Studium. Filosofía y Teología. Vol. XXIX 57 (2026) 73-87
ISSN 0329-8930 - ISSNL 2591-426X

General Artificial Intelligence needs Consciousness

La Inteligencia Artificial General necesita conciencia

George Zarkadakis
Independent researcher
zarkadakis@gmail.com
ORCID: 0000-0001-5191-4337

DOI: https://doi.org/10.53439/stdfyt57.29.2026.73-87

Abstract: Human-like, Artificial General Intelligence (AGI) must autonomously explore reality, set its own goals, and demonstrate genuine creativity, such as making original scientific discoveries. Creativity in humans is rooted in high-level consciousness and the subjective feeling of personhood, making it unattainable for machines that are merely intelligent but not conscious, because creativity arises from stochastic processes machines can only emulate, not reproduce it. The frame and halting problems further limit conventional computational systems, preventing them from transcending their operational boundaries. Consciousness, on the other hand, can overcome operational boundaries set by evolution and access deeper layers of reality. Consequently, sentient AGI is not possible using conventional computer science approaches. Alternative approaches—such as biocomputing or quantum-based models of consciousness—might yield sentient machines in the future, though such hypothetical advancements raise profound ethical questions about creating new forms of awareness.

Resumen: Una Inteligencia Artificial General (AGI) de tipo humano debe explorar la realidad de manera autónoma, establecer sus propios objetivos y demostrar una creatividad genuina, como la realización de descubrimientos científicos originales. La creatividad en los seres humanos está arraigada en la conciencia de alto nivel y en la experiencia subjetiva de la propia identidad personal, lo que la hace inalcanzable para máquinas que son meramente inteligentes pero no conscientes, dado que la creatividad surge de procesos estocásticos que las máquinas solo pueden emular, pero no reproducir. Los problemas del marco y de la detención limitan aún más a los sistemas computacionales convencionales, impidiéndoles trascender sus propios límites operativos. La conciencia, en cambio, puede superar los límites operativos establecidos por la evolución y acceder a niveles más profundos de la realidad. En consecuencia, una AGI sintiente no es posible mediante los enfoques convencionales de la ciencia de la computación. Enfoques alternativos –como la biocomputación o los modelos de la conciencia basados en la mecánica cuántica– podrían dar lugar a máquinas sintientes en el futuro; sin embargo, tales avances hipotéticos plantean profundas

cuestiones éticas acerca de la creación de nuevas formas de conciencia.

Keywords: Artificial General Intelligence, consciousness, creativity, biocomputing, quantum-based models

Palabras clave: Inteligencia Artificial General, conciencia, creatividad, biocomputación, modelos basados en la mecánica cuántica

Recibido: 27/10/2025
Aceptado: 10/03/2026

Introduction: Why We Need a Much Tougher Test for AGI

Before Aristotle, the implicit assumption was that intelligence and consciousness were interchangeable concepts. In the *Organon* Aristotle demonstrated how at least one category of intelligence –logical thinking– could exist without the need for sentience. His proposal ushered a long process of logical reasoning about phenomena to increase human knowledge that culminated in the scientific revolution and the advent of "computer" machines that could perform step-by-step logical inferences with great success. It was an astonishing feat and for a while electronic computers were referred to as "electronic brains". Nowadays, it is not uncommon to refer to brains as "biological computers", a metaphor that partly explains why the word "Intelligence" in "Artificial Intelligence" has been the source of much misunderstanding (Zarkadakis, 2015).

The founding fathers of AI used the word "intelligence" to suggest an electronic computer capable of logical thinking in the Aristotelian sense. Phenomenal consciousness and its elusive side-effects such as personhood, emotions, and feelings were not implicated in the original conceptualisation of AI. Indeed, given the context of Cold War in the late 1950s, the undeclared intention of the AI project was to develop decision-support systems that would give the USA a decisive advantage in a possible thermonuclear war. An "AI general" would be a supreme commander, one that would take life and death decisions in split seconds free of feelings. The same AI advantage could be exploited in business too. The "AI CEO" would be a superior captain of industry exactly because the system will be one of pure intelligence and no consciousness –a supersmart psychopath.

Nearly seven decades after the seminal Dartmouth Summer Research Project on AI we now have intelligent systems that easily pass the Turing Test. Advanced Large Language Models (LLMs) can fool us into thinking that their intelligence is on a par with our own, especially when they behave in quin-

tessential conscious ways and try to cheat us (Park et al., 2024), or solve gold medal level questions in the International Mathematical Olympiad (Trinh et al., 2024), or succeed in passing very hard tests across science and engineering disciplines (Kung et al., 2023). Theory of mind, our instinctive tendency to assume minds in other persons, animals, as well as inanimate things like computers, is being exploited as Big Tech companies tout the imminent arrival of "Artificial General Intelligence" (AGI), or "human-level" AI.

Although there is no consensus on what "Artificial General Intelligence" means, we will assume for the purpose of this paper a computational system capable of matching and surpassing humans in every cognitive task. The most advanced AI systems presently, such as LLMs, surely do not fall under the category of "narrow" or "weak" AI, which is when systems lack general cognitive abilities. LLMs are endowed with high degree of generality, as they can be pre-trained to apply across several domains of knowledge. Impressive as they are, however, they are also riddled with serious limitations. Their understanding of the real world is lacking –a weakness that demands considerable human curation to limit their tendency for confabulation (also called "hallucination"). Arguably, their limited understanding of the world is because they have been trained on data produced by humans. They have not learned things about the world by themselves, through self-exploration and trial-and-error, as humans and other animals do. They were never *babies*. One might therefore suggest that to get to *human-level* AGI we ought to embody those systems and release them into the real world so that they may discover it for themselves. We already have a proven method for furnishing machines with the ability to teach themselves called "Reinforcement Learning" (RL), whereby a sensory input leads to action and a reward (or punishment). RL is at the heart of advanced AI systems capable of self-learning. Driverless cars are an example of embodied autonomous AI that uses RL and which, after many years of self-exploration and learning into the real world, is finally achieving productivity and reliability. But driverless cars are "narrow AI", not "General AI". In another example, an LLM called DeepSeek-R1 was recently trained using RL and was rewarded with a high score when it solved mathematical problems correctly and penalized when it got the answer wrong. As a result, it learned that stepwise reasoning (also called "chain-of-thought") was more likely to lead to a correct answer. It changed its behaviour and started to self-verify and self-reflect, checking its performance before giving answers to new questions and thereby improving its performance in coding and graduate-level science problems (Guo et al., 2025).

Although the system *invented* a successful strategy thanks to RL this is not enough evidence for AGI, because for RL to produce such impressive results needs a value-based goal –and in this case the goal of solving of mathematical and coding problems that was set by its human creators.

Animals also use a reward system to explore and learn about the world, and develop successful strategies for survival and reproduction, whereby pain is a cost and pleasure a reward. At an abstract level their goal-seeking behaviour, as shaped by evolution, is governed by maximising pleasure and minimising pain. One could therefore suggest that AI systems could also be equipped with a self-learning computational module that takes sensory inputs from its environment and sets itself goals for survival while aiming to maximise/minimise pleasure/pain similarly to animals –if only the objective was to produce "Artificial Animal Intelligence"! But could that type of AI naturally *evolve* to human-level?

What distinguishes humans from the rest of the animals is that we also have *higher goals* beyond survival and reproduction. These higher goals drive our unique ability for creativity that manifests in civilization-building. We are not only curious about exploring the world that surrounds us, but we invent ways to transcend the limits of our senses. Because of this ability we have discovered science and maths and created artistic masterpieces. AGI must be able to have similar instincts and abilities to qualify as *human-level*. It must be able for genuine creativity instead of merely imitating pre-existing human ideas picked up from its training data. AGI must intuit like a human. Most proposed tests of AGI are therefore inadequate: in the Turing Test, Robot College Student Test, Employment test, Ikea test, Coffee test, Modern Turing test, high-level goals are set by humans.

Thus, the ultimate test for AGI should be one where the system discovers something significant –a new scientific theory, let us say– out of curiosity and a desire to do something worthy in pushing the boundaries of science. Until a system does so for these reasons, all iterations in AI will be producing better and more powerful AI systems but not human-level. In 2016 the chief executive of Sony, Hiroaki Kitano, suggested exactly such a test; and called it the *Nobel Turing Challenge*. He challenged researchers to develop an AI system that would make a discovery worthy of a Nobel Prize (Lu et al., 2024). But, could such a system ever be possible? Dough Downey, a researcher at The Allen Institute for Artificial Intelligence (2025), and his colleagues recently studied 57 AI agents and found that although the agents could fully complete specific science-related tasks about 70% of the time, that figure dropped to merely 1%

when they attempted to generate an idea, plan and execute an experiment and analyse data for a full report. End-to-end scientific discovery is a formidable challenge because it needs something very important that conventional computation systems seem to lack: creativity. But could AI systems achieve that capability soon, as the current consensus among AI researchers suggests?

Creativity Needs Sentience

The neurobiological basis of creativity in humans involves dynamic interplay between distinct brain networks, rather than a single *creativity centre*. Key regions include the prefrontal cortex (especially the frontopolar cortex and medial prefrontal cortex), which is crucial for cognitive control and evaluation, and the medial temporal lobe (including the hippocampus and amygdala), involved in idea generation and memory integration. The Default Mode Network (DMN), active during rest and spontaneous thought, is also vital for divergent thinking, working in coordination with the prefrontal cortex's executive functions. Finally, spontaneous fluctuations in neuronal activity and the integration of random neuronal noise with learned information contribute to the novelty and meaningfulness of creative output (Park et al., 2016). Summarizing, creativity in humans involves a unique blend of random neuronal processes (stochasticity) and organized, knowledge-based information (order), leading to both original and meaningful outcomes (Liu et al., 2024). Stochasticity is a key characteristic of creativity and designers of LLMs have taken this into account by introducing *random noise* into the attention mechanism of LLMs. One can calibrate an LLM to be more *creative* by increasing the stochasticity of the system. And yet what one gets by doing so is a *creative output* filled with evermore confabulation. The more stochastic the LLM, the more divergent from reality its outputs are. It is also important to note that stochasticity is not a physical thing or a process. It is a mathematical way to quantify our ignorance about something. As such it can be emulated in a computer but not reproduced. This important difference makes biological brains ontologically different from computers. To discover why the biological brains of humans are stochastic and yet their outputs are grounded to the experience of the real world –while that of AIs do not– one must look at the evolution of our species.

It has been an amazing coincidence that intelligence co-evolved with consciousness. We do not really know why or how that happened yet. To be clear, there are many levels of consciousness, but for the purpose of this exa-

mination let us consider the *highest* level only, often referred to as *phenomenal consciousness* or *sentience* that includes the strange feeling of personhood. Intriguingly, we only become conscious in this way when we must pay attention to something. It is only at such instances that sentience kicks in and we get that unitary feeling of selfhood. Apart from those instances, our bodies are guided by unconscious processes, and rather effectively too. We are mostly mindless biological automata, our brains executing complex computations –such as predicting whether the car in front of us will stop or accelerate– without us needing to be fully conscious. And yet, occasionally, evolution has decided to present us with a strong feeling of being inside reality as conscious agents and purposeful actors. This feeling is particularly vivid when an unexpected event occurs, and deeper thinking is required and not a merely a learned reaction. So maybe consciousness evolved as a mechanism for animals to deal with uncertainty. To speculate further, the more complex the world becomes the more uncertainty there is, and therefore the reward function of evolution drives even more complex adaptive behaviours that require high level consciousness. In the definition of an "evolving world" we should include social structures –which could be an explanation for us humans having evolved our consciousness further than other species on this planet. But consciousness does more than help us deal with uncertainty in a complex and dynamic environment. It also makes us *doubt* the phenomenal reality that surrounds us. Because of high-level consciousness we suspect that the world may be very different from what our senses tell us. That there is a *hidden* reality. Mental phenomena such as dreams, religious experiences, drug-induced hallucinations, and out-of-body experiences are usually dismissed as irrelevant, marginally important, delusions caused by chemical imbalances in the brain. Nevertheless, such experiences are at the root of human creativity and exploration. Often, they set off a quest for *meaning* as we start to seek the reason for our existence, an ultimate purpose perhaps.

Because of such *altered* mental experiences we have obtained the curious intellectual capacity to doubt our senses, and in the process managed to transcend the phenomenal reality that surrounds us and discover philosophy, science and mathematics. We invented technologies and created art. We set lofty goals such as to build pyramids and cathedrals, fly to the stars, and create sentient AI systems. These self-setting goals are well outside what was required for the mere survival of our species. We are explorers and inventors only because we can transcend biological automation thanks to this strange ability of our species for high-level consciousness.

The Mysterious I

The reason phenomenal consciousness is *strange* is because we have no scientific evidence that it exists. And yet, we all *feel* that we have it. The current scientific consensus is that self-awareness emerges from somewhere inside the brain, perhaps as the result of complex interactions between the various parts of the brain –but no one is certain. That's not because of lack of trying. Plenty of probing instruments have been invented that can detect electrical activity and blood flow in the brain and associate it with various stages of consciousness. For instance, fMRI can tell whether a person is conscious or not –even if that person happens to be in a vegetative state (Owen, 2013); or whether they are dreaming (Raduga & Shashkov, 2023). Neural links have been developed that can read internal dialogue of paralysed patients and translate it into language that can be generated and voiced through a computer. (Kunz et al., 2025). Nevertheless, no scientific instrument currently exists that can lift the veil of self-awareness and reveal the so-called "neural correlates" of high-level consciousness. We do not know how electrochemical excitations of neurons produce the unitary feeling of personhood. We have no idea why some things smell sweet and others acrid, or why we see colours. These so-called *qualia* remain irritatingly inexplicable despite the considerable technological advancement of the past few years. The "hard problem" of consciousness remains unassailable by the scientific method (Chalmers, 2007). And yet none of us needs an instrument to tell them who they are. Having an individual identity is as real as anything in the world around us. That feeling of self is tethered to our memories –however fluid and mutable– and our senses, feelings, emotions, and moral values, convincing us that we exist in the world as unique and separate entities. Without consciousness we are helpless. The world vanishes. We are as good as dead. No wonder the ancient Greeks considered Thanatos (death) and Hypnos (sleep) to be twin brothers.

Stochastic Creativity Meets Probabilistic Reality

There is an inextricable link between consciousness and reality that has been explored by many philosophers and scientists, from the ancients to the moderns. From Plato to Kant and Berkeley there have been strong arguments that a material universe independent of an observant mind is metaphysically impossible. Qualities, such as shape, colour, smell, texture, depend on minds. Surely, the idealists have claimed, there must exist some

objective reality of *things-in-themselves*, however that noumenal dimension is unknowable. Scientific positivists have pushed back against this notion claiming that objective reality is ultimately knowable, and that the scientific method was the surest way to do so. The scientific revolution that was ignited in the 16th century and the marvels of engineering that ushered the industrial revolution two centuries later seemed to vindicate them. Aristotelian empiricism was marching triumphant until the early 20th century when he hit upon the limits of knowing and understanding.

The discovery of quantum physics in early 20th century provided hard evidence for what used to be purely philosophical debate. We now *know* that reality as perceived by our minds does not exist. There is no *real* space or time. Quantum reality –what we think is the deeper and *truer* level of reality– is a weird, mathematical construct in Hilbert space suspended in probabilistic superposition that occasionally interacts with the four-dimensional world of our perception. That four-dimensional world that we experience is in effect an illusion propagated by evolution. Our minds, like the minds of other animals, evolved so that they can successfully manipulate objects at this "fake reality interface", similarly to how we manipulate digital objects on a computer screen (Hoffman, 2019). As in the computer metaphor the *real* system is not the virtual things floating on the screen but the hard-wired electronics inside the box.

So, what is *real*? At the quantum level everything exists as entangled probabilities, which is a meaningless statement albeit mathematically precise. Saying that everything can be everywhere and nowhere always is not different that saying nothing ever exists. Then, suddenly, when intelligent observers try to measure something that something becomes. Quantum strangeness commands that spacetime emerges out of nothingness only when observed. The similarities between attention begetting the unitary experience of consciousness and measurement begetting macroscopic reality are too intriguing to ignore. There are of course various interpretations trying to resolve this *measurement problem*, all of them ending up in absurdities. Either there are infinite universes, or *hidden variables* that we will never discover, or reality is truly non-existent unless someone observes it. Our minds tell us that this vast universe made of billions of stars and trillions of planets must exist without the need of us looking at it through our telescopes. We fall asleep at night certain that the Moon outside our window is shedding its silver rays on our face while we dream. And yet, quantum physics says otherwise. Phenomena are all we can get. And to get them we need to be awake and pay attention. The

universe is the creation of attentive minds. It is atop that perceived reality that we have built our technologies, including digital computers. These computers manipulate states on the same perceived *sensory interface* to perform their calculations. They are *virtual* machines operating in our *virtual* world. And yet our creative and curious brains have managed to penetrate the *interface* and discover a deeper layer of reality –the quantum reality. We could not have possibly done so without the freedom of thought and imagination enabled by high-level consciousness.

What about Conscious Machines?

Let's recap what has been discussed so far. We must apply a much more stringent and rigorous test when deciding whether an AI system possesses human-level intelligence: the system must be genuinely creative while setting up its own goals. It must be a self-motivated explorer, an inventor, a scientist and an artist of excellence and originality, at least on a par with the best of us. *Creativity* was selected as a criterion because it has permitted humans to transcend the boundaries of their experience and discover deeper layers of reality. Creativity requires sentience and is a stochastic process, which means that we will always have limited knowledge of its cause. Because of stochasticity, creativity can only be *emulated* in a machine but not *reproduced*.

This argument poses a serious problem for the dominant paradigm in consciousness (as well as *machine consciousness*) studies, which is the idea that consciousness is the result of reproducible neural processes in the brain. This "functionalist" view purports that there is functional equivalence between computers and brains. Thus, a functionalist would hasten to propose that there is nothing special about our biological brains, being machines of a different, wetter kind. That a reductive approach based on neuroscience will sooner or later decode the algorithm of biological mental processes –the *neural correlates*– that are the cause of consciousness so that they can be replicated computationally to any another physical medium, such as one made of silicon and thus furnish machines with personhood, self-awareness, and creativity too. For a computational functionalist, consciousness *is* an algorithm that can run on any physical system capable of computations. Thus, consciousness in AI could be detected and verified through brain scans and other techniques developed by neuroscientists (Butlin et al., 2023).

Functionalists may refute the argument that creativity is non-reproducible, and therefore non-computable, on the grounds that nothing exists be-

yond the *physical world* where computations take place. All physical processes can be ultimately simulated using a Universal Turing Machine, and therefore consciousness should be *computable* too. However, as demonstrated by quantum physics, the physical world is an illusion created by our brains –an *interface* where animals like us can interact with objects and processes that exist in four-dimensional spacetime. There is a deeper layer of reality that is completely different from the perceived reality, where spacetime does not exist. Our computations can only take place at the illusionary, macroscopic interface of chemistry, Newtonian physics, and transistors. But even if one insists that consciousness is a process restricted to the interface with no relation to the deeper –and rather mysterious– reality of mathematical entities suspended in Hilbert space, the problem with computational functionalism is that computability in this physical reality interface has considerable limits too.

In 1969 John McCarthy and Patrick J. Hayes laid out the "frame problem" in AI, which is the difficulty to teach a machine to make smart decision based on relevant information without having it consider every irrelevant detail. Such details are endless. Knowledge is forever incomplete, which is one of the main reasons why training AI with data will never be enough for those systems to get a satisfactory understanding of the real world. Developing a robust, internal *world model* grounded on truth is not a trivial problem. Indeed, the only way to develop such an internal representation is through self-directed exploration –as animals do.

Moreover, as Alan Turing, proved in 1936, there is no algorithm that can always determine whether another algorithm, given some input, will stop or run forever –the so-called "halting problem" (Turing, 1937). Computers running on algorithms are prone to getting stuck in loops, while biological brains do not. For instance, LLMs will soon finish consuming all available human generated data. At the same time, they are already polluting the web with their own generated output data that are being recycled into their training. This will lead to a gradual model degradation process that risks future AI systems having even less understanding about the real world (Shumailov et al., 2024). The reason why those AI systems are stuck in this downward spiral is because they are unable to comprehend the problem of using their own outputs for training and break out of the loop. Computers are not aware of the halting problem because they lack consciousness, and they will forever lack consciousness, regardless of how many hacks software engineers come up with, because consciousness in non-computable.

Biological supremacy beats functionalism

We must accept our ignorance about the primal cause of consciousness, about how and why it co-evolved with intelligence, and about why the stochastic nature of human creativity makes it capable of transcending conventional experience and discover deeper levels of reality. All these questions remain deeply mysterious. Nevertheless, denying their validity and importance while insisting that computational functionalism will inexorably lead to machine consciousness is short-sighted. We must doubt the functional equivalence between brains and computers. We have plenty of empirical evidence to do so. Human brains consume less than 20 Watts of energy to perform cognitive processes that take computers a million times more energy to do so (Balasubramanian, 2021). Humans can learn new tasks after being shown a few examples while a computer needs significantly more data, sometimes millions, to do the same (Flesch et al., 2018). Biocomputing is exploring this unique ability of brain cells for cognition at low energies. Brain organoids are tiny, lab-grown mini-brains, made from living stem cells which have been cultured to become clusters of neurons and supporting cells. In a much-discussed experiment a cluster of organoids was connected to a computer and learned how to play ping pong in less than 10 minutes, all by itself and without code. (Kagan et al., 2022). This is a very exciting area of research with many open ethical questions. Would a more sophisticated assembly of organoids exhibit consciousness? How must we treat the new sentient entities that we will have created in the lab? What laws should apply to protect them from suffering and harm? How far should we develop them? How intelligent we would like them to be?

Alternative computing approaches may also help explore non-computable stochastic cognitive processes where, evidently, time and entropy are essential conditions. Thermodynamics is already being explored for AI applications (Melanson et al., 2025). The free energy principle proposed by Karl Friston (2010) aims to explain how organoids (and brains) function by combining ideas from thermodynamics, control theory and Bayes Theorem. This mathematical theory describes how biological systems maintain their existence by minimizing *surprise* (or prediction error) over time through a process of continuous perception and action. It posits that to survive any living system must resist a natural tendency toward disorder by always acting to minimize the difference between its predictions about the world and the sensory input it receives. The element of surprise is another way of articulating what has already been discussed in this paper as the incomplete-

ness of knowledge about the real world –and, speculatively, the reason why evolution furnished us with consciousness.

Other researchers speculate that consciousness may be more fundamental than biology. Most notably among them, Stuart Hameroff in collaboration with Sir Roger Penrose have proposed a quantum theory of consciousness that suggests that consciousness is not an emergent, but a fundamental property of reality. Mental properties including qualia accompany self-collapse of a quantum wave function by a mechanism called "objective reduction (OR)", a threshold in the structure of spacetime geometry (Hameroff, 2023). Penrose's theory of objective reduction (OR) is considered *non-computable* because it posits that the collapse of a quantum superposition into a definite classical state is a spontaneous, self-collapse event influenced by fundamental spacetime geometry and a non-algorithmic process, making its outcome inherently unpredictable by any finite set of algorithmic steps, unlike standard quantum theory which relies on observer-induced collapses. In other words, no *observer* is necessary for the quantum system to collapse. There has been much criticism of the quantum theory of consciousness, including the fact that the theory is grounded on ideas about quantum gravity that are very speculative. Nevertheless, it would be too early to dismiss the idea, as it opens an enormous range of exciting possibilities and has been shown to make experimentally falsifiable predictions (Wiest, 2025).

If consciousness is a fundamental property of quantum reality, then it should be incorporated in our understanding of physics and biology. Such a unification of mental, biological and physical processes goes beyond the current paradigm of functionalism that reduces consciousness to an algorithm (Zarkadakis, 2001). We are in the very early stages of quantum engineering with a very limited scope of applications, such as quantum computing and teleportation. As our physics and our quantum technology advance further, we might be able to experimentally test a quantum theory of consciousness in the future. If such theory were to be verified it would be nothing less than a new Copernican revolution, as it would introduce a general theory of intelligence in a physical theory of everything.

Conclusions

In this paper the following arguments have been put forward:

Argument 1: We need a more stringent definition of Artificial General Intelligence. Human-level AGI systems must be able to self-explore reality, set their own goals, and be genuinely creative.

Argument 2: Creativity requires high-level consciousness that includes the unitary feeling of personhood. Therefore, the only way to satisfy Argument 1 is for AIs to become conscious and not simply *superintelligent*. Creativity in the human brain is based on stochasticity. As such, creativity can only be emulated in a computer but not reproduced.

Argument 3: The frame and halting problems impose insurmountable problems on computers being able to transcend their functional limitations, unlike brains that do. Therefore, there is no functionalist equivalence between brains and computers.

Argument 4: Phenomenal consciousness can transcend perceived reality and has allowed humans to discover in quantum physics a deeper layer of reality; a discovery that is beyond our biologically evolved senses and understanding. Intriguingly, both consciousness and that deeper layer of reality require attention to manifest.

Based on the above arguments, the following conclusions can be drawn: Human-like AGI is impossible with current computer technology. Our AI technology based on conventional computer architectures will improve further and develop *non-human-like* superintelligence but will never reach sentient level. Given the constraints for human-like AGI it is very likely that AI companies will rebrand *AGI* as *superintelligence* or some other similar term.

Alternative paths to sentient machines, most notably thermodynamic computing and biocomputing, as showing promise. As we further advance quantum engineering, we may also be able to experimentally test theories of consciousness as a fundamental property of reality. All these paths raise serious ethical questions about creating new forms of sentience that must be considered.

References

Balasubramanian, V. (2021). Brain Power. *Proc. Natl. Acad. Sci. U. S. A.*, *118*(32), e2107022118. https://doi.org/10.1073/pnas.2107022118

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A. et al. (2023). Consciousness in Artificial Intelligence: insights from the Science of Consciousness. *arXiv preprint arXiv:2308.08708*. https://doi.org/10.48550/arXiv.2308.08708

Chalmers, D. (2007). The Hard Problem of Consciousness. In M. Velmans & S. Schneider (Eds.), *The Blackwell Companion to Consciousness* (pp. 32-42). Wiley-Blackwell.

Flesch, T., Balaguer, J., Dekker, R., Nili, H. & Summerfield, C. (2018). Comparing Continual Task Learning in Minds and Machines, *Proc. Natl. Acad. Sci. U.S.A, 115*(44) E10313-E10322. https://doi.org/10.1073/pnas.1800755115

Friston, K. (2010). The Free-energy Principle: A Unified Brain Theory? *Nat Rev Neurosci*, *11*, 127-138. https://doi.org/10.1038/nrn2787

Guo, D., Yang, D., Zhang, H. et al. (2025). DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning. *Nature, 645*, 633-638. https://doi.org/10.1038/s41586-025-09422-z

Hameroff, S. (2023). Consciousness Is Quantum State Reduction Which Creates the Flow of Time. *Timing & Time Perception*, *12*(2), 158-167. https://doi.org/10.1163/22134468-bja10098

Hoffman, D. (2019). *The Case Against Reality: How Evolution Hid the Truth from Our Eyes*. Penguin.

Kagan, B. J., Kitchen, A. C., Tran, N. T., Habibollahi, F., Khajehnejad, M., Parker, B. J., Bhat, A., Rollo, B., Razi, A. & Friston, K. J. (2022). In Vitro Neurons Learn and Exhibit Sentience when embodied in a Simulated Game-world. *Neuron*, *110*(23), 3952-3969.e8. http://doi.org/10.1016/j.neuron.2022.09.001

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J. & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted Medical Education using Large Language Models. *PLOS. Digital Health*, *2*(2), e0000198. http://doi.org/10.1371/journal.pdig.0000198

Kunz, E. M., Krasa B. A., Kamdrar F., et al. (2025). Inner Speech in Motor Cortex and Implications for Speech Neuroprostheses. *Cell*, *188*(17), 4658-4673.e17. https://doi.org/10.1016/j.cell.2025.06.015

Liu, C., Zhuang, K., Zeitlen, D.C. et al. (2024). Neural, Genetic, and Cognitive Signatures of Creativity. *Commun Biol*, *7*, 1324. https://doi.org/10.1038/s42003-024-07007-6

Lu, Ch., Lu, C., Lange, R. T., Foerster, J., Clune, J. & Ha, D. (2024). The AI Scientists: Towards Fully Automated Open-Ended Scientific Discovery. *arXiv preprint arXiv:2408.06292*. https://doi.org/10.48550/arXiv.2408.06292

Melanson, D., Abu Khater, M., Aifer, M., Donatella, K., Gordon, M. H., Ahle, T., Crooks, G., Martinez, A. J., Sbahi, F. & Coles, P. J. (2025). Thermodynamic Computing System for AI Applications. *Nat Commun, 16*, 3757. https://doi.org/10.1038/s41467-025-59011-x

McCarthy, J. & Hayes, P. J. (1969). Some Philosophical Problems from the Standpoint of Artificial Intelligence. *Machine Intelligence*, *4*, 463-502.

Owen A. M. (2013). Detecting Consciousness: a Unique Role for Neuroimaging. *Annual Review Psychology*, *64*, 109-133. http://doi.org/10.1146/annurev-psych-113011-143729

Park, P. S., Goldstein, S., O'Gara, A., Chen, M. & Hendrycks, D. (2023). AI Deception: A Survey of Examples, Risks, and Potential Solutions. *Patterns*, *5*(5). https://doi.org/10.1016/j.patter.2024.100988

Park S. H., Kim, K. K. & Hahm, J. (2016). Neuro-Scientific Studies of Creativity. *Dement Neurocogn Disord*, *15*(4), 110-114. https://doi.org/10.12779/dnd.2016.15.4.110

Raduga, M. & Shashkov, A. (2023). Detecting Lucid Dreams by Electroencephalography and Eyebrow Movements. *Sleep Science, 16*(4), e408-e416. http://doi.org/10.1055/s-0043-1776749

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R. & Gal, Y. (2024). AI Models Collapse when Trained on Recursively Generated Data. *Nature, 631*, 755-759. https://doi.org/10.1038/s41586-024-07566-y

The Allen Institute for Artificial Intelligence. (2025, August 26). AstaBench: Rigorous benchmarking of AI agents with a holistic scientific research suite. *Ai2*. https://allenai.org/blog/astabench

Trinh, T.H., Wu, Y., Le, Q.V., He, H. & Loung, T. (2024). Solving Olympiad Geometry without Human Demonstrations. *Nature, 625*, 476-482. https://doi.org/10.1038/s41586-023-06747-5

Turing, A. M. (1937). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society. s2-42*(1), 230-265. https://doi.org/10.1112/plms/s2-42.1.230

Wiest, M. C. (2025). A Quantum Microtubule Substrate of Consciousness is Experimentally Supported and Solves the Binding and Epiphenomenalism Problems. *Neuroscience of Consciousness*, *2025*(1), niaf011. https://doi.org/10.1093/nc/niaf011

Zarkadakis, G. (2001). Noetics: A Proposal for a Theoretical approach to Consciousness. *Proceedings of International Conference "Toward a Science of Consciousness: Consciousness and its place in Nature*". University of Skövde, Sweden, 7-11 August.

--. (2015). *In Our Own Image: Will Artificial Intelligence Save Us or Destroy Us?* Rider Books.