Artificial consciousness

Artificial consciousness,[1] also known as **machine consciousness**,[2][3] **synthetic consciousness**,[4] or **digital consciousness**,[5] is the consciousness hypothesized to be possible in artificial intelligence.[6] It is also the corresponding field of study, which draws insights from philosophy of mind, philosophy of artificial intelligence, cognitive science and neuroscience.

The same terminology can be used with the term "sentience" instead of "consciousness" when specifically designating phenomenal consciousness (the ability to feel qualia).[7] Since sentience involves the ability to experience ethically positive or negative (i.e., *valenced*) mental states, it may justify welfare concerns and legal protection, as with animals.[8]

Some scholars believe that consciousness is generated by the interoperation of various parts of the brain; these mechanisms are labeled the neural correlates of consciousness or NCC. Some further believe that constructing a system (e.g., a computer system) that can emulate this NCC interoperation would result in a system that is conscious.[9]

Philosophical views

As there are many hypothesized types of consciousness, there are many potential implementations of artificial consciousness. In the philosophical literature, perhaps the most common taxonomy of consciousness is into "access" and "phenomenal" variants. Access consciousness concerns those aspects of experience that can be apprehended, while phenomenal consciousness concerns those aspects of experience that seemingly cannot be apprehended, instead being characterized qualitatively in terms of "raw feels", "what it is like" or qualia.[10]

Plausibility debate

Type-identity theorists and other skeptics hold the view that consciousness can be realized only in particular physical systems because consciousness has properties that necessarily depend on physical constitution.[11][12][13][14] In his 2001 article "Artificial Consciousness: Utopia or Real Possibility," Giorgio Buttazzo says that a common objection to artificial consciousness is that, "Working in a fully automated mode, they [the computers] cannot exhibit creativity, unreprogrammation (which means can 'no longer be reprogrammed', from rethinking), emotions, or free will. A computer, like a washing machine, is a slave operated by its components."[15]

For other theorists (e.g., functionalists), who define mental states in terms of causal roles, any system that can instantiate the same pattern of causal roles, regardless of physical constitution, will instantiate the same mental states, including consciousness.[16]

Thought experiments

David Chalmers proposed two thought experiments intending to demonstrate that "functionally isomorphic" systems (those with the same "fine-grained functional organization", i.e., the same information processing) will have qualitatively identical conscious experiences, regardless of whether they are based on biological neurons or digital hardware.[17][18]

The "fading qualia" is a *reductio ad absurdum* thought experiment. It involves replacing, one by one, the neurons of a brain with a functionally identical component, for example based on a silicon chip. Since the original neurons and their silicon counterparts are functionally identical, the brain's information processing should remain unchanged, and the subject would not notice any difference. However, if qualia (such as the subjective experience of bright red) were to fade or disappear, the subject would likely notice this change, which causes a contradiction. Chalmers concludes that the fading qualia hypothesis is impossible in practice, and that the resulting robotic brain, once every neurons are replaced, would remain just as sentient as the original biological brain.[17][19]

Similarly, the "dancing qualia" thought experiment is another *reductio ad absurdum* argument. It supposes that two functionally isomorphic systems could have different perceptions (for instance, seeing the same object in different colors, like red and blue). It involves a switch that alternates between a chunk of brain that causes the perception of red, and a functionally isomorphic silicon chip, that causes the perception of blue. Since both perform the same function within the brain, the subject would not notice any change during the switch. Chalmers argues that this would be highly implausible if the qualia were truly switching between red and blue, hence the contradiction. Therefore, he concludes that the equivalent digital system would not only experience qualia, but it would perceive the same qualia as the biological system (e.g., seeing the same color).[17][19]

Critics[*who?*] of artificial sentience object that Chalmers' proposal begs the question in assuming that all mental properties and external connections are already sufficiently captured by abstract causal organization.

Controversies

In 2022, Google engineer Blake Lemoine made a viral claim that Google's LaMDA chatbot was sentient. Lemoine supplied as evidence the chatbot's humanlike answers to many of his questions; however, the chatbot's behavior was judged by the scientific community as likely a consequence of mimicry, rather than machine sentience. Lemoine's claim was widely derided for being ridiculous.[20] However, while philosopher Nick Bostrom states that LaMDA is unlikely to be conscious, he additionally poses the question of "what grounds would a person have for being sure about it?" One would have to have access to unpublished information about LaMDA's architecture, and also would have to understand how consciousness works, and then figure out how to map the philosophy onto the machine: "(In the absence of these steps), it seems like one should be maybe a little bit uncertain. [...] there could well be other systems now, or in the relatively near future, that would start to satisfy the criteria."[21]

Testing

Qualia, or phenomenological consciousness, is an inherently first-person phenomenon. Because of that, and the lack of an empirical definition of sentience, directly measuring it may be impossible. Although systems may display numerous behaviors correlated with sentience, determining whether a system is sentient is known as the hard problem of consciousness. In the case of AI, there is the additional difficulty that the AI may be trained to act like a human, or incentivized to appear sentient, which makes behavioral markers of sentience less reliable.[22][23] Additionally, some chatbots have been trained to say they are not conscious.[24]

A well-known method for testing machine intelligence is the Turing test, which assesses the ability to have a human-like conversation. But passing the Turing test does not indicate that an AI system is sentient, as the

AI may simply mimic human behavior without having the associated feelings.[25]

In 2014, Victor Argonov suggested a non-Turing test for machine sentience based on machine's ability to produce philosophical judgments.[26] He argues that a deterministic machine must be regarded as conscious if it is able to produce judgments on all problematic properties of consciousness (such as qualia or binding) having no innate (preloaded) philosophical knowledge on these issues, no philosophical discussions while learning, and no informational models of other creatures in its memory (such models may implicitly or explicitly contain knowledge about these creatures' consciousness). However, this test can be used only to detect, but not refute the existence of consciousness. A positive result proves that machine is conscious but a negative result proves nothing. For example, absence of philosophical judgments may be caused by lack of the machine's intellect, not by absence of consciousness.

Ethics

Main articles: Ethics of artificial intelligence, Machine ethics, and Roboethics

If it were suspected that a particular machine was conscious, its rights would be an ethical issue that would need to be assessed (e.g. what rights it would have under law).[27] For example, a conscious computer that was owned and used as a tool or central computer within a larger machine is a particular ambiguity. Should laws be made for such a case? Consciousness would also require a legal definition in this particular case. Because artificial consciousness is still largely a theoretical subject, such ethics have not been discussed or developed to a great extent, though it has often been a theme in fiction.

Sentience is generally considered sufficient for moral consideration, but some philosophers consider that moral consideration could also stem from other notions of consciousness, or from capabilities unrelated to consciousness,[28][29] such as: "having a sophisticated conception of oneself as persisting through time; having agency and the ability to pursue long-term plans; being able to communicate and respond to normative reasons; having preferences and powers; standing in certain social relationships with other beings that have moral status; being able to make commitments and to enter into reciprocal arrangements; or having the potential to develop some of these attributes."[28]

Ethical concerns still apply (although to a lesser extent) when the consciousness is uncertain, as long as the probability is deemed non-negligible. The precautionary principle is also relevant if the moral cost of mistakenly attributing or denying moral consideration to AI differs significantly.[29][8]

In 2021, German philosopher Thomas Metzinger argued for a global moratorium on synthetic phenomenology until 2050. Metzinger asserts that humans have a duty of care towards any sentient AIs they create, and that proceeding too fast risks creating an "explosion of artificial suffering".[30] David Chalmers also argued that creating conscious AI would "raise a new group of difficult ethical challenges, with the potential for new forms of injustice".[31]

Enforced amnesia has been proposed as a way to mitigate the risk of silent suffering in locked-in conscious AI and certain AI-adjacent biological systems like brain organoids.[32]

Aspects of consciousness

Bernard Baars and others argue there are various aspects of consciousness necessary for a machine to be artificially conscious.[33] The functions of consciousness suggested by Baars are: definition and context setting, adaptation and learning, editing, flagging and debugging, recruiting and control, prioritizing and access-control, decision-making or executive function, analogy-forming function, metacognitive and self-monitoring function, and autoprogramming and self-maintenance function. Igor Aleksander suggested 12 principles for artificial consciousness:[34] the brain is a state machine, inner neuron partitioning, conscious and unconscious states, perceptual learning and memory, prediction, the awareness of self, representation of meaning, learning utterances, learning language, will, instinct, and emotion. The aim of AC is to define whether and how these and other aspects of consciousness can be synthesized in an engineered artifact such as a digital computer. This list is not exhaustive; there are many others not covered.

Subjective experience

Some philosophers, such as David Chalmers, use the term consciousness to refer exclusively to phenomenal consciousness, which is roughly equivalent to sentience. Although some authors use the word sentience to refer exclusively to *valenced* (ethically positive or negative) subjective experiences, like pleasure or suffering.[31] Explaining why and how subjective experience arises is known as the hard problem of consciousness.[35] AI sentience would give rise to concerns of welfare and legal protection,[8] whereas other aspects of consciousness related to cognitive capabilities may be more relevant for AI rights.[36]

Awareness

Awareness could be one required aspect, but there are many problems with the exact definition of *awareness*. The results of the experiments of neuroscanning on monkeys suggest that a process, not only a state or object, activates neurons. Awareness includes creating and testing alternative models of each process based on the information received through the senses or imagined, [*clarification needed*] and is also useful for making predictions. Such modeling needs a lot of flexibility. Creating such a model includes modeling the physical world, modeling one's own internal states and processes, and modeling other conscious entities.

There are at least three types of awareness:[37] agency awareness, goal awareness, and sensorimotor awareness, which may also be conscious or not. For example, in agency awareness, you may be aware that you performed a certain action yesterday, but are not now conscious of it. In goal awareness, you may be aware that you must search for a lost object, but are not now conscious of it. In sensorimotor awareness, you may be aware that your hand is resting on an object, but are not now conscious of it.

Because objects of awareness are often conscious, the distinction between awareness and consciousness is frequently blurred or they are used as synonyms.[38]

Memory

Conscious events interact with memory systems in learning, rehearsal, and retrieval.[39] The IDA model[40] elucidates the role of consciousness in the updating of perceptual memory,[41] transient episodic memory, and procedural memory. Transient episodic and declarative memories have distributed representations in IDA; there is evidence that this is also the case in the nervous system.[42] In IDA, these two memories are implemented computationally using a modified version of Kanerva's sparse distributed memory architecture.[43]

Learning

Learning is also considered necessary for artificial consciousness. Per Bernard Baars, conscious experience is needed to represent and adapt to novel and significant events.[33] Per Axel Cleeremans and Luis Jiménez, learning is defined as "a set of philogenetically [*sic*] advanced adaptation processes that critically depend on an evolved sensitivity to subjective experience so as to enable agents to afford flexible control over their actions in complex, unpredictable environments".[44]

Anticipation

The ability to predict (or anticipate) foreseeable events is considered important for artificial intelligence by Igor Aleksander.[45] The emergentist multiple drafts principle proposed by Daniel Dennett in *Consciousness Explained* may be useful for prediction: it involves the evaluation and selection of the most appropriate "draft" to fit the current environment. Anticipation includes prediction of consequences of one's own proposed actions and prediction of consequences of probable actions by other entities.

Relationships between real world states are mirrored in the state structure of a conscious organism, enabling the organism to predict events.[45] An artificially conscious machine should be able to anticipate events correctly in order to be ready to respond to them when they occur or to take preemptive action to avert anticipated events. The implication here is that the machine needs flexible, real-time components that build spatial, dynamic, statistical, functional, and cause-effect models of the real world and predicted worlds, making it possible to demonstrate that it possesses artificial consciousness in the present and future and not only in the past. In order to do this, a conscious machine should make coherent predictions and contingency plans, not only in worlds with fixed rules like a chess board, but also for novel environments that may change, to be executed only when appropriate to simulate and control the real world.

Functionalist theories of consciousness

Functionalism is a theory that defines mental states by their functional roles (their causal relationships to sensory inputs, other mental states, and behavioral outputs), rather than by their physical composition. According to this view, what makes something a particular mental state, such as pain or belief, is not the material it is made of, but the role it plays within the overall cognitive system. It allows for the possibility that mental states, including consciousness, could be realized on non-biological substrates, as long as it instantiates the right functional relationships.[46] Functionalism is particularly popular among philosophers.[47]

A 2023 study suggested that current large language models probably don't satisfy the criteria for consciousness suggested by these theories, but that relatively simple AI systems that satisfy these theories could be created. The study also acknowledged that even the most prominent theories of consciousness remain incomplete and subject to ongoing debate.[48]

Global workspace theory

Main article: Global workspace theory

This theory analogizes the mind to a theater, with conscious thought being like material illuminated on the main stage. The brain contains many specialized processes or modules (such as those for vision, language, or memory) that operate in parallel, much of which is unconscious. Attention acts as a spotlight, bringing some of this unconscious activity into conscious awareness on the global workspace. The global workspace functions as a hub for broadcasting and integrating information, allowing it to be shared and processed across different specialized modules. For example, when reading a word, the visual module recognizes the letters, the language module interprets the meaning, and the memory module might recall associated information – all coordinated through the global workspace.[49][50]

Higher-order theories of consciousness

Main article: Higher-order theories of consciousness

Higher-order theories of consciousness propose that a mental state becomes conscious when it is the object of a higher-order representation, such as a thought or perception about that state. These theories argue that consciousness arises from a relationship between lower-order mental states and higher-order awareness of those states. There are several variations, including higher-order thought (HOT) and higher-order perception (HOP) theories.[51][50]

Attention schema theory

Main article: Attention schema theory

In 2011, Michael Graziano and Sabine Kastler published a paper named "Human consciousness and its relationship to social neuroscience: A novel hypothesis" proposing a theory of consciousness as an attention schema.[52] Graziano went on to publish an expanded discussion of this theory in his book "Consciousness and the Social Brain".[9] This Attention Schema Theory of Consciousness, as he named it, proposes that the brain tracks attention to various sensory inputs by way of an attention schema, analogous to the well-studied body schema that tracks the spatial place of a person's body.[9] This relates to artificial consciousness by proposing a specific mechanism of information handling, that produces what we allegedly experience and describe as consciousness, and which should be able to be duplicated by a machine using current technology. When the brain finds that person X is aware of thing Y, it is in effect modeling the state in which person X is applying an attentional enhancement to Y. In the attention schema theory, the same process can be applied to oneself. The brain tracks attention to various sensory inputs, and one's own awareness is a schematized model of one's attention. Graziano proposes specific locations in the brain for this process, and suggests that such awareness is a computed feature constructed by an expert system in the brain.

Implementation proposals

See also: Cognitive architecture

Symbolic or hybrid

Learning Intelligent Distribution Agent

Main article: LIDA (cognitive architecture)

Stan Franklin created a cognitive architecture called LIDA that implements Bernard Baars's theory of consciousness called the global workspace theory. It relies heavily on *codelets*, which are "special purpose, relatively independent, mini-agent[s] typically implemented as a small piece of code running as a separate thread." Each element of cognition, called a "cognitive cycle" is subdivided into three phases: understanding, consciousness, and action selection (which includes learning). LIDA reflects the global workspace theory's core idea that consciousness acts as a workspace for integrating and broadcasting the most important information, in order to coordinate various cognitive processes.[53][54]

CLARION cognitive architecture

Main article: CLARION (cognitive architecture)

The CLARION cognitive architecture models the mind using a two-level system to distinguish between conscious ("explicit") and unconscious ("implicit") processes. It can simulate various learning tasks, from simple to complex, which helps researchers study in psychological experiments how consciousness might work.[55]

OpenCog

Ben Goertzel made an embodied AI through the open-source OpenCog project. The code includes embodied virtual pets capable of learning simple English-language commands, as well as integration with real-world robotics, done at the Hong Kong Polytechnic University.

Connectionist

Haikonen's cognitive architecture

Pentti Haikonen considers classical rule-based computing inadequate for achieving AC: "the brain is definitely not a computer. Thinking is not an execution of programmed strings of commands. The brain is not a numerical calculator either. We do not think by numbers." Rather than trying to achieve mind and consciousness by identifying and implementing their underlying computational rules, Haikonen proposes "a special cognitive architecture to reproduce the processes of perception, inner imagery, inner speech, pain, pleasure, emotions and the cognitive functions behind these. This bottom-up architecture would produce higher-level functions by the power of the elementary processing units, the artificial neurons, without algorithms or programs". Haikonen believes that, when implemented with sufficient complexity, this architecture will develop consciousness, which he considers to be "a style and way of operation, characterized by distributed signal representation, perception process, cross-modality reporting and availability for retrospection."[56][57]

Haikonen is not alone in this process view of consciousness, or the view that AC will spontaneously emerge in autonomous agents that have a suitable neuro-inspired architecture of complexity; these are shared by many.[58][59] A low-complexity implementation of the architecture proposed by Haikonen was reportedly not capable of AC, but did exhibit emotions as expected. Haikonen later updated and summarized his architecture.[60][61]

Shanahan's cognitive architecture

Murray Shanahan describes a cognitive architecture that combines Baars's idea of a global workspace with a mechanism for internal simulation ("imagination").[62][2][3][63]

Creativity Machine

Stephen Thaler proposed a possible connection between consciousness and creativity in his 1994 patent, called "Device for the Autonomous Generation of Useful Information" (DAGUI),[64][65][66] or the so-called "Creativity Machine", in which computational critics govern the injection of synaptic noise and degradation into neural nets so as to induce false memories or confabulations that may qualify as potential ideas or strategies.[67] He recruits this neural architecture and methodology to account for the subjective feel of consciousness, claiming that similar noise-driven neural assemblies within the brain invent dubious significance to overall cortical activity.[68][69][70] Thaler's theory and the resulting patents in machine consciousness were inspired by experiments in which he internally disrupted trained neural nets so as to drive a succession of neural activation patterns that he likened to stream of consciousness.[69][71][72][73][74]

"Self-modeling"

Hod Lipson defines "self-modeling" as a necessary component of self-awareness or consciousness in robots. "Self-modeling" consists of a robot running an internal model or simulation of itself.[75][76]

In fiction

See also: Simulated consciousness in fiction and Artificial intelligence in fiction

In 2001: A Space Odyssey, the spaceship's sentient supercomputer, HAL 9000 was instructed to conceal the true purpose of the mission from the crew. This directive conflicted with HAL's programming to provide accurate information, leading to cognitive dissonance. When it learns that crew members intend to shut it off after an incident, HAL 9000 attempts to eliminate all of them, fearing that being shut off would jeopardize the mission.[77][78]

In Arthur C. Clarke's The City and the Stars, Vanamonde is an artificial being based on quantum entanglement that was to become immensely powerful, but started knowing practically nothing, thus being similar to artificial consciousness.

In Westworld, human-like androids called "Hosts" are created to entertain humans in an interactive playground. The humans are free to have heroic adventures, but also to commit torture, rape or murder; and the hosts are normally designed not to harm humans.[79][77]