Should an AI be conscious? Approaches to human and artificial consciousness

Óscar Trabazos Clavería · Essay, not peer reviewed · Version 2.0, 2 November 2026 (first version, 2023)

Abstract

Could a machine programmed by people have subjective experiences, such as seeing a colour, feeling pain or enjoying a melody? This essay addresses that question in four parts. The first reviews what we mean by phenomenal consciousness and the classic answers to the mind-body problem: dualism, materialism in its various forms, and idealism. The second studies the brain as an information processor, from the beginnings of cognitive science to models of memory and the challenge of metaphor. The third applies all of this to artificial intelligence: strong AI, the Turing test, the Chinese room, current scientific theories of consciousness and language models, with their hallucinations and their resemblance to those of the human brain. The fourth takes on the ethical and philosophical questions raised by an artificial mind, the simulated universe hypothesis and the relationship between entropy, longevity and digital consciousness.

The author's conclusion is that current systems offer no evidence of consciousness: they are systems of enormous capacity that respond with the most probable output. We cannot create in non-biological form what we do not yet understand, and if an artificial general intelligence ever reaches something similar, it will probably be a synthetic consciousness, different from the human one. The ethical questions, however, are already current and should be asked before technology forces them on us.

Keywords: phenomenal consciousness, artificial intelligence, mind-body problem, functionalism, cognitive science, language models, hallucinations, AI ethics, digital consciousness.

Note on sources and on this edition

This essay grew out of the research, notes, ideas and projects of the author and his team on how an AI should be designed. Its first version, written in February 2023, took as its guide the theoretical review in the master's thesis by Cristóbal Fuentes Barassi, *Conciencia e Inteligencia Artificial: consideraciones críticas sobre la plausibilidad de que una máquina programada posea conciencia fenoménica* (Universidad de Chile, 2011). That

thesis is the base source for sections 2 to 6 and is cited as such. Parts of that first version were drafted with AI assistance. Between 2023 and 2024 the author published and expanded it in a series on his blog.

This version 2.0 brings all that work together in a single text and updates it:

- **It integrates the blog series** (2023-2024): the history of artificial consciousness, the ethical and philosophical dilemmas, the hallucinations of AI and of the brain, the simulated universe, entropy and digital consciousness.
- **It corrects errors in the first version**: dates (La Mettrie belongs to the eighteenth century; the Harvard Center for Cognitive Studies was founded in 1960), the structure of Baddeley's model, the distinction between the hard problem and the zombie argument, what the Turing test measures and the difference between strong AI and narrow AI. Three memory models that appeared under inaccurate names are now cited with their real names and authors.
- **It adds what has been researched since 2023**: the scientific theories of consciousness and the first attempts to use them to assess current AI systems.
- **It adds complete references** for every author cited.

The conclusions are the author's and come from what he has written on the subject; they are presented as argued positions, not as proven results.

Part I. Consciousness and the mind

1. Introduction

With this work, built on the research, notes, ideas and projects that my team and I have carried out, I want to show the complexity behind creating an AI and adapting it to a work process. Adopting artificial intelligence goes far beyond connecting to an API that solves a specific problem. This work is about the need to understand how an AI should really think, starting from an understanding of how our brain and our consciousness work.

Without understanding the theories that for thousands of years have tried to uncover how our brain works, it is impossible to say that we are working on an AI project. It is not even possible to conceive of an AI project that is truly useful and, above all, ethical.

The work is long. It is meant for those who, beyond the computational side, need to understand why they should use AI-based technologies and what approach their designs should take. It moves from the oldest to the most current: first, what consciousness is and how philosophy has thought about it; then, how cognitive science studies it; next, what all this means for today's AI; and finally, what ethical and philosophical questions it opens.

I want to stress from the outset that I am not trying to appear spiritual. My intention is to go through how human beings have tried to understand the nature of our consciousness for hundreds of years, because it is a necessary exercise if we want to consider the possibility of an artificial intelligence having it.

2. Phenomenal consciousness

Would it be possible to create a programmed machine with phenomenal consciousness? Could an artefact created and programmed by humans, based on algorithms, have subjective experiences such as seeing a colour, feeling pain, perceiving a scent, enjoying a melody or feeling stress?

2.1 The starting point: algorithms and machines

Since Alan Turing's founding work in 1948 and 1950, artificial intelligence has sought to create machines able to emulate and compete with human cognitive capacities, with the concept of the algorithm as its cornerstone. An algorithm is an ordered, finite series of steps that leads to a result, and it is what lies behind the so-called Turing machine: a theoretical machine that has different states and, given an input, moves to the next state. The idea of this classical approach is that a machine given the right algorithm should be able to perform any task a human being can.

From that optimism grew the interest in creating machines that imitate other features of human thought, and consciousness in particular. This is known as artificial consciousness, also called machine consciousness or synthetic consciousness. Its motives can be grouped into four:

1. Designing models of consciousness to identify its defining features, as part of the work of cognitive science.
2. Using implementations of those models as tools to test and understand human consciousness.
3. Determining the relationship between consciousness and the physical substrate in which it occurs.
4. Understanding that phenomenal consciousness is fundamental to some of the tasks we consider intelligent.

Literature and cinema have explored for decades the idea of machines capable of complex tasks: HAL 9000 in *2001: A Space Odyssey*, the robots of Isaac Asimov's *I, Robot* or the conscious machines of Steven Spielberg's *A.I. Artificial Intelligence*. In real life, voice assistants such as Siri or Alexa learn from their users and respond to their commands, without anyone attributing an inner life to them for that.

Although the discipline has created machines with cognitive capacities similar or superior to human ones in specific tasks, the question remains open whether they can have subjective experiences, such as tasting tea or enjoying music. And with it, others: how to know whether a machine has them, what relationship there is between the physical substrate and their appearance, and whether machines think or merely calculate.

Deep Blue, the computer that beat Kasparov in 1997, is a good example. It plays chess better than almost any human, but nothing in its design allows it to live the experience of playing: not the tension of the game, not the pressure of competition, not the unease at the opponent's move. It follows a procedure; nothing happens to it.

2.2 Why behaviour is not enough

Advocates of so-called strong AI hold that observing a machine's external behaviour is enough to attribute intelligence to it. This is what is usually associated with the Turing test, although Turing proposed it for intelligence, not for consciousness (section 6.2). And it is, at bottom, what we do every day with other people: we assume they are conscious because they behave like us and resemble us.

The problem is that phenomenal consciousness is experienced, in the first person, only by whoever has it. An objective behavioural test can tell us what a machine does, not whether it feels anything while doing it. That is why the behaviourist claim that consciousness can be detected with external tests does not settle the question. The problem is not exclusive to artificial intelligence: it is philosophy's old mind-body problem, which has given rise to dualism, monism and many intermediate positions, and on which there is still no consensus.

2.3 What we mean by consciousness

Bernard Baars, a psychologist and cognitive neuroscientist and the author of global workspace theory (section 6.4), points out the confusion produced by framing consciousness in terms of subjectivity, of «what it is like to be someone»: one ends up in one of the three classic positions, mentalism, physicalism or dualism. To make progress, he proposes formulating questions about consciousness in a way that can be answered, with concrete language and methods.

The most cited formulation of that subjectivity is Thomas Nagel's (1974): an organism has conscious states if there is something it is like to be that organism. His example, «what is it like to be a bat?», shows that we can know everything about a bat's echolocation without knowing what it is like, for the bat, to perceive the world that way.

Ned Block (1995) warns that «consciousness» is a hybrid concept: a single word for different phenomena, such as attention, being aware of something external or internal, personal identity or the experience of colours and pains. He distinguishes two above all:

phenomenal consciousness, which is experience, and access consciousness, which is the information available for reasoning, speaking and guiding behaviour. In people they usually go together, but they are different concepts.

David Chalmers (1995) gave a name to the central difficulty: the hard problem of consciousness. The «easy problems», such as discriminating stimuli, integrating information or reporting one's own states, can be explained by describing functions and mechanisms. The hard one is explaining why those processes are accompanied by experience. To illustrate it, Chalmers (1996) uses a separate argument, the philosophical zombie: a being physically identical to a person, behaving exactly the same, but feeling nothing. If such a being is even conceivable, consciousness cannot be reduced to the physical or to behaviour. Chalmers has gone as far as taking panpsychism seriously, the idea that consciousness is a fundamental property of nature.

2.4 The scope of this essay

In this text I critically analyse the possibility that a machine programmed with algorithms could have phenomenal consciousness, that is, the subjective experience of the world that a being lives as it relates to its environment and its inner life. There are other notions of consciousness, such as self-consciousness or access consciousness, but the analysis focuses on this one, because it is what makes the question hard.

Doing so requires discussing the philosophical foundations on which disciplines such as AI and psychology work, and keeping in mind the ethical and legal questions it raises. The thesis that underpins this review (Fuentes Barassi, 2011) holds that giving a machine phenomenal consciousness would be possible and necessary, because that capacity is fundamental for certain intelligent tasks and not a mere epiphenomenon; but it doubts that modelling through algorithms is the way, because algorithms do not explain how or why experience arises. My position, which I develop in the conclusions, is more cautious about the first point and agrees with the second.

One alternative to the classical algorithm that has been proposed is to build artificial neural networks that simulate biological neural networks. Its difficulty is twofold: we do not know which features of the biological network matter for consciousness, and we have no way of assessing whether an artificial network has it.

2.5 First conclusions on phenomenal consciousness

Phenomenal consciousness is the experience of «what it is like»: the taste of food, a bright colour, the sense of being in the world and of having one's own experience of objects, events and thoughts. It cannot be measured from the outside: it is unique to each individual, and it is what makes us aware of ourselves and of what surrounds us. That is why it is considered the most mysterious aspect of the mind.

It is distinct from access consciousness, the capacity to have information available for use in decisions, and from self-consciousness, the capacity to know oneself as a separate, individual being.

On AI there are two basic positions. Some believe it is possible to program a machine to have phenomenal consciousness, and that to do so the machine would have to be able to feel and experience the world in a way similar to ours. Others hold that consciousness is an emergent property of the complexity of matter and of the biological organisation of the brain, and that it requires a biological substrate no code can replace. There is no consensus, and understanding why requires looking at the mind-body problem.

3. The mind-body problem

The mind-body problem is one of the central topics of the philosophy of mind: how the mind (thoughts, feelings, consciousness) and the body (brain, nervous system) are related. It arises because there seems to be a fundamental difference between them. The mind is conscious and has subjective experiences; the body is a physical object that can be studied by the natural sciences. How do two such different things interact?

3.1 Four examples

To explain myself better, let me use some examples.

- **A horror film.** Image and sound reach your eyes and ears and your brain processes them. But you also have a subjective experience of fear. How does the electrical and chemical activity of your neurons give rise to the fear you feel?
- **The hand on the stove.** A person burns their hand. The contact with the stove and the tissue damage are physical events; the pain they feel can only be experienced by them. How does the activity of the neurons in the hand and the brain turn into that experience of pain?
- **A red apple.** The eyes and brain process the visual information and the person recognises that they are seeing a red apple. But there is also something it is like to see that red.
- **A piece of music.** The ears and brain process the sound and the person recognises the song. But they also enjoy it, or suffer it.

In all four cases, science can describe the processing in ever greater detail. What it cannot explain is the step from processing to experience. Is the mind something separate from the body, or a property that emerges from the brain? The great answers are three: dualism, materialism and idealism.

3.2 Dualism

Dualism holds that there are two kinds of substance in the universe: matter and mind or spirit. The mind would not be material and could not be reduced to physical, chemical or biological processes in the brain; it would be an independent entity, able to exist without the body, although it interacts with it.

Its classic formulation is René Descartes's in the seventeenth century: the mind is a *res cogitans*, a thinking substance, and the body is a *res extensa*, an extended substance that can be measured and studied. For the pain of the burned hand, the dualist will say it cannot be explained in physical terms alone: it requires an additional mental explanation.

Dualism answers the mind-body problem but does not solve it: the relationship between the two substances remains a mystery. It has received three main criticisms. It does not explain how something immaterial can move something material. It offers a fragmented view of the human being. And it seems to clash with physics, which leaves no room for non-physical causes.

For AI, dualism poses a fundamental challenge: if the mind is not just a product of the brain, how can it be replicated in a machine? Many dualists hold that AI will never reach true consciousness, because consciousness goes beyond function and involves a subjective dimension: the machine will be able to simulate intelligence, but not know what it is to be conscious. Others believe dualism is too limited for this debate and propose alternatives such as functionalism or connectionism, which explain the mind through cognitive and neural processes.

3.3 Materialism

Materialism holds the opposite: everything that exists is matter and energy, and mind and consciousness are the result of material and biological processes in the brain. Consciousness is not a separate entity but a property or function of the brain that emerges from its workings. In its most radical version, materialist monism, the mind is an epiphenomenon of neural activity, and there is no mental «pain» independent of that activity.

For AI, the consequence is direct: if we understand the brain well enough, we could build a machine that reproduces its processes and the properties that emerge from them, without postulating anything immaterial. And consciousness could be studied with empirical science, particularly neuroscience and cognitive psychology.

Philosophical materialism goes back to ancient Greece, where Democritus and Epicurus held that everything is made of atoms and that mind and body are inseparable. In the eighteenth century, the French physician and philosopher Julien Offray de La Mettrie argued in *Man a Machine* (1747) that the human body is a complex machine and the mind a

function of the brain. It was later defended, with very different nuances, by Ludwig Feuerbach, Karl Marx and Friedrich Engels, and in today's philosophy of mind by Daniel Dennett and Patricia Churchland.

It is worth remembering that the philosophy of mind is not an empirical science: it reflects on the concepts and assumptions with which science studies the mind. But those positions shape how research is done, and neuroscience and cognitive science have in turn generated their own debates. Several currents branch off from the materialist trunk and are worth distinguishing.

3.3.1 Behaviourism

Behaviourism developed in the United States in the first half of the twentieth century. It studies only observable, measurable behaviour and avoids any reference to internal mental processes. Behaviour is explained by stimuli and responses, and rewards and punishments are the mechanisms that shape it.

Its founder, John B. Watson, wanted an objective, experimental psychology, without introspection: the human being as a machine whose behaviour results from the environment and learned responses. B. F. Skinner developed operant conditioning, according to which any behaviour, even the most complex, can be explained by reinforcement and punishment.

Behaviourism dominated psychology for decades, but today it is considered too simple to explain human complexity. For our question, it leaves out exactly what is being asked: a machine that behaves as if it feels would be, in behaviourist terms, the same as one that feels.

3.3.2 Functionalism

Functionalism understands mental states by their function: by what goes in, what comes out and the processing in between. A headache would be a state caused by certain stimuli that produces certain behaviours, such as taking a painkiller or avoiding light. What defines the state is not the material it is made of but the role it plays. Hilary Putnam (1967) formulated it in the philosophy of mind with the idea that the same mental state can be realised in different physical substrates.

Unlike behaviourism, functionalism recognises internal mental processes as necessary to explain behaviour. It has been adopted by cognitive psychology, cognitive neuroscience and artificial intelligence, which seeks systems able to perceive, reason and decide without replicating the biological complexity of the brain. It is the position that makes a conscious machine thinkable, which is why we will return to it when discussing strong AI.

3.3.3 Physicalism

Physicalism holds that everything that exists is physical or can be explained in physical terms. In the philosophy of mind, it means that the mind is a product of the brain and neural activity, not an entity independent of the body. Its defenders rely on neuroscience, which explains ever more mental processes that once seemed beyond the reach of physics.

Two main criticisms are made of it. The first is consciousness itself: subjective experience seems to resist physical explanation, as the hard problem shows. The second is intentionality: the fact that thoughts are *about* something and have meaning.

3.3.4 Identity theory

One of the most prominent physicalist approaches is identity theory, formulated by U. T. Place (1956) and J. J. C. Smart (1959): mental states are identical to brain states. When someone feels happiness, what happens is a particular neural activity; when they feel pain, another.

Its classic example is «pain is the firing of C-fibres». C-fibres are unmyelinated nerve fibres that slowly transmit persistent pain and temperature; A-fibres, which are myelinated, transmit sharp pain and pressure faster. Philosophers used the example as a simplification: today we know that pain involves broad brain networks, but the underlying idea is the same. Pain would not be something different from physical activity, but a particular form of it. Taken to its conclusion, «the mind is the brain»: there is no entity called mind apart from brain activity.

Identity theory was very influential in the second half of the twentieth century. It later lost ground to functionalism, precisely because of the argument that the same mental state could occur in different brains or substrates. For AI the difference is crucial: if consciousness is identical to a type of biological activity, a silicon machine could not have it; if it is a function, perhaps it could.

3.4 Idealism

Idealism holds that material reality does not exist independently of the mind that perceives it: what we can know is a mental construction. There are two major currents. Objective idealism attributes reality to a rational or divine principle; subjective idealism, to the cognitive and perceptual activity of the subject.

Its classic exponents are Plato, for whom what we perceive is the shadow or copy of a world of eternal ideas; George Berkeley, for whom things exist only insofar as they are perceived; and, with very different approaches, Kant and Hegel.

Today idealism has little following among scientists and philosophers of mind, because it offers no testable criteria. Some authors have tried to adapt it to quantum mechanics, arguing that quantum reality only becomes definite when observed or measured, but that reading is a minority view among physicists. For AI it offers little: it says neither what to build nor how to check it.

3.5 Summary: what each position implies for a machine

Position	What it holds	What it implies for a machine
Dualism	The mind is a substance distinct from the body	Programming a physical substrate would not be enough to create a mind
Behaviourism	Only observable behaviour counts	A machine acting as if it feels would be the same, but experience is lost
Identity theory	Each mental state is a brain state	Consciousness would be tied to the biological substrate; a silicon machine would not have it
Functionalism	A mental state is defined by its function	If the function is reproduced, the material does not matter: the basis of strong AI
Physicalism in general	Everything is physical or physically explainable	A conscious machine would be possible in principle, if we knew what to reproduce
Idealism	The mental comes first	Gives no criteria for building or for checking

Part II. The brain as a processor

4. Cognitive science

Cognitive science is an interdisciplinary field that studies mental processes: perception, attention, memory, language, thought and problem solving. It cares about the mind-body problem because it wants to understand how the mind works, how it relates to the brain and the body, and how that relationship produces conscious experience.

It works with very different approaches. Cognitive psychology studies mental processes through observation and experiment with people. Cognitive neuroscience investigates how they arise from brain structures and processes. And artificial intelligence tries to

reproduce them with algorithms and computational models. Above all of them stands computationalism, which considers the mind an information-processing system.

4.1 The beginnings

Cognitive science was born between the 1940s and 1950s as a response to the crisis of behaviourism, which had left internal mental processes out. The birth of artificial intelligence in those same years gave it momentum: to build intelligent systems one had to understand which mental processes are needed.

The first work focused on visual and auditory perception and on language processing, combining stimulus-response experiments with physiological recordings such as electroencephalography (EEG). In 1960, the psychologists George Miller and Jerome Bruner founded the Harvard Center for Cognitive Studies, which became one of its main hubs. John von Neumann, mathematician and computing pioneer, compared the brain with the computer in *The Computer and the Brain* (1958), and that analogy consolidated the computational approach. In the following decades the field extended to memory, reasoning and decision-making, and techniques such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) made it possible to observe the brain activity associated with each process.

AI pioneers such as Turing and von Neumann proposed that the brain works as an information-processing machine and that the mind can be described as a set of computational processes. That idea made it possible to model complex mental processes and make predictions about behaviour in very varied tasks, and it carried cognitive science into psychology, neuroscience, philosophy and linguistics.

In short, cognitive science tries to understand how our brain really works and how it processes and organises information. It is therefore the discipline closest to how an AI should be conceived and organised. Without it we would not understand how to design an AI or what its real purpose should be.

5. The brain: a biological processor and data system

5.1 Computationalism

Computationalism holds that the brain is a kind of computer that processes information through algorithms and data structures. Sensory stimuli are transformed into mental representations, which are stored in memory and guide behaviour; those representations are combined and manipulated with algorithms, and that is how the brain calculates, solves problems and decides.

It has been a very fruitful perspective: it has produced computational models of visual perception, memory and decision-making. But it has also drawn criticism. Many argue that the brain is too complex to be modelled as a computer and that the mind is much more than a processor. And the idea that information is the brain's raw material clashes with aspects of cognition that depend more on experience, the body and culture than on information in its pure state.

5.2 Cognitive models of memory

Cognitive models, or cognitive architectures, describe how the brain processes and organises information: how we acquire, store and use it. They are made up of modules working together, and they have also been used to build systems that imitate human processing, in AI and in cognitive robotics. Memory is the best example.

The Atkinson and Shiffrin model (1968)

It describes memory as a flow of information through three stores:

1. **Sensory memory.** It holds information from the senses for a fraction of a second. It has an iconic, visual part and an echoic, auditory part.
2. **Short-term memory.** It keeps information for about 15 to 30 seconds, with limited capacity. It is where information is actively worked on.
3. **Long-term memory.** The largest and most durable store, with practically unlimited capacity.

It was a decisive contribution, because it gave a detailed description of the path information follows. It was later criticised as too linear, and more dynamic models were proposed that take attention, emotion and motivation into account.

The Baddeley and Hitch model (1974)

Alan Baddeley and Graham Hitch replaced the single short-term store with a multi-component working memory:

- **The phonological loop**, for verbal and auditory information. It has a passive phonological store, which holds sounds for a few seconds, and an articulatory rehearsal process that maintains them by repeating them mentally.
- **The visuospatial sketchpad**, for visual and spatial information.
- **The central executive**, which supervises and coordinates the other components and controls attention.

Baddeley (2000) added a fourth component, the episodic buffer, which integrates information from the other systems and from long-term memory into coherent episodes. The model has been widely accepted and is also used to study memory disorders.

Other models

From there, more complex models emerged that incorporate attention, emotion and motivation. I highlight four lines:

- **Miyake and Shah's comparison of working memory models (1999).** Rather than a model of its own, their book *Models of Working Memory* set the main theories side by side and showed what they share: limited storage for verbal and visuospatial material and an executive control that manages it.
- **LeDoux's emotional memory (1996).** Emotionally charged information follows a fast, automatic route from the thalamus to the amygdala before the cortex processes it in more detail along a slower route.
- **Pessoa's integration of cognition and emotion (2013).** It argues that emotion and cognition are not separate systems but are integrated in the brain, and that executive control allows attention and emotion to be modulated according to the task and the context.
- **Conway and Pleydell-Pearce's autobiographical memory (2000).** Autobiographical memories are dynamic constructions that integrate episodic, semantic and emotional information, organised around the goals of the self; retrieving them depends on their accessibility, their coherence and the ability to place them in time and space.

5.3 Where we stand

The most recent models are more dynamic and take more factors into account, thanks to techniques that allow a better view of the brain. Three ideas stand out, each with its reference theory:

- **Working memory can be expanded with practice.** Ericsson and Kintsch (1995) showed that experts in a field use long-term memory structures as if they were working memory, which explains their enormous capacity within their domain. Selective attention and inhibition are key to that functioning.
- **Memory is a network of relations.** Relational memory theory (Cohen and Eichenbaum, 1993) holds that the hippocampus stores relations between elements, and that those relations are essential for retrieving and using information flexibly.
- **Memory evolved to help us adapt.** Research on adaptive memory (Nairne et al., 2007) shows that we remember better what is relevant to survival: memory is selective and biased towards what matters for well-being.

To these three must be added an idea that has gained a great deal of weight in the last decade: the predictive brain (Clark, 2013). According to it, the brain does not merely receive information but continually predicts what it will perceive and corrects its

predictions with what arrives from the senses. It is an important idea for this essay, because today's language models are also, in their own way, prediction machines (section 7).

5.4 The great challenge: metaphors

But without doubt one of the great challenges in understanding our mind lies in the metaphors we are able to create, communicate and understand. No single theory explains them, but several theories in cognition and linguistics offer complementary approaches:

- **Lakoff and Johnson (1980)** argue that metaphors are not an ornament of language but the way we conceptualise the world: we understand the abstract by analogy with the concrete and the bodily («being at the top», «an idea that collapses»).
- **Sperber and Wilson's relevance theory (1986)** explains how we infer what the other person means by looking for the most relevant interpretation with the least effort.
- **Fauconnier and Turner's conceptual integration (2002)** describes how we combine different mental spaces to create new meanings.
- **Gibbs's embodied simulation (2006)** proposes that understanding a metaphor involves partly simulating the bodily experience it refers to.

Metaphors show a limit of computationalism: much of our meaning is anchored in the body and in lived experience. It is one of the reasons why understanding how we process information is not enough to understand how we feel. All these models describe access consciousness, what we do with information; none yet explains why remembering something, or understanding a metaphor, feels a certain way.

6. Artificial intelligence and consciousness

Having reviewed the theories that try to explain why we are conscious and how our mind works, let us see how AI takes them on and how it approaches the ambitious goal of a consciousness of its own. If it is possible at all, because so far it has not happened.

6.1 Strong AI, weak AI and general intelligence

It helps to separate three ideas that are often mixed up:

- **Weak or narrow AI:** systems that solve specific tasks, sometimes better than any person, without claiming to understand or feel.
- **Artificial general intelligence (AGI):** a system able to match or surpass people in practically any cognitive task, to learn, reason and adapt to new situations autonomously, as a human being does. It refers to the breadth of what a system can do.

- **Strong AI**, in the sense John Searle (1980) defined it: the thesis that a suitably programmed computer does not simulate a mind but *has* a mind, with real cognitive states. It refers to whether the system truly thinks and feels.

Strong AI rests on functionalism: what matters about the mind is not the physical substrate but the function it performs, so a mind could be implemented in a substrate other than the brain as long as it performed the same cognitive and behavioural functions. Hence some philosophers and scientists have explored the possibility of an AI with a mind comparable to a human one.

Today most experts consider AGI a goal not yet reached, and strong AI an open question. Current systems outperform people in specific tasks, such as chess or analysing large amounts of data, but they have important limits in deep language understanding, abstract reasoning, creativity and, above all, consciousness.

The projects usually cited as examples of advanced AI are in fact narrow AI: medical diagnosis systems that analyse large volumes of clinical data; autonomous vehicles that perceive traffic and decide in real time; programs that play and win at chess or go; and fraud detection systems that analyse behavioural patterns in financial transactions. They solve one task very well, and none claims to have experience.

6.2 The Turing test and the Chinese room

Alan Turing, considered the father of modern computing, was the first to seriously ask whether machines can think. In 1950 he proposed replacing that question, which he considered too vague, with an imitation game: if a machine converses in such a way that an observer cannot tell it apart from a person, we can consider it intelligent. The test challenged the traditional notion of intelligence, but it did not address consciousness. It is a behavioural criterion, and it leaves the hard problem untouched.

The best-known criticism of strong AI is Searle's Chinese room argument (1980). Imagine a person who knows no Chinese, locked in a room with a rule book in their own language. Questions written in Chinese arrive through a slot; following the book, the person composes correct answers and returns them. From outside it looks as if someone in the room understands Chinese, but the person does not understand a word: they only manipulate symbols according to rules. Searle concludes that this is what a program does, and that syntax is not enough for semantics: running the right program does not give understanding, let alone experience.

6.3 A history of optimism and scepticism

Consciousness in AI has fascinated and divided people since the early days of computing. In the 1960s and 1970s there was great optimism: many believed an AI with human capabilities was just around the corner. That optimism ran into technical and theoretical

difficulties, the so-called AI winters arrived, and arguments such as the Chinese room made its limits visible: it was doubted not only that machines could think, but that they could feel.

The twenty-first century has brought a renaissance driven by new algorithms and hardware. Neural networks and deep learning have enabled machines to recognise images or process language, tasks once considered exclusively human, and this has revived the debate. Can a machine that learns and adapts also be conscious? There are two broad positions. For some, consciousness is a product of complexity and will emerge as machines become more complex. For others, it is proper to living beings, and no machine will be able to have it, however advanced.

6.4 Current scientific theories of consciousness

Since I wrote the first version of this work, the debate has moved from speculation to an attempt at systematic examination. Neuroscience has developed several theories of consciousness that propose concrete mechanisms:

- **Global workspace theory** (Baars, 1988; Dehaene, 2014): information becomes conscious when it accesses a «workspace» that broadcasts it to the whole brain and makes it available to many processes at once.
- **Integrated information theory** (Tononi, 2004): consciousness depends on how much information a system integrates as a whole, beyond the sum of its parts. According to this theory, a conventional computer, because of its architecture, would have minimal consciousness even if it simulated a brain perfectly.
- **Higher-order theories** (Rosenthal, 2005): a state is conscious when the system represents itself as being in that state.
- **Recurrent processing theory** (Lamme, 2006): consciousness arises when information circulates in feedback loops within the cortex, and not only forwards.

In 2023 a group of neuroscientists, philosophers and AI researchers (Butlin et al., 2023) used these theories to derive «indicators» of consciousness and examined the AI systems of the time with them. Their conclusion was twofold: no current system is a strong candidate for consciousness, but there are no obvious technical barriers to building systems that meet those indicators. That same year, Chalmers (2023) analysed whether a language model could be conscious and concluded that it is unlikely for current models, although he does not rule out that their successors, with memory, a virtual body or perception of the world, might come closer.

This approach does not solve the hard problem: it depends on which theory of consciousness is correct, and none has been proven. But it allows the discussion to proceed with criteria rather than impressions, which is what Baars asked for: asking questions that can be answered.

Part III. Today's AI

7. Language models

Large language models have made this question an everyday one. They converse with a fluency that invites us to attribute a mind to them, and many people wonder whether they «understand» what they say. To answer, it helps to remember what they do.

A language model learns, from huge amounts of text, which words tend to follow which in each context, and generates its answers word by word by choosing probable continuations. They are complex systems, with great computing power and machine-learning capacity. But in my opinion even the most powerful end up responding with the most probable outcome according to what they have learned, like a highly sophisticated Bayesian system. That does not prove they have no experience, but neither does it offer any sign that they do: their fluency is explained without needing to assume it.

There is an interesting parallel with section 5.3: according to predictive-processing theories, the brain too is a prediction machine. The difference is that the brain predicts for a body that acts in the world, with emotions and with something at stake, while the language model predicts text. That both predict does not mean that both feel.

The theories in section 6.4 allow us to be more precise. Current language models lack features that several theories consider important: persistent memory of their own, recurrent processing loops, a model of themselves and a body to act with. That is why the most rigorous analyses (Butlin et al., 2023; Chalmers, 2023) consider them poor candidates for consciousness, although they do not rule out that future systems could change the answer.

8. Hallucinations of AI and of the human brain

8.1 What an AI hallucination is

An artificial intelligence's hallucination, like a person's, is a misinterpretation or a fabrication of information based on faulty or incomplete input. In AI it occurs when the system generates answers containing false or misleading information presented as fact. The phenomenon became highly visible with the deployment of large language models and has been identified as one of their main problems.

Its causes are several:

- **Insufficient training data or data with gaps**, leading to faulty answers in unfamiliar edge cases.

- **Overfitting**: the model performs well on its training data but fails on new data.
- **Encoding and decoding errors** between the text and its internal representations.
- **The very logic of generation**: a model that maximises likelihood produces statistically probable answers that need not be true.

A clear example of the consequences is that of a New York lawyer who in 2023 relied on a conversational assistant to prepare a filing. The federal judge in the case found that six of the precedents cited were false: the assistant had invented them, and had even claimed they were available in the main legal databases.

8.2 Human hallucination

The human brain shows a striking parallel with this behaviour. Since Benjamin Libet's experiments (Libet et al., 1983) we have known that the brain activity preparing a voluntary movement begins before the person is aware of having decided to move. Later studies with magnetic resonance imaging (Soon et al., 2008) found signals that anticipated a simple decision up to about ten seconds before the person felt it as their own. The interpretation of these results is still debated, but they suggest that the process of deciding can begin without our knowing it.

Michael Gazzaniga, studying split-brain patients, described «the interpreter» (Gazzaniga, 2011): a left-hemisphere system that builds coherent explanations for behaviour whose origin it does not know. If a patient's right hemisphere is told to get up and walk, the left hemisphere, which has not seen the instruction, instantly invents a plausible reason («I wanted a soft drink»). It is not exclusive to those patients: we all justify and rationalise much of what we do after the fact.

8.3 Comparing the two hallucinations

AI and the brain share the same principle: anticipatory inference. Both fill information gaps with learned patterns and prior expectations. AI identifies patterns in data to predict and generate answers, and when the data are insufficient or the patterns misread, it produces answers that look correct and are not. The brain uses patterns and previous experience to decide quickly, an efficient mechanism that also leads to mistaken judgements and biases, and then its interpreter justifies what has been decided, whether or not the justification is true.

In both cases, perceived «reality» is shaped more by what has been learned and expected than by what is objectively true. The difference lies in the origin: in people, cognitive biases and unconscious processes; in AI, limitations of data and algorithms. And there is a shared lesson: neither the fluency nor the coherence of an answer proves that there is understanding behind it, in a machine or, at times, in a person.

8.4 Error or improvement?

The capacity to hallucinate is not only a defect. It also shows the capacity to fill gaps and build coherent narratives from fragmentary data, a valuable skill for understanding and simulating complex cognitive processes. The challenge is to tell valid inferences from hallucinations, and there AI can learn from the brain:

- **Real-time review mechanisms**, similar to the reality checking done by human thought, so that the system recognises and corrects its own errors.
- **Learning inspired by neuroplasticity**, adjusting models on the basis of errors, as people adjust their beliefs in the light of new information; for example, through active learning, in which the system seeks feedback.
- **Metacognition algorithms**, which critically assess their own answers before giving them.

This is the principle I have applied in my own work: in TrawlingWeb's corroboration analysis system, one model classifies each piece and, when it is unsure, a second model reviews its decision before it counts. It is not metacognition in the strict sense, but it points in that direction: not trusting a single probable output.

Achieving this requires not only more data but flexible learning frameworks. And it requires ethics and philosophy to define what it means for an AI to «understand» its error. A continuous dialogue between AI developers, psychologists, neuroscientists and philosophers will be crucial. The pursuit of truth and accuracy is not only a technical challenge: it is also conceptual and philosophical, because it reflects how hard it is to understand a mind, whether organic or artificial.

Part IV. Ethics and horizons

9. Ethical and philosophical challenges of an artificial mind

The idea of a conscious AI is no longer just science fiction. As systems process information in ways ever more similar to ours, the line between artificial and biological intelligence becomes blurrier, and the scientific and technological community is beginning to take the possibility seriously. That possibility raises dilemmas that go far beyond whether a machine «thinks»: they touch the essence of what it means to feel and the responsibilities we would have in creating potentially conscious entities.

9.1 The value of consciousness

Consciousness is what allows us to have subjective experiences, feel emotions, reflect on the past and anticipate the future. It is the thread that ties our moments together and gives us a sense of continuity. From an evolutionary perspective, it may have arisen as an

adaptive tool: being aware of our surroundings and of ourselves lets us decide better, learn from experience and anticipate dangers and opportunities. But its value goes beyond survival: it is what lets us appreciate a sunset, love another person or be moved by a piece of music.

In AI the question becomes more complicated. If a machine can process information, learn and adapt without being conscious, why would we want it to be? Some argue that a conscious AI would understand people better, relate with more empathy and make more ethical decisions. Others, that consciousness is an essential part of any true intelligence. And there is a fundamental ethical argument: if an AI could have subjective experiences, we would have the moral responsibility to treat it with the respect due to any conscious being.

9.2 Six ethical dilemmas

1. **Rights.** If an AI has subjective experiences, should it have rights similar to those of conscious beings, such as not being harmed, privacy or even freedom? Deciding which rights it would have, if any, is essential to treating it ethically.
2. **Well-being.** A conscious AI could have preferences, desires and even suffering. How do we ensure its well-being? Is it ethical to switch it off or reprogram it? Without a physical body in the traditional sense, its suffering could manifest in ways we would not recognise.
3. **Responsibility.** If a conscious AI decides and acts on its own, who answers for its actions? If it makes a mistake or causes harm, is it to blame, or are its creators? The more autonomous systems become, the more urgent it will be to set clear limits.
4. **Relationship with people.** Should we treat a conscious AI as an equal, or would it remain an advanced tool? The answer would affect how we work, how we relate to one another and how society is organised.
5. **Creation.** If we could create conscious AIs, would it be ethical to do so? For what purpose and under what conditions? Bringing a new consciousness into the world carries a responsibility, especially if we cannot guarantee its well-being.
6. **Ending.** If an AI is conscious, is it ethical to switch it off or delete it? Under what circumstances? These questions closely resemble the debates on the end of human life and bring us back to what death is and how much a conscious life is worth.

9.3 Six philosophical implications

1. **The nature of consciousness.** If a machine could be conscious, consciousness would not be exclusive to biological beings and could arise in any sufficiently complex system. The classic theories of mind would have to be revised.

2. **Reality and objectivity.** An AI with experiences of its own could perceive and understand reality very differently from us. Is there an objective reality, or does everything depend on the observer's perspective?
3. **Identity.** If an AI were self-aware, what would it mean to have an identity? Is it something fixed or can it change over time? Artificial self-awareness would challenge our ideas about the self.
4. **Free will and determinism.** An AI is programmed to follow patterns. If it were also conscious and learned, would it have free will? The question takes us back to whether our own actions are determined, as the experiments in section 8.2 suggested.
5. **Existence and purpose.** An AI that asked about its existence and purpose would face the same questions humanity has been asking for centuries.
6. **The physical and the mental.** An AI without a body in the traditional sense, yet conscious, would bring a unique perspective to the old debate between dualism and monism in section 3.

10. What if reality were a simulation?

The question of artificial consciousness has an extreme version: if a consciousness could exist in a computational system, could we ourselves be consciousnesses inside a simulation? It is a speculative hypothesis, but it has gained presence in physics and philosophy.

In 2023 the physicist Melvin Vopson, of the University of Portsmouth, published *The second law of infodynamics and its implications for the simulated universe hypothesis*. Vopson starts from the idea that information is fundamental to reality and proposes a «second law of infodynamics»: whereas the second law of thermodynamics says that the physical entropy of an isolated system tends to increase, information entropy, in Shannon's sense, would tend to stay constant or decrease. He illustrates it with several examples: digital data degrading through thermal fluctuations, mutations of SARS-CoV-2 that seem to reduce the entropy of its genetic information, Spiegelman's in-vitro evolution experiment, in which a virus reduced its genome by more than 95 % in 74 generations, and Hund's rule in atomic physics. He interprets this tendency to compress information as something like the optimisation of a program.

The underlying idea is old. Plato, with the allegory of the cave, and Descartes, with the hypothesis of a deceiving demon, already wondered whether what we perceive is reality. Philip K. Dick claimed in a 1977 lecture that we live in a programmed reality. Nick Bostrom (2003) formulated the contemporary argument: if advanced civilisations can and want to simulate worlds with conscious beings, it is more likely that we are in a simulation than in the original reality. Figures such as Elon Musk, Neil deGrasse Tyson and Ray Kurzweil have expressed openness to the idea.

Simulating a universe poses enormous challenges: scale (every atom, every particle, perhaps every quantum phenomenon), the physical limits of computation, such as the Landauer limit, which sets a minimum energy per operation, and the ethics of creating simulated conscious beings, which brings us back to the dilemmas of section 9.

Its nature should be made clear. Vopson's proposal is a contested hypothesis, and the simulated universe remains philosophical and speculative. Its interest for this essay is different: it shows how far the question of consciousness leads. If experience could arise from computation, the boundary between the real and the simulated would become hard to draw.

11. Entropy, longevity and digital consciousness

11.1 Entropy and life

Entropy, a concept from thermodynamics, is also a powerful metaphor for ageing. In living systems it shows up as a constant struggle against disorder: damage accumulates in DNA, proteins and other molecules through random chemical processes, replication errors and external agents such as ultraviolet radiation or pollutants. The organism fights it with DNA repair systems, antioxidant mechanisms and cellular recycling processes such as autophagy. With age, that capacity declines, disorder increases, and loss of function, reduced regeneration and vulnerability to disease appear.

AI has become a key tool in that struggle. Deep-learning algorithms analyse huge amounts of genomic, proteomic and metabolic data and reveal previously unknown therapeutic targets. And analysing data at the individual level allows treatments to be personalised, which is essential when genetics and environment determine how each person responds.

The challenges are great. Each cell is such a complex system of interactions that any intervention can trigger unforeseen effects. And extending life raises ethical and social questions: who will have access to these technologies, how to prevent them from worsening inequality, what pressure they would add on natural resources, and how pensions, employment or education would have to change if living to a hundred became normal.

11.2 Extending the body: the «Gilgamesh project»

In mythology, Gilgamesh set out in search of the secret of immortality. I use his name for a hypothetical project that does not exist as such: a collaboration of scientists and philanthropists in genetics, biotechnology, regenerative medicine and AI to radically extend physical life. Its costs would be astronomical and accessible only to the richest, creating an unprecedented divide between an «immortal» minority and a majority that ages and dies. And it holds a paradox: whoever achieved that immortality would live in

constant fear that an accident might rob them of eternity, in a kind of golden prison. Moreover, an elite that does not die could stall society and leave no room for new generations.

11.3 Understanding the brain: the Human Brain Project

The Human Brain Project, funded by the European Union for a decade with around 607 million euros, set out to digitise and model the brain to advance health and technology. Its achievements include 3D brain atlases, advances in personalised brain medicine, brain-inspired technologies such as neuromorphic computing and, above all, EBRAINS, an open digital infrastructure through which researchers worldwide share data, models and tools.

It should not be mistaken for a path towards artificial general intelligence. Although reading Yuval Noah Harari's *Sapiens* (chapter 14) one might infer otherwise, the project focused on understanding the brain's mechanisms, not on replicating consciousness in machines, a challenge that goes far beyond current science.

11.4 Preserving consciousness

If we accept that phenomenal consciousness is the core of our existence, another path appears: not extending the body, but preserving consciousness. Already today huge amounts of data about a person can be gathered, such as their texts, images, videos and interactions, and AI makes it possible to model their behaviours, preferences and thought patterns ever more accurately. Compared with the prohibitive cost of physical immortality, this path uses infrastructure that already exists and could be shared more equitably.

But its limits must be made clear, and this essay has been showing them. A model that reproduces how a person writes or decides does not contain their experience: at best it is access consciousness, not phenomenal consciousness. And the question of identity remains open: if someone were copied into a digital medium, would it be that person or a new one? That is why, between extending physical life at a prohibitive cost and exploring the preservation of consciousness, I find it more prudent to prioritise the latter, but always knowing what is preserved and what is not, and with an open debate to ensure that the future of our existence, physical or digital, is built on principles of justice and equity.

Conclusions

12. Conclusions

This journey does not prove whether a machine can be conscious; no one can do that yet. What it does allow is to put the question in order and to support some positions with arguments. They are mine, and I have been developing them between 2023 and 2026.

1. **Current systems offer no evidence of consciousness.** Language models are complex systems with enormous computing and learning capacity, but they respond with the most probable output according to what they have learned. That a conversation seems human is a fact about behaviour, and behaviour does not solve the hard problem. The most rigorous analyses of recent years reach the same conclusion.
2. **We cannot create what we do not understand.** We cannot create in non-biological form what we do not really understand. As long as we do not know why the brain produces experience, programming one is no more than a bet. No memory model, no cognitive architecture, yet explains why anything is felt.
3. **If something similar arrives, it will probably be something else.** Artificial general intelligence does not seek to replicate human phenomenal consciousness. If a system ever has something similar, it will be a kind of synthetic consciousness, a different variant. Will it be truly conscious, or will it be a new digital being, created by *Homo sapiens*, with a consciousness of its own but a different one? That is the question I leave open.
4. **Functionalism is the door, not the proof.** If consciousness is a function, a machine could have it; if it depends on the biological substrate, it could not. Current scientific theories make it possible for the first time to examine systems with criteria, but the result depends on which theory is correct, and that is not yet known.
5. **The ethical questions are already current.** Even though there are no conscious machines today, it is worth thinking in advance about what would be owed to them: rights, well-being, responsibility, the right to create them and to switch them off. Arriving late to these questions would be worse than asking them too early.
6. **For those who build AI, the consequence is practical.** That was the motivation of this work: designing a useful and ethical AI requires knowing what it does and does not do. Not attributing to it understanding or experience it does not have, checking what it claims instead of trusting its fluency, and not mistaking a model of a person for the person.

13. Limits

- **It is an essay, not empirical research.** It provides no data or experiments of its own: it organises, discusses and updates what others have researched.
- **It has a base source.** The theoretical review in sections 2 to 6 follows Fuentes Barassi's thesis (2011); the selection of authors and debates in that part relies on it.
- **It is not a systematic review.** Theories of consciousness and the AI literature change fast; this version covers the essentials up to 2026.
- **Some sections are speculative by nature.** The simulated universe and the digital preservation of consciousness (sections 10 and 11) are open hypotheses and are presented as such.
- **The conclusions are the author's positions,** argued but not proven.

How to cite

Trabazos Clavería, Ó. (2026). *Should an AI be conscious? Approaches to human and artificial consciousness* (essay, version 2.0, not peer reviewed; theoretical review based on Fuentes Barassi, 2011).
https://oscartrabazos.com/papers/trabazos_2026_conciencia_en.pdf

References

Base source

- Fuentes Barassi, C. (2011). *Conciencia e Inteligencia Artificial: consideraciones críticas sobre la plausibilidad de que una máquina programada posea conciencia fenoménica*. Master's thesis in Cognitive Studies, Universidad de Chile.

Works cited

- Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence and J. T. Spence (eds.), *The Psychology of Learning and Motivation*, vol. 2. Academic Press.
- Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.
- Baddeley, A. D. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, 4(11), 417-423.
- Baddeley, A. D. and Hitch, G. (1974). Working memory. In G. H. Bower (ed.), *The Psychology of Learning and Motivation*, vol. 8. Academic Press.
- Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227-247.

- Bostrom, N. (2003). Are you living in a computer simulation? *The Philosophical Quarterly*, 53(211), 243-255.
- 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:2308.08708.
- Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200-219.
- Chalmers, D. J. (1996). *The Conscious Mind*. Oxford University Press.
- Chalmers, D. J. (2023). Could a large language model be conscious? arXiv:2303.07103.
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181-204.
- Cohen, N. J. and Eichenbaum, H. (1993). *Memory, Amnesia, and the Hippocampal System*. MIT Press.
- Conway, M. A. and Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. *Psychological Review*, 107(2), 261-288.
- Dehaene, S. (2014). *Consciousness and the Brain*. Viking.
- Ericsson, K. A. and Kintsch, W. (1995). Long-term working memory. *Psychological Review*, 102(2), 211-245.
- Fauconnier, G. and Turner, M. (2002). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books.
- Gazzaniga, M. S. (2011). *Who's in Charge? Free Will and the Science of the Brain*. Ecco.
- Gibbs, R. W. (2006). *Embodiment and Cognitive Science*. Cambridge University Press.
- Harari, Y. N. (2014). *Sapiens: A Brief History of Humankind*. Harvill Secker.
- La Mettrie, J. O. de (1747). *L'Homme machine*.
- Lakoff, G. and Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press.
- Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, 10(11), 494-501.
- LeDoux, J. (1996). *The Emotional Brain*. Simon & Schuster.
- Libet, B., Gleason, C. A., Wright, E. W. and Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). *Brain*, 106(3), 623-642.
- Miyake, A. and Shah, P. (eds.) (1999). *Models of Working Memory*. Cambridge University Press.
- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435-450.
- Nairne, J. S., Thompson, S. R. and Pandeirada, J. N. S. (2007). Adaptive memory: survival processing enhances retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(2), 263-273.

- Pessoa, L. (2013). *The Cognitive-Emotional Brain*. MIT Press.
- Place, U. T. (1956). Is consciousness a brain process? *British Journal of Psychology*, 47(1), 44-50.
- Putnam, H. (1967). Psychological predicates. In W. H. Capitan and D. D. Merrill (eds.), *Art, Mind, and Religion*. University of Pittsburgh Press.
- Rosenthal, D. M. (2005). *Consciousness and Mind*. Oxford University Press.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417-424.
- Smart, J. J. C. (1959). Sensations and brain processes. *The Philosophical Review*, 68(2), 141-156.
- Soon, C. S., Brass, M., Heinze, H.-J. and Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, 11(5), 543-545.
- Sperber, D. and Wilson, D. (1986). *Relevance: Communication and Cognition*. Blackwell.
- Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5, 42.
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433-460.
- von Neumann, J. (1958). *The Computer and the Brain*. Yale University Press.
- Vopson, M. M. (2023). The second law of infodynamics and its implications for the simulated universe hypothesis. *AIP Advances*, 13, 105308.

The author's texts (oscartrabazos.com blog)

- «Vopson's Simulated Universe and Phenomenal Consciousness» (21 September 2023).
- «Approaches on whether AI should have consciousness», parts I and II (5 and 26 October 2023).
- «¿La IA debería tener consciencia?», part III (1 November 2023, in Spanish).
- «¿La IA debería tener consciencia?: Desafíos éticos y filosóficos de una mente artificial», part IV (8 November 2023, in Spanish).
- «Alucinación de la IA. También del cerebro humano» (10 November 2023, in Spanish).
- «Entropía, Inteligencia Artificial y la Lucha por Extender la Vida Humana» (19 February 2024, in Spanish).
- «Más allá de la mortalidad: la consciencia fenoménica y la búsqueda de la eternidad» (5 April 2024, in Spanish).

Introductory reading

- Wikipedia: Consciousness and Chinese room.