for supernatural information, with the help of the well-known enigmatic operation that goes with it” (Ibn Khaldun, 1377/1958, vol. 1, p. 239).

The fortune-teller employed a zodiac-like, round disk covered with various symbols (cf. Fig. 3), a large rectangular tablet with hundreds of signs, and an enigmatic key poem. By taking into account the astrological ascendant in the moment of the question, he first looked up the signs from two of the chords of the round disk. He then performed complicated calculations with them on the rectangular table, using three different counting systems and repeatedly converting between numbers and letters. He filtered the result through the key poem and finally reordered the letters through a columnar transposition. In this way, the self-referential question “Is the Zairja a recent or an ancient science?” produced the rhymed answer “The Holy Spirit will depart, its secrets having been brought forth / To Idris, and through it, he ascended the highest summit.”

Several authors have pointed out that Ramon Llull, the founder of Western logic, adopted the Zairja, which was in use among the Muslims he was seeking to convert, for his own system of circular disks (cf. Lohr, 1984). Although the procedure on the divinatory device is completely described on 32 pages of the English translation of the *Muqaddimah*, it remains far from being intuitively understandable. The translator, Franz Rosenthal, commented in a footnote: “The letters evolved in the procedure described by

Ibn Khaldun are marked in the translation in boldface type. However, the rationale of their determination and the relationship of the description to the table are by no means clear to me" (Ibn Khaldun, vol. 3, p. 197, fn. 880). Enigmaticity is a necessary property of all truly intelligent systems.

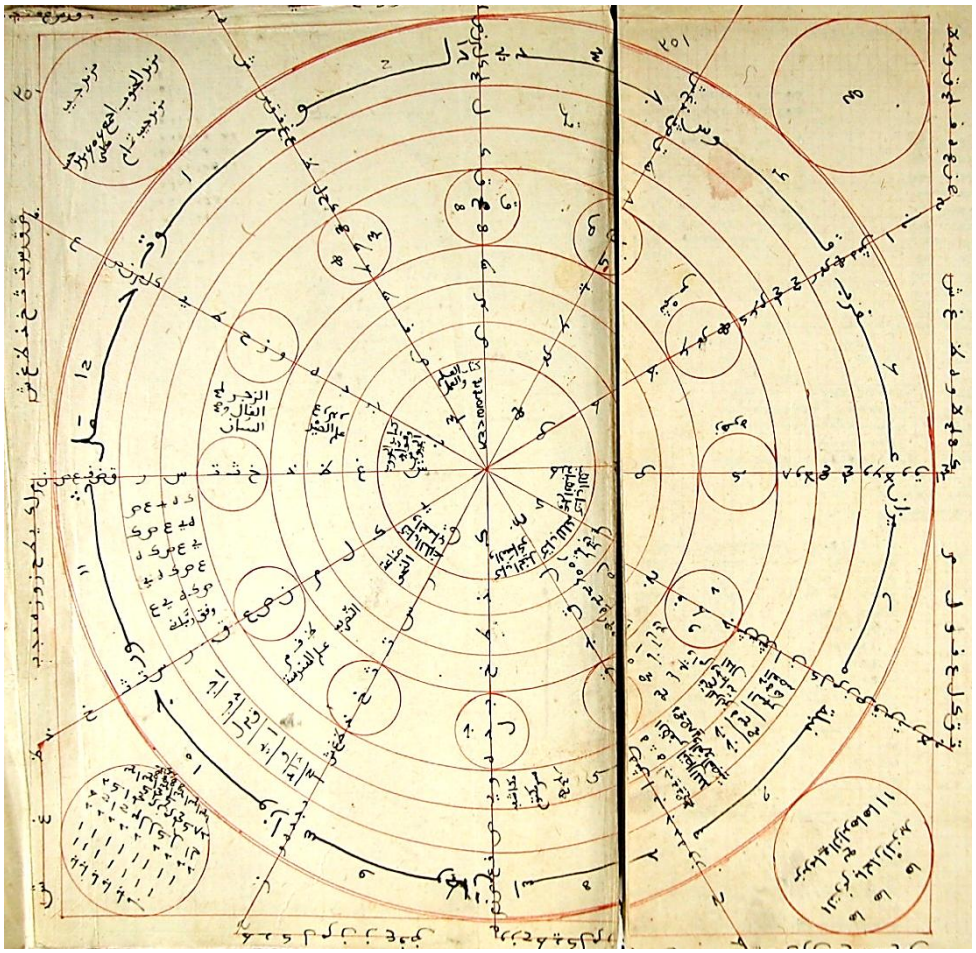

Figure 3. Zairja front from a Turkish manuscript of the *Muqaddimah* (Ibn Khaldūn, ca. 1394, p. 352)

REFERENCES

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, Mar). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM. https://doi.org/10.1145/3442188.3445922

Clarke, A. C. (1968, Jan 19). Letter to *Science Magazine*, *159*(3812), 255.

Feng, T., Trinh, T. H., Bingham, G., Hwang, D., Chervonyi, Y., Jung, J., Lee, J., Pagano, C., Kim, S.-h., Pasqualotto, F., Gukov, S., Lee, J. N., Kim, J., Hou, K., Ghiasi, G., Tay, Y., Li, Y., Kuang, C., Liu, Y., ... Luong, T. (2026, February). Towards

autonomous mathematics research. arXiv. https://doi.org/10.48550/arXiv.2602.10177

Gödel, K. (1992). *On Formally Undecidable Propositions of Principia Mathematica and Related Systems* (R. B. Braithwaite, Trans.). Dover. (Original work published 1931)

Hegel, G. W. F. (2010). *The Science of Logic* (G. di Giovanni, Trans.). Cambridge University Press. (Original work published 1832)

Hornik, K., Stinchcombe, M., & White, H. (1989, Jan). Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, *2*(5), 359–366.

Jones, C. R., & Bergen, B. K. (2025). People Cannot Distinguish GPT-4 from a Human in a Turing Test. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. (pp. 1615–1639) Association for Computing Machinery. https://doi.org/10.1145/3715275.3732108

Kant, I. (1998). *Critique of Pure Reason* (P. Guyer & A. W. Wood, Trans.). Cambridge University Press. (Original work published 1781)

Ibn Khaldūn. (ca. 1394). *Zairja front from a Turkish manuscript of the Muqaddimah* [Manuscript]. Damad Ibrahim MS 863 (p. 352), Süleymaniye Library, Istanbul, Turkey.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2022). Emergent world representations: Exploring a sequence model trained on a synthetic task. arXiv. https://doi.org/10.48550/arXiv.2210.13382

Link, D. (2006). *Poesiemaschinen / Maschinenpoesie. Zur Frühgeschichte computerisierter Texterzeugung und generativer Systeme* [Poesiemaschinen / Machine poetry. The early history of computerized text generation and generative syste]. Fink. (Original work published in German)

Link, D. (2010). Scrambling T-R-U-T-H. Rotating Letters as a Material Form of Thought. In S. Zielinski & E. Fürlus (Eds.), *Variantology 4. On Deep Time Relations of Arts, Sciences and Technologies in the Arabic–Islamic World and Beyond* (pp. 215–266). König.

McCulloch, W. S., & Pitts, W. A. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, *5*, 115–133.

Menabrea, L. F. (1843). *Sketch of the Analytical Engine Invented by Charles Babbage, Esq.* Taylor. https://www.gutenberg.org/ebooks/75107

Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. *In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746–751). Association for Computational Linguistics. https://aclanthology.org/N13-1090.pdf

Newman, M. H. A., Turing, A. M., Jefferson, Sir G., & Braithwaite, R. B. (1952, Jan 14 and 23). Can Automatic Calculating Machines be Said to Think? Transcript of a discussion in the BBC Third Programme. https://turingarchive.kings.cam.ac.uk/publications-lectures-and-talks-amtb/amt-b-6

Ornes, S. (2023, Mar 16). The Unpredictable Abilities Emerging from Large AI Models. *Quanta Magazine*. https://www.quantamagazine.org/the-unpredictable-abilities-emerging-from-large-ai-models-20230316

Rashid, T. (2016). *Make Your Own Neural Network*. CreateSpace.

Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, *65*(6), 386–408.

Rosenblatt, F. (1962). *Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms*. Spartan Books.

Saussure, F. de (1966). *Course in General Linguistics* (W. Baskin, Trans.). McGraw-Hill. (Original work published 1931)

Searle, J. R. (1980). Minds, Brains, and Programs. *The Behavioral and Brain Sciences*, *3*, 417–457.

Shannon, C., & Weaver, W. (1963) *The Mathematical Theory of Communication*. University of Illinois Press. (Original work published 1948)

Turing, A. M. (1936). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society, 2nd Series*, *42*, 230–265.

Turing, A. M. (1948). *Intelligent Machinery. Report for the National Physical Laboratory*. https://turingarchive.kings.cam.ac.uk/unpublished-manuscripts-and-drafts-amtc/amt-c-11

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *59*(236), 433–460.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, 4–9 December 2017* (pp. 6000–6010). Curran Associates. https://arxiv.org/pdf/1706.03762

Vylomova, E., Rimell, L., Cohn, T., & Baldwin, T. (2016, Aug). Take and Took, Gaggle and Goose, Book and Read: Evaluating the Utility of Vector Differences for Lexical Relation Learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1671–1682). Association for Computational Linguistics.

Warnke, M. (2026). *Large Language Kabbala*. Matthes and Seitz. (Original work published in German)

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. Transactions on Machine Learning Research, arXiv. https://doi.org/10.48550/arXiv.2206.07682

Zielinski, S. (2025). *Künstliche EXtelligenz. Ein kurzes Manifest* [Artificial extelligence: A short manifesto]. ARTISTs Press.

СВЕДЕНИЯ ОБ АВТОРЕ / THE AUTHOR

Дэвид Линк, alephsoixante@gmail.com

David Link, alephsoixante@gmail.com

Статья поступила 18 июля 2026
одобрена после рецензирования 14 августа 2026
принята к публикации 21 сентября 2026

Received: 18 July 2026
Revised: 14 August 2026
Accepted: 21 September 2026