Philosophy of artificial intelligence

See also: Ethics of artificial intelligence

The **philosophy of artificial intelligence** is a branch of the philosophy of mind and the philosophy of computer science[1] that explores artificial intelligence and its implications for knowledge and understanding of intelligence, ethics, consciousness, epistemology, and free will.[2][3] Furthermore, the technology is concerned with the creation of artificial animals or artificial people (or, at least, artificial creatures; see artificial life) so the discipline is of considerable interest to philosophers.[4] These factors contributed to the emergence of the philosophy of artificial intelligence.

The philosophy of artificial intelligence attempts to answer such questions as follows:[5]

- Can a machine act intelligently? Can it solve *any* problem that a person would solve by thinking?
- Are human intelligence and machine intelligence the same? Is the human brain essentially a computer?
- Can a machine have a mind, mental states, and consciousness in the same sense that a human being can? Can it *feel how things are*? (i.e. does it have qualia?)

Questions like these reflect the divergent interests of AI researchers, cognitive scientists and philosophers respectively. The scientific answers to these questions depend on the definition of "intelligence" and "consciousness" and exactly which "machines" are under discussion.

Important propositions in the philosophy of AI include some of the following:

- Turing's "polite convention": If a machine behaves as intelligently as a human being, then it is as intelligent as a human being.[6]
- The Dartmouth proposal: "Every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."[7]
- Allen Newell and Herbert A. Simon's physical symbol system hypothesis: "A physical symbol system has the necessary and sufficient means of general intelligent action."[8]
- John Searle's strong AI hypothesis: "The appropriately programmed computer with the right inputs and outputs would thereby have a mind in exactly the same sense human beings have minds."[9]
- Hobbes' mechanism: "For 'reason' ... is nothing but 'reckoning,' that is adding and subtracting, of the consequences of general names agreed upon for the

'marking' and 'signifying' of our thoughts..."[10]

Can a machine display general intelligence?

Is it possible to create a machine that can solve *all* the problems humans solve using their intelligence? This question defines the scope of what machines could do in the future and guides the direction of AI research. It only concerns the *behavior* of machines and ignores the issues of interest to psychologists, cognitive scientists and philosophers, evoking the question: does it matter whether a machine is *really* thinking, as a person thinks, rather than just producing outcomes that appear to result from thinking?[11]

The basic position of most AI researchers is summed up in this statement, which appeared in the proposal for the Dartmouth workshop of 1956:

- "Every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."[7]

Arguments against the basic premise must show that building a working AI system is impossible because there is some practical limit to the abilities of computers or that there is some special quality of the human mind that is necessary for intelligent behavior and yet cannot be duplicated by a machine (or by the methods of current AI

research). Arguments in favor of the basic premise must show that such a system is possible.

It is also possible to sidestep the connection between the two parts of the above proposal. For instance, machine learning, beginning with Turing's infamous child machine proposal,[12] essentially achieves the desired feature of intelligence without a precise design-time description as to how it would exactly work. The account on robot tacit knowledge[13] eliminates the need for a precise description altogether.

The first step to answering the question is to clearly define "intelligence".

Intelligence

Turing test

Main article: Turing test

Alan Turing[15] reduced the problem of defining intelligence to a simple question about conversation. He suggests that: if a machine can answer *any* question posed to it, using the same words that an ordinary person would, then we may call that machine intelligent. A modern version of his experimental design would use an online chat room, where one of the participants is a real person and one of the participants is a computer program. The program passes the test if no one can tell which of the two participants is human.[6] Turing notes that no one

(except philosophers) ever asks the question "can people think?" He writes "instead of arguing continually over this point, it is usual to have a polite convention that everyone thinks".[16] Turing's test extends this polite convention to machines:

- If a machine acts as intelligently as a human being, then it is as intelligent as a human being.

One criticism of the Turing test is that it only measures the "humanness" of the machine's behavior, rather than the "intelligence" of the behavior. Since human behavior and intelligent behavior are not exactly the same thing, the test fails to measure intelligence. Stuart J. Russell and Peter Norvig write that "aeronautical engineering texts do not define the goal of their field as 'making machines that fly so exactly like pigeons that they can fool other pigeons'".[17]

Intelligence as achieving goals

Twenty-first century AI research defines intelligence in terms of goal-directed behavior. It views intelligence as a set of problems that the machine is expected to solve – the more problems it can solve, and the better its solutions are, the more intelligent the program is. AI founder John McCarthy defined intelligence as "the computational part of the ability to achieve goals in the world."[18]

Stuart Russell and Peter Norvig formalized this definition

using abstract intelligent agents. An "agent" is something which perceives and acts in an environment. A "performance measure" defines what counts as success for the agent.[19]

- "If an agent acts so as to maximize the expected value of a performance measure based on past experience and knowledge then it is intelligent."[20]

Definitions like this one try to capture the essence of intelligence. They have the advantage that, unlike the Turing test, they do not also test for unintelligent human traits such as making typing mistakes.[21] They have the disadvantage that they can fail to differentiate between "things that think" and "things that do not". By this definition, even a thermostat has a rudimentary intelligence.[22]

Arguments that a machine can display general intelligence

The brain can be simulated

Main article: Artificial brain

Hubert Dreyfus describes this argument as claiming that "if the nervous system obeys the laws of physics and chemistry, which we have every reason to suppose it does, then ... we ... ought to be able to reproduce the behavior of the nervous system with some physical device".[23] This argument, first introduced as early as

1943[24] and vividly described by Hans Moravec in 1988,[25] is now associated with futurist Ray Kurzweil, who estimates that computer power will be sufficient for a complete brain simulation by the year 2029.[26] A non-real-time simulation of a thalamocortical model that has the size of the human brain (10^{11} neurons) was performed in 2005,[27] and it took 50 days to simulate 1 second of brain dynamics on a cluster of 27 processors.

Even AI's harshest critics (such as Hubert Dreyfus and John Searle) agree that a brain simulation is possible in theory.[a] However, Searle points out that, in principle, *anything* can be simulated by a computer; thus, bringing the definition to its breaking point leads to the conclusion that any process at all can technically be considered "computation". "What we wanted to know is what distinguishes the mind from thermostats and livers," he writes.[30] Thus, merely simulating the functioning of a living brain would in itself be an admission of ignorance regarding intelligence and the nature of the mind, like trying to build a jet airliner by copying a living bird precisely, feather by feather, with no theoretical understanding of aeronautical engineering.[31]

Human thinking is symbol processing

Main article: Physical symbol system

In 1963, Allen Newell and Herbert A. Simon proposed that "symbol manipulation" was the essence of both human

and machine intelligence. They wrote:

- "A physical symbol system has the necessary and sufficient means of general intelligent action."[8]

This claim is very strong: it implies both that human thinking is a kind of symbol manipulation (because a symbol system is *necessary* for intelligence) and that machines can be intelligent (because a symbol system is *sufficient* for intelligence).[32] Another version of this position was described by philosopher Hubert Dreyfus, who called it "the psychological assumption":

- "The mind can be viewed as a device operating on bits of information according to formal rules."[33]

The "symbols" that Newell, Simon and Dreyfus discussed were word-like and high level—symbols that directly correspond with objects in the world, such as <dog> and <tail>. Most AI programs written between 1956 and 1990 used this kind of symbol. Modern AI, based on statistics and mathematical optimization, does not use the high-level "symbol processing" that Newell and Simon discussed.

Arguments against symbol processing

These arguments show that human thinking does not consist (solely) of high level symbol manipulation. They do *not* show that artificial intelligence is impossible, only that more than symbol processing is required.

Gödelian anti-mechanist arguments

Main article: Mechanism (philosophy): Gödelian arguments

In 1931, Kurt Gödel proved with an incompleteness theorem that it is always possible to construct a "Gödel statement" that a given consistent formal system of logic (such as a high-level symbol manipulation program) could not prove. Despite being a true statement, the constructed Gödel statement is unprovable in the given system. (The truth of the constructed Gödel statement is contingent on the consistency of the given system; applying the same process to a subtly inconsistent system will appear to succeed, but will actually yield a false "Gödel statement" instead.)[citation needed] More speculatively, Gödel conjectured that the human mind can eventually correctly determine the truth or falsity of any well-grounded mathematical statement (including any possible Gödel statement), and that therefore the human mind's power is not reducible to a *mechanism*.[34] Philosopher John Lucas (since 1961) and Roger Penrose (since 1989) have championed this philosophical anti-mechanist argument.[35]

Gödelian anti-mechanist arguments tend to rely on the innocuous-seeming claim that a system of human mathematicians (or some idealization of human mathematicians) is both consistent (completely free of error) and believes fully in its own consistency (and can

make all logical inferences that follow from its own consistency, including belief in its Gödel statement) [citation needed]. This is probably impossible for a Turing machine to do (see Halting problem); therefore, the Gödelian concludes that human reasoning is too powerful to be captured by a Turing machine, and by extension, any digital mechanical device.

However, the modern consensus in the scientific and mathematical community is that actual human reasoning is inconsistent; that any consistent "idealized version" *H* of human reasoning would logically be forced to adopt a healthy but counter-intuitive open-minded skepticism about the consistency of *H* (otherwise *H* is provably inconsistent); and that Gödel's theorems do not lead to any valid argument that humans have mathematical reasoning capabilities beyond what a machine could ever duplicate.[36][37][38] This consensus that Gödelian anti-mechanist arguments are doomed to failure is laid out strongly in *Artificial Intelligence*: "*any* attempt to utilize (Gödel's incompleteness results) to attack the computationalist thesis is bound to be illegitimate, since these results are quite consistent with the computationalist thesis."[39]

Stuart Russell and Peter Norvig agree that Gödel's argument does not consider the nature of real-world human reasoning. It applies to what can theoretically be proved, given an infinite amount of memory and time. In practice, real machines (including humans) have finite

resources and will have difficulty proving many theorems. It is not necessary to be able to prove everything in order to be an intelligent person.[40]

Less formally, Douglas Hofstadter, in his Pulitzer prize winning book *Gödel, Escher, Bach: An Eternal Golden Braid*, states that these "Gödel-statements" always refer to the system itself, drawing an analogy to the way the Epimenides paradox uses statements that refer to themselves, such as "this statement is false" or "I am lying".[41] But, of course, the Epimenides paradox applies to anything that makes statements, whether it is a machine *or* a human, even Lucas himself. Consider:

- Lucas can't assert the truth of this statement.[42]

This statement is true but cannot be asserted by Lucas. This shows that Lucas himself is subject to the same limits that he describes for machines, as are all people, and so Lucas's argument is pointless.[43]

After concluding that human reasoning is non-computable, Penrose went on to controversially speculate that some kind of hypothetical non-computable processes involving the collapse of quantum mechanical states give humans a special advantage over existing computers. Existing quantum computers are only capable of reducing the complexity of Turing computable tasks and are still restricted to tasks within the scope of Turing machines. [*citation needed*] [*clarification needed*]. By Penrose and

Lucas's arguments, the fact that quantum computers are only able to complete Turing computable tasks implies that they cannot be sufficient for emulating the human mind.[citation needed] Therefore, Penrose seeks for some other process involving new physics, for instance quantum gravity which might manifest new physics at the scale of the Planck mass via spontaneous quantum collapse of the wave function. These states, he suggested, occur both within neurons and also spanning more than one neuron.[44] However, other scientists point out that there is no plausible organic mechanism in the brain for harnessing any sort of quantum computation, and furthermore that the timescale of quantum decoherence seems too fast to influence neuron firing.[45]

Dreyfus: the primacy of implicit skills

Main article: Hubert Dreyfus's views on artificial intelligence

Hubert Dreyfus argued that human intelligence and expertise depended primarily on fast intuitive judgements rather than step-by-step symbolic manipulation, and argued that these skills would never be captured in formal rules.[46]

Dreyfus's argument had been anticipated by Turing in his 1950 paper Computing machinery and intelligence, where he had classified this as the "argument from the informality of behavior."[47] Turing argued in response

that, just because we do not know the rules that govern a complex behavior, this does not mean that no such rules exist. He wrote: "we cannot so easily convince ourselves of the absence of complete laws of behaviour ... The only way we know of for finding such laws is scientific observation, and we certainly know of no circumstances under which we could say, 'We have searched enough. There are no such laws.'"[48]

Russell and Norvig point out that, in the years since Dreyfus published his critique, progress has been made towards discovering the "rules" that govern unconscious reasoning.[49] The situated movement in robotics research attempts to capture our unconscious skills at perception and attention.[50] Computational intelligence paradigms, such as neural nets, evolutionary algorithms and so on are mostly directed at simulated unconscious reasoning and learning. Statistical approaches to AI can make predictions which approach the accuracy of human intuitive guesses. Research into commonsense knowledge has focused on reproducing the "background" or context of knowledge. In fact, AI research in general has moved away from high level symbol manipulation, towards new models that are intended to capture more of our *intuitive* reasoning.[49]

Cognitive science and psychology eventually came to agree with Dreyfus' description of human expertise. Daniel Kahnemann and others developed a similar theory where they identified two "systems" that humans use to

solve problems, which he called "System 1" (fast intuitive judgements) and "System 2" (slow deliberate step by step thinking).[51]

Although Dreyfus' views have been vindicated in many ways, the work in cognitive science and in AI was in response to specific problems in those fields and was not directly influenced by Dreyfus. Historian and AI researcher Daniel Crevier wrote that "time has proven the accuracy and perceptiveness of some of Dreyfus's comments. Had he formulated them less aggressively, constructive actions they suggested might have been taken much earlier."[52]

Can a machine have a mind, consciousness, and mental states?

This is a philosophical question, related to the problem of other minds and the hard problem of consciousness. The question revolves around a position defined by John Searle as "strong AI":

- A physical symbol system can have a mind and mental states.[9]

Searle distinguished this position from what he called "weak AI":

- A physical symbol system can act intelligently.[9]

Searle introduced the terms to isolate strong AI from weak

AI so he could focus on what he thought was the more interesting and debatable issue. He argued that *even if we assume* that we had a computer program that acted exactly like a human mind, there would still be a difficult philosophical question that needed to be answered.[9]

Neither of Searle's two positions are of great concern to AI research, since they do not directly answer the question "can a machine display general intelligence?" (unless it can also be shown that consciousness is *necessary* for intelligence). Turing wrote "I do not wish to give the impression that I think there is no mystery about consciousness… [b]ut I do not think these mysteries necessarily need to be solved before we can answer the question [of whether machines can think]."[53] Russell and Norvig agree: "Most AI researchers take the weak AI hypothesis for granted, and don't care about the strong AI hypothesis."[54]

There are a few researchers who believe that consciousness is an essential element in intelligence, such as Igor Aleksander, Stan Franklin, Ron Sun, and Pentti Haikonen, although their definition of "consciousness" strays very close to "intelligence". (See artificial consciousness.)

Before we can answer this question, we must be clear what we mean by "minds", "mental states" and "consciousness".

Consciousness, minds, mental states, meaning

The words "mind" and "consciousness" are used by different communities in different ways. Some new age thinkers, for example, use the word "consciousness" to describe something similar to Bergson's "élan vital": an invisible, energetic fluid that permeates life and especially the mind. Science fiction writers use the word to describe some essential property that makes us human: a machine or alien that is "conscious" will be presented as a fully human character, with intelligence, desires, will, insight, pride and so on. (Science fiction writers also use the words "sentience", "sapience", "self-awareness" or "ghost"—as in the *Ghost in the Shell* manga and anime series—to describe this essential human property). For others [*who?*], the words "mind" or "consciousness" are used as a kind of secular synonym for the soul.

For philosophers, neuroscientists and cognitive scientists, the words are used in a way that is both more precise and more mundane: they refer to the familiar, everyday experience of having a "thought in your head", like a perception, a dream, an intention or a plan, and to the way we *see* something, *know* something, *mean* something or *understand* something.[55] "It's not hard to give a commonsense definition of consciousness" observes philosopher John Searle.[56] What is mysterious and fascinating is not so much *what* it is but *how* it is: how

does a lump of fatty tissue and electricity give rise to this (familiar) experience of perceiving, meaning or thinking?

Philosophers call this the hard problem of consciousness. It is the latest version of a classic problem in the philosophy of mind called the "mind-body problem".[57] A related problem is the problem of *meaning* or *understanding* (which philosophers call "intentionality"): what is the connection between our *thoughts* and *what we are thinking about* (i.e. objects and situations out in the world)? A third issue is the problem of *experience* (or "phenomenology"): If two people see the same thing, do they have the same experience? Or are there things "inside their head" (called "qualia") that can be different from person to person?[58]

Neurobiologists believe all these problems will be solved as we begin to identify the neural correlates of consciousness: the actual relationship between the machinery in our heads and its collective properties; such as the mind, experience and understanding. Some of the harshest critics of artificial intelligence agree that the brain is just a machine, and that consciousness and intelligence are the result of physical processes in the brain.[59] The difficult philosophical question is this: can a computer program, running on a digital machine that shuffles the binary digits of zero and one, duplicate the ability of the neurons to create minds, with mental states (like understanding or perceiving), and ultimately, the experience of consciousness?

Arguments that a computer cannot have a mind and mental states

Searle's Chinese room

Main article: Chinese room

John Searle asks us to consider a thought experiment: suppose we have written a computer program that passes the Turing test and demonstrates general intelligent action. Suppose, specifically that the program can converse in fluent Chinese. Write the program on 3x5 cards and give them to an ordinary person who does not speak Chinese. Lock the person into a room and have him follow the instructions on the cards. He will copy out Chinese characters and pass them in and out of the room through a slot. From the outside, it will appear that the Chinese room contains a fully intelligent person who speaks Chinese. The question is this: is there anyone (or anything) in the room that understands Chinese? That is, is there anything that has the mental state of understanding, or which has conscious awareness of what is being discussed in Chinese? The man is clearly not aware. The room cannot be aware. The *cards* certainly are not aware. Searle concludes that the Chinese room, or *any* other physical symbol system, cannot have a mind.[60]

Searle goes on to argue that actual mental states and consciousness require (yet to be described) "actual physical-chemical properties of actual human brains."[61]

He argues there are special "causal properties" of brains and neurons that gives rise to minds: in his words "brains cause minds."[62]

Related arguments: Leibniz' mill, Davis's telephone exchange, Block's Chinese nation and Blockhead

Gottfried Leibniz made essentially the same argument as Searle in 1714, using the thought experiment of expanding the brain until it was the size of a mill.[63] In 1974, Lawrence Davis imagined duplicating the brain using telephone lines and offices staffed by people, and in 1978 Ned Block envisioned the entire population of China involved in such a brain simulation. This thought experiment is called "the Chinese Nation" or "the Chinese Gym".[64] Ned Block also proposed his Blockhead argument, which is a version of the Chinese room in which the program has been re-factored into a simple set of rules of the form "see this, do that", removing all mystery from the program.

Responses to the Chinese room

Responses to the Chinese room emphasize several different points.

- **The systems reply** and the **virtual mind reply**:[65] This reply argues that *the system*, including the man, the program, the room, and the cards, is what understands Chinese. Searle claims that the man in the room is the only thing which could possibly "have

a mind" or "understand", but others disagree, arguing that it is possible for there to be *two* minds in the same physical place, similar to the way a computer can simultaneously "be" two machines at once: one physical (like a Macintosh) and one "virtual" (like a word processor).

- **Speed, power and complexity replies**:[66] Several critics point out that the man in the room would probably take millions of years to respond to a simple question, and would require "filing cabinets" of astronomical proportions. This brings the clarity of Searle's intuition into doubt.
- **Robot reply**:[67] To truly understand, some believe the Chinese Room needs eyes and hands. Hans Moravec writes: "If we could graft a robot to a reasoning program, we wouldn't need a person to provide the meaning anymore: it would come from the physical world."[68]
- **Brain simulator reply**:[69] What if the program simulates the sequence of nerve firings at the synapses of an actual brain of an actual Chinese speaker? The man in the room would be simulating an actual brain. This is a variation on the "systems reply" that appears more plausible because "the system" now clearly operates like a human brain, which strengthens the intuition that there is something besides the man in the room that could understand Chinese.
- **Other minds reply** and the **epiphenomena reply**:

[70] Several people have noted that Searle's argument is just a version of the problem of other minds, applied to machines. Since it is difficult to decide if people are "actually" thinking, we should not be surprised that it is difficult to answer the same question about machines.

A related question is whether "consciousness" (as Searle understands it) exists. Searle argues that the experience of consciousness cannot be detected by examining the behavior of a machine, a human being or any other animal. Daniel Dennett points out that natural selection cannot preserve a feature of an animal that has no effect on the behavior of the animal, and thus consciousness (as Searle understands it) cannot be produced by natural selection. Therefore, either natural selection did not produce consciousness, or "strong AI" is correct in that consciousness can be detected by suitably designed Turing test.

Is thinking a kind of computation?

Main article: Computational theory of mind

The computational theory of mind or "computationalism" claims that the relationship between mind and brain is similar (if not identical) to the relationship between a *running program* (software) and a computer (hardware). The idea has philosophical roots in Hobbes (who claimed

reasoning was "nothing more than reckoning"), Leibniz (who attempted to create a logical calculus of all human ideas), Hume (who thought perception could be reduced to "atomic impressions") and even Kant (who analyzed all experience as controlled by formal rules).[71] The latest version is associated with philosophers Hilary Putnam and Jerry Fodor.[72]

This question bears on our earlier questions: if the human brain is a kind of computer then computers can be both intelligent and conscious, answering both the practical and philosophical questions of AI. In terms of the practical question of AI ("Can a machine display general intelligence?"), some versions of computationalism make the claim that (as Hobbes wrote):

- Reasoning is nothing but reckoning.[10]

In other words, our intelligence derives from a form of *calculation*, similar to arithmetic. This is the physical symbol system hypothesis discussed above, and it implies that artificial intelligence is possible. In terms of the philosophical question of AI ("Can a machine have mind, mental states and consciousness?"), most versions of computationalism claim that (as Stevan Harnad characterizes it):

- Mental states are just implementations of (the right) computer programs.[73]

This is John Searle's "strong AI" discussed above, and it

is the real target of the Chinese room argument (according to Harnad).[73]

Other related questions

Can a machine have emotions?

If "emotions" are defined only in terms of their effect on behavior or on how they function inside an organism, then emotions can be viewed as a mechanism that an intelligent agent uses to maximize the utility of its actions. Given this definition of emotion, Hans Moravec believes that "robots in general will be quite emotional about being nice people".[74] Fear is a source of urgency. Empathy is a necessary component of good human computer interaction. He says robots "will try to please you in an apparently selfless manner because it will get a thrill out of this positive reinforcement. You can interpret this as a kind of love."[74] Daniel Crevier writes "Moravec's point is that emotions are just devices for channeling behavior in a direction beneficial to the survival of one's species."[75]

Can a machine be self-aware?

"Self-awareness", as noted above, is sometimes used by science fiction writers as a name for the essential human property that makes a character fully human. Turing strips away all other properties of human beings and reduces the question to "can a machine be the subject of its own thought?" Can it *think about itself*? Viewed in this way, a

program can be written that can report on its own internal states, such as a debugger.[76]

Can a machine be original or creative?

Turing reduces this to the question of whether a machine can "take us by surprise" and argues that this is obviously true, as any programmer can attest.[77] He notes that, with enough storage capacity, a computer can behave in an astronomical number of different ways.[78] It must be possible, even trivial, for a computer that can represent ideas to combine them in new ways. (Douglas Lenat's Automated Mathematician, as one example, combined ideas to discover new mathematical truths.) Kaplan and Haenlein suggest that machines can display scientific creativity, while it seems likely that humans will have the upper hand where artistic creativity is concerned.[79]

In 2009, scientists at Aberystwyth University in Wales and the U.K's University of Cambridge designed a robot called Adam that they believe to be the first machine to independently come up with new scientific findings.[80] Also in 2009, researchers at Cornell developed Eureqa, a computer program that extrapolates formulas to fit the data inputted, such as finding the laws of motion from a pendulum's motion.

Can a machine be benevolent or hostile?

Main article: Ethics of artificial intelligence

This question (like many others in the philosophy of artificial intelligence) can be presented in two forms. "Hostility" can be defined in terms function or behavior, in which case "hostile" becomes synonymous with "dangerous". Or it can be defined in terms of intent: can a machine "deliberately" set out to do harm? The latter is the question "can a machine have conscious states?" (such as intentions) in another form.[53]

The question of whether highly intelligent and completely autonomous machines would be dangerous has been examined in detail by futurists (such as the Machine Intelligence Research Institute). The obvious element of drama has also made the subject popular in science fiction, which has considered many differently possible scenarios where intelligent machines pose a threat to mankind; see Artificial intelligence in fiction.

One issue is that machines may acquire the autonomy and intelligence required to be dangerous very quickly. Vernor Vinge has suggested that over just a few years, computers will suddenly become thousands or millions of times more intelligent than humans. He calls this "the Singularity".[81] He suggests that it may be somewhat or possibly very dangerous for humans.[82] This is discussed by a philosophy called Singularitarianism.

In 2009, academics and technical experts attended a conference to discuss the potential impact of robots and computers and the impact of the hypothetical possibility

that they could become self-sufficient and able to make their own decisions. They discussed the possibility and the extent to which computers and robots might be able to acquire any level of autonomy, and to what degree they could use such abilities to possibly pose any threat or hazard. They noted that some machines have acquired various forms of semi-autonomy, including being able to find power sources on their own and being able to independently choose targets to attack with weapons. They also noted that some computer viruses can evade elimination and have achieved "cockroach intelligence". They noted that self-awareness as depicted in science-fiction is probably unlikely, but that there were other potential hazards and pitfalls.[81]

Some experts and academics have questioned the use of robots for military combat, especially when such robots are given some degree of autonomous functions.[83] The US Navy has funded a report which indicates that as military robots become more complex, there should be greater attention to implications of their ability to make autonomous decisions.[84][85]

The President of the Association for the Advancement of Artificial Intelligence has commissioned a study to look at this issue.[86] They point to programs like the Language Acquisition Device which can emulate human interaction.

Some have suggested a need to build "Friendly AI", a term coined by Eliezer Yudkowsky, meaning that the advances

which are already occurring with AI should also include an effort to make AI intrinsically friendly and humane.[87]

Can a machine imitate all human characteristics?

Turing said "It is customary ... to offer a grain of comfort, in the form of a statement that some peculiarly human characteristic could never be imitated by a machine. ... I cannot offer any such comfort, for I believe that no such bounds can be set."[88]

Turing noted that there are many arguments of the form "a machine will never do X", where X can be many things, such as:

> Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humor, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make someone fall in love with it, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behaviour as a man, do something really new.[76]

Turing argues that these objections are often based on naive assumptions about the versatility of machines or are "disguised forms of the argument from consciousness". Writing a program that exhibits one of these behaviors "will not make much of an impression."[76] All of these arguments are tangential to the basic premise of AI,

unless it can be shown that one of these traits is essential for general intelligence.

Can a machine have a soul?

Finally, those who believe in the existence of a soul may argue that "Thinking is a function of man's immortal soul." Alan Turing called this "the theological objection". He writes:

> In attempting to construct such machines we should not be irreverently usurping His power of creating souls, any more than we are in the procreation of children: rather we are, in either case, instruments of His will providing mansions for the souls that He creates.[89]

The discussion on the topic has been reignited as a result of recent claims made by Google's LaMDA artificial intelligence system that it is sentient and had a "soul".[90]

LaMDA (Language Model for Dialogue Applications) is an artificial intelligence system that creates chatbots—AI robots designed to communicate with humans—by gathering vast amounts of text from the internet and using algorithms to respond to queries in the most fluid and natural way possible.

The transcripts of conversations between scientists and LaMDA reveal that the AI system excels at this, providing answers to challenging topics about the nature of

emotions, generating Aesop-style fables on the moment, and even describing its alleged fears.[91] Pretty much all philosophers doubt LaMDA's sentience.[92]

Views on the role of philosophy

Some scholars argue that the AI community's dismissal of philosophy is detrimental. In the *Stanford Encyclopedia of Philosophy*, some philosophers argue that the role of philosophy in AI is underappreciated.[4] Physicist David Deutsch argues that without an understanding of philosophy or its concepts, AI development would suffer from a lack of progress.[93]

Conferences and literature

The main conference series on the issue is "Philosophy and Theory of AI" (PT-AI), run by Vincent C. Müller.

The main bibliography on the subject, with several sub-sections, is on PhilPapers.

A recent survey for *Philosophy of AI* is Müller (2023).[3]

See also

- AI takeover
- Artificial brain
- Artificial consciousness
- Artificial intelligence
- Artificial neural network

- Chatbot
- Computational theory of mind
- Computing Machinery and Intelligence
- Existential risk from artificial general intelligence
- Functionalism
- Hubert Dreyfus's views on artificial intelligence
- Multi-agent system
- Philosophy of computer science
- Philosophy of information
- Philosophy of mind
- Physical symbol system
- Simulated reality
- *Superintelligence: Paths, Dangers, Strategies*
- Synthetic intelligence
- Wireheading

Notes

1. Hubert Dreyfus writes: "In general, by accepting the fundamental assumptions that the nervous system is part of the physical world and that all physical processes can be described in a mathematical formalism which can, in turn, be manipulated by a digital computer, one can arrive at the strong claim that the behavior which results from human 'information processing,' whether directly formalizable or not, can always be indirectly reproduced on a digital machine." .[28] John Searle writes: "Could a man made machine think? Assuming

it possible produce artificially a machine with a nervous system, ... the answer to the question seems to be obviously, yes ... Could a digital computer think? If by 'digital computer' you mean anything at all that has a level of description where it can be correctly described as the instantiation of a computer program, then again the answer is, of course, yes, since we are the instantiations of any number of computer programs, and we can think."[29]

References

1. *"Philosophy of Computer Science". obo.*
2. *McCarthy, John. "The Philosophy of AI and the AI of Philosophy". jmc.stanford.edu. Archived from the original on 2018-10-23. Retrieved 2018-09-18.*
3. ^ **a b** *Müller, Vincent C. (2023-07-24). "Philosophy of AI: A structured overview". Nathalie A. Smuha (Ed.), Cambridge Handbook on the Law, Ethics and Policy of Artificial Intelligence.*
4. ^ **a b** *Bringsjord, Selmer; Govindarajulu, Naveen Sundar (2018), "Artificial Intelligence", in Zalta, Edward N. (ed.), The Stanford Encyclopedia of Philosophy (Fall 2018 ed.), Metaphysics Research Lab, Stanford University, archived from the original on 2019-11-09, retrieved 2018-09-18*
5. Russell & Norvig 2003, p. 947 define the philosophy of AI as consisting of the first two questions, and the additional question of the ethics of artificial

intelligence. Fearn 2007, p. 55 writes "In the current literature, philosophy has two chief roles: to determine whether or not such machines would be conscious, and, second, to predict whether or not such machines are possible." The last question bears on the first two.

6. ^ *a b* This is a paraphrase of the essential point of the Turing test. Turing 1950, Haugeland 1985, pp. 6–9, Crevier 1993, p. 24, Russell & Norvig 2003, pp. 2–3, 948.
7. ^ *a b* McCarthy et al. 1955. This assertion was printed in the program for the Dartmouth Conference of 1956, widely considered the "birth of AI." also Crevier 1993, p. 28.
8. ^ *a b* Newell & Simon 1976 and Russell & Norvig 2003, p. 18.
9. ^ *a b c d* This version is from Searle (1999), and is also quoted in Dennett 1991, p. 435. Searle's original formulation was "The appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states." (Searle 1980, p. 1). Strong AI is defined similarly by Russell & Norvig (2003, p. 947): "The assertion that machines could possibly act intelligently (or, perhaps better, act as if they were intelligent) is called the 'weak AI' hypothesis by philosophers, and the assertion that machines that do so are actually thinking (as opposed to simulating thinking) is called

the 'strong AI' hypothesis."
10. ^ *a b* Hobbes 1651, chpt. 5
11. See Russell & Norvig 2003, p. 3, where they make the distinction between *acting* rationally and *being* rational, and define AI as the study of the former.
12. *Turing, Alan M. (1950). "Computing Machinery and Intelligence". Mind. **49** (236): 433–460. doi:10.1093/mind/LIX.236.433. Archived from the original on 2021-12-22. Retrieved 2020-10-18 – via cogprints.*
13. *Heder, Mihaly; Paksi, Daniel (2012). "Autonomous Robots and Tacit Knowledge". Appraisal. **9** (2): 8–14 – via academia.edu.*
14. Saygin 2000.
15. Turing 1950 and see Russell & Norvig 2003, p. 948, where they call his paper "famous" and write "Turing examined a wide variety of possible objections to the possibility of intelligent machines, including virtually all of those that have been raised in the half century since his paper appeared."
16. Turing 1950, under "The Argument from Consciousness".
17. Russell & Norvig 2003, p. 3.
18. McCarthy 1999.
19. Russell & Norvig 2003, pp. 4–5, 32, 35, 36 and 56.
20. Russell and Norvig would prefer the word "rational" to "intelligent".
21. *"Artificial Stupidity". The Economist. Vol. 324, no. 7770. 1 August 1992. p. 14.*

22. Russell & Norvig (2003, pp. 48–52) consider a thermostat a simple form of intelligent agent, known as a reflex agent. For an in-depth treatment of the role of the thermostat in philosophy see Chalmers (1996, pp. 293–301) "4. Is Experience Ubiquitous?" subsections *What is it like to be a thermostat?*, *Whither panpsychism?*, and *Constraining the double-aspect principle*.
23. Dreyfus 1972, p. 106.
24. Pitts & McCullough 1943.
25. Moravec 1988.
26. Kurzweil 2005, p. 262. Also see Russell & Norvig, p. 957 and Crevier 1993, pp. 271 and 279. The most extreme form of this argument (the brain replacement scenario) was put forward by Clark Glymour in the mid-1970s and was touched on by Zenon Pylyshyn and John Searle in 1980.
27. *Eugene Izhikevich (2005-10-27). "Eugene M. Izhikevich, Large-Scale Simulation of the Human Brain". Vesicle.nsi.edu. Archived from the original on 2009-05-01. Retrieved 2010-07-29.*
28. Dreyfus 1972, pp. 194–5.
29. Searle 1980, p. 11.
30. Searle 1980, p. 7.
31. Yudkowsky 2008.
32. Searle writes "I like the straight forwardness of the claim." Searle 1980, p. 4.
33. Dreyfus 1979, p. 156.
34. Gödel, Kurt, 1951, *Some basic theorems on the*

foundations of mathematics and their implications in Solomon Feferman, ed., 1995. *Collected works / Kurt Gödel, Vol. III*. Oxford University Press: 304-23. - In this lecture, Gödel uses the incompleteness theorem to arrive at the following disjunction: (a) the human mind is not a consistent finite machine, or (b) there exist Diophantine equations for which it cannot decide whether solutions exist. Gödel finds (b) implausible, and thus seems to have believed the human mind was not equivalent to a finite machine, i.e., its power exceeded that of any finite machine. He recognized that this was only a conjecture, since one could never disprove (b). Yet he considered the disjunctive conclusion to be a "certain fact".

35. Lucas 1961, Russell & Norvig 2003, pp. 949–950, Hofstadter 1979, pp. 471–473, 476–477.
36. *Oppy, Graham (20 January 2015). "Gödel's Incompleteness Theorems". Stanford Encyclopedia of Philosophy. Archived from the original on 3 May 2021. Retrieved 27 April 2016. "These Gödelian anti-mechanist arguments are, however, problematic, and there is wide consensus that they fail."*
37. *Russell, Stuart J.; Norvig, Peter (2010). "26.1.2: Philosophical Foundations/Weak AI: Can Machines Act Intelligently?/The mathematical objection". Artificial Intelligence: A Modern Approach (3rd ed.). Upper Saddle River, NJ: Prentice Hall. ISBN 978-0-13-604259-4. "...even if we grant that computers have limitations on what they can prove, there is no*

evidence that humans are immune from those limitations."

38. Mark Colyvan. An Introduction to the Philosophy of Mathematics. Cambridge University Press, 2012. From 2.2.2, 'Philosophical significance of Gödel's incompleteness results': "The accepted wisdom (with which I concur) is that the Lucas-Penrose arguments fail."
39. LaForte, G.; Hayes, P. J.; Ford, K. M., 1998. Why Gödel's theorem cannot refute computationalism. Artificial Intelligence, 104: 265–286, 1998.
40. Russell & Norvig 2003, p. 950. They point out that real machines with finite memory can be modeled using propositional logic, which is formally decidable, and Gödel's argument does not apply to them.
41. Hofstadter 1979.
42. According to Hofstadter 1979, pp. 476–477, this statement was first proposed by C. H. Whiteley.
43. Hofstadter 1979, pp. 476–477, Russell & Norvig 2003, p. 950, Turing 1950 under "The Argument from Mathematics" where he writes "although it is established that there are limitations to the powers of any particular machine, it has only been stated, without sort of proof, that no such limitations apply to the human intellect."
44. Penrose 1989.
45. *Litt, Abninder; Eliasmith, Chris; Kroon, Frederick W.; Weinstein, Steven; Thagard, Paul (6 May 2006). "Is the Brain a Quantum Computer?". Cognitive Science.*

30 *(3): 593–603.*
doi:10.1207/s15516709cog0000_59.
PMID 21702826.

46. Dreyfus 1972, Dreyfus 1979, Dreyfus & Dreyfus 1986. See also Russell & Norvig 2003, pp. 950–952, Crevier 1993, pp. 120–132 and Fearn 2007, pp. 50–51.
47. Russell & Norvig 2003, pp. 950–951.
48. Turing 1950, under "(8) The Argument from the Informality of Behavior".
49. ^ ***a*** ***b*** Russell & Norvig 2003, p. 52.
50. See Brooks 1990, and Moravec 1988.
51. *Kahneman, Daniel (2011). Thinking, Fast and Slow. Macmillan. ISBN 978-1-4299-6935-2. Archived from the original on March 15, 2023. Retrieved April 8, 2012.*
52. Crevier 1993, p. 125.
53. ^ ***a*** ***b*** Turing 1950, under "(4) The Argument from Consciousness". See also Russell & Norvig 2003, pp. 952–953, where they identify Searle's argument with Turing's "Argument from Consciousness."
54. Russell & Norvig 2003, p. 947.
55. Blackmore 2005, p. 1.
56. "[P]eople always tell me it was very hard to define consciousness, but I think if you're just looking for the kind of commonsense definition that you get at the beginning of the investigation, and not at the hard nosed scientific definition that comes at the end, it's not hard to give commonsense definition of

consciousness." The Philosopher's Zone: The question of consciousness, Archived 2007-11-28 at the Wayback Machine. Also see Dennett 1991.

57. Blackmore 2005, p. 2.
58. Russell & Norvig 2003, pp. 954–956.
59. For example, John Searle writes: "Can a machine think? The answer is, obvious, yes. We are precisely such machines." (Searle 1980, p. 11).
60. Searle 1980. See also Cole 2004, Russell & Norvig 2003, pp. 958–960, Crevier 1993, pp. 269–272 and Hearn 2007, pp. 43–50.
61. Searle 1980, p. 13.
62. Searle 1984.
63. Cole 2004, 2.1, Leibniz 1714, 17.
64. Cole 2004, 2.3.
65. Searle 1980, under "1. The Systems Reply (Berkeley)", Crevier 1993, p. 269, Russell & Norvig 2003, p. 959, Cole 2004, 4.1. Among those who hold to the "system" position (according to Cole) are Ned Block, Jack Copeland, Daniel Dennett, Jerry Fodor, John Haugeland, Ray Kurzweil and Georges Rey. Those who have defended the "virtual mind" reply include Marvin Minsky, Alan Perlis, David Chalmers, Ned Block and J. Cole (again, according to Cole 2004).
66. Cole 2004, 4.2 ascribes this position to Ned Block, Daniel Dennett, Tim Maudlin, David Chalmers, Steven Pinker, Patricia Churchland and others.
67. Searle 1980, under "2. The Robot Reply (Yale)". Cole

2004, 4.3 ascribes this position to Margaret Boden, Tim Crane, Daniel Dennett, Jerry Fodor, Stevan Harnad, Hans Moravec and Georges Rey.
68. Quoted in Crevier 1993, p. 272.
69. Searle 1980, under "3. The Brain Simulator Reply (Berkeley and M.I.T.)", Cole 2004, ascribes this position to Paul and Patricia Churchland and Ray Kurzweil.
70. Searle 1980 under "5. The Other Minds Reply", Cole 2004, 4.4. Turing 1950 makes this reply under "(4) The Argument from Consciousness." Cole ascribes this position to Daniel Dennett and Hans Moravec.
71. Dreyfus 1979, p. 156, Haugeland 1985, pp. 15–44.
72. Horst 2005.
73. ^ *a b* Harnad 2001.
74. ^ *a b* Quoted in Crevier 1993, p. 266.
75. Crevier 1993, p. 266.
76. ^ *a b c* Turing 1950 under "(5) Arguments from Various Disabilities".
77. Turing 1950 under "(6) Lady Lovelace's Objection".
78. Turing 1950 under "(5) Argument from Various Disabilities".
79. *"Kaplan Andreas; Michael Haenlein". Business Horizons.* ***62*** *(1): 15–25. January 2019. doi:10.1016/j.bushor.2018.08.004. S2CID 158433736.*
80. *Katz, Leslie (2009-04-02). "Robo-scientist makes gene discovery-on its own | Crave". News.cnet.com. Archived from the original on July 12, 2012. Retrieved*

2010-07-29.
81. ^ *a b* John Markoff, Scientists Worry Machines May Outsmart Man, Archived 2017-07-01 at the Wayback Machine, New York Times, July 26, 2009.
82. Vernor Vinge. The Coming Technological Singularity: How to Survive in the Post-Human Era, Department of Mathematical Sciences, San Diego State University, (c) 1993 by Vernor Vinge.
83. Jason Palmer. Call for debate on killer robots Archived 2009-08-07 at the Wayback Machine, BBC News, 8/3/09.
84. Jason Mick (blog). Science New Navy-funded Report Warns of War Robots Going "Terminator", Archived 2009-07-28 at the Wayback Machine, dailytech.com, February 17, 2009.
85. Joseph L. Flatley. Navy report warns of robot uprising, suggests a strong moral compass, Archived 2011-06-04 at the Wayback Machine, engadget.com, February 18th, 2009.
86. AAAI Presidential Panel on Long-Term AI Futures 2008-2009 Study, Archived 2009-08-28 at the Wayback Machine, Association for the Advancement of Artificial Intelligence, Accessed 7/26/09.
87. Article at Asimovlaws.com, July 2004, accessed 7/27/09. Archived June 30, 2009, at the Wayback Machine.
88. 'Can digital computers think?'. Talk broadcast on BBC Third Programme, 15 May 1951. http://www.turingarchive.org/viewer/?id=459&title=8.

89. Turing 1950 under "(1) The Theological Objection", although he also writes, "I am not very impressed with theological arguments whatever they may be used to support".
90. *Specktor, Brandon (2022-06-13). "Google AI 'is sentient,' software engineer claims before being suspended". livescience.com. Archived from the original on 2022-06-14. Retrieved 2022-06-14.*
91. *Lemoine, Blake (2022-06-11). "Is LaMDA Sentient? — an Interview". Medium. Archived from the original on 2022-06-13. Retrieved 2022-06-14.*
92. M. Morioka et al. (2023-01-15) *Artificial Intelligence, Robots, and Philosophy*, *Archived 2022-12-28 at the Wayback Machine*, pp. 2–4.
93. *Deutsch, David (2012-10-03). "Philosophy will be the key that unlocks artificial intelligence | David Deutsch". the Guardian. Archived from the original on 2013-09-27. Retrieved 2018-09-18.*

Works cited

- Adam, Alison (1989). Artificial Knowing: Gender and the Thinking Machine. *Routledge & CRC Press*. ISBN 978-0-415-12963-3
- Benjamin, Ruha (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Wiley. ISBN 978-1-509-52643-7
- *Blackmore, Susan (2005), Consciousness: A Very Short Introduction, Oxford University Press*

- Bostrom, Nick (2014), Superintelligence: Paths, Dangers, Strategies, Oxford University Press, ISBN 978-0-19-967811-2
- Brooks, Rodney (1990), "Elephants Don't Play Chess" (PDF), Robotics and Autonomous Systems, **6** (1–2): 3–15, CiteSeerX 10.1.1.588.7539, doi:10.1016/S0921-8890(05)80025-9, retrieved 2007-08-30
- Bryson, Joanna (2019). The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation, p. 34
- Chalmers, David J (1996), The Conscious Mind: In Search of a Fundamental Theory, Oxford University Press, New York, ISBN 978-0-19-511789-9
- Cole, David (Fall 2004), "The Chinese Room Argument", in Zalta, Edward N. (ed.), The Stanford Encyclopedia of Philosophy
- Crawford, Kate (2021). Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press
- Crevier, Daniel (1993). AI: The Tumultuous Search for Artificial Intelligence. New York, NY: BasicBooks. ISBN 0-465-02997-3.
- Dennett, Daniel (1991), Consciousness Explained, The Penguin Press, ISBN 978-0-7139-9037-9
- Dreyfus, Hubert (1972), What Computers Can't Do, New York: MIT Press, ISBN 978-0-06-011082-6
- Dreyfus, Hubert (1979), What Computers Still Can't Do, New York: MIT Press
- Dreyfus, Hubert; Dreyfus, Stuart (1986), Mind over

Machine: The Power of Human Intuition and Expertise in the Era of the Computer, Oxford, UK: Blackwell

- *Fearn, Nicholas (2007), The Latest Answers to the Oldest Questions: A Philosophical Adventure with the World's Greatest Thinkers, New York: Grove Press*
- *Gladwell, Malcolm (2005), Blink: The Power of Thinking Without Thinking, Boston, Massachusetts: Little, Brown, ISBN 978-0-316-17232-5*
- *Harnad, Stevan (2001), "What's Wrong and Right About Searle's Chinese Room Argument?", in Bishop, M.; Preston, J. (eds.), Essays on Searle's Chinese Room Argument, Oxford University Press*
- Haraway, Donna (1985). *A Cyborg Manifesto*
- *Haugeland, John (1985), Artificial Intelligence: The Very Idea, Cambridge, Massachusetts: MIT Press*
- *Hofstadter, Douglas (1979), Gödel, Escher, Bach: an Eternal Golden Braid*
- *Horst, Steven (2009), "The Computational Theory of Mind", in Zalta, Edward N. (ed.), The Stanford Encyclopedia of Philosophy, Metaphysics Research Lab, Stanford University*
- *Kaplan, Andreas; Haenlein, Michael (2018), "Siri, Siri in my Hand, who's the Fairest in the Land? On the Interpretations, Illustrations and Implications of Artificial Intelligence", Business Horizons, **62**: 15–25, doi:10.1016/j.bushor.2018.08.004, S2CID 158433736*
- *Kurzweil, Ray (2005), The Singularity is Near, New York: Viking Press, ISBN 978-0-670-03384-3*

- Lucas, John (1961), "Minds, Machines and Gödel", in Anderson, A. R. (ed.), Minds and Machines
- Malabou, Catherine (2019). *Morphing Intelligence: From IQ Measurement to Artificial Brains*. (C. Shread, Trans.). Columbia University Press
- McCarthy, John; Minsky, Marvin; Rochester, Nathan; Shannon, Claude (1955), A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, archived from the original on 2008-09-30
- McCarthy, John (1999), What is AI?, archived from the original on 4 December 2022, retrieved 4 December 2022
- McDermott, Drew (May 14, 1997), "How Intelligent is Deep Blue", New York Times, archived from the original on October 4, 2007, retrieved October 10, 2007
- Moravec, Hans (1988), Mind Children, Harvard University Press
- Penrose, Roger (1989), The Emperor's New Mind: Concerning Computers, Minds, and The Laws of Physics, Oxford University Press, ISBN 978-0-14-014534-2c
- Rescorla, Michael, "The Computational Theory of Mind", in:Edward N. Zalta (ed.), The Stanford Encyclopedia of Philosophy (Fall 2020 Edition)
- Russell, Stuart J.; Norvig, Peter (2003), Artificial Intelligence: A Modern Approach (2nd ed.), Upper Saddle River, New Jersey: Prentice Hall, ISBN 0-13-

790395-2
- Searle, John (1980), "Minds, Brains and Programs" (PDF), Behavioral and Brain Sciences, **3** (3): 417–457, doi:10.1017/S0140525X00005756, S2CID 55303721, archived from the original (PDF) on 2015-09-23
- Searle, John (1992), The Rediscovery of the Mind, Cambridge, Massachusetts: M.I.T. Press
- Searle, John (1999), Mind, language and society, New York, NY: Basic Books, ISBN 978-0-465-04521-1, OCLC 231867665
- Turing, Alan (October 1950). "Computing Machinery and Intelligence". Mind. **59** (236): 433–460. doi:10.1093/mind/LIX.236.433. ISSN 1460-2113. JSTOR 2251299. S2CID 14636783.

- v
- t
- e

Philosophy of mind

- G. E. M. Anscombe
- Aristotle
- Armstrong
- Thomas Aquinas
- J. L. Austin
- Alexander Bain
- George Berkeley
- Henri Bergson
- Ned Block
- Franz Brentano
- C. D. Broad
- Tyler Burge

Philosophers	• David Chalmers • Patricia Churchland • Paul Churchland • Andy Clark • Dharmakirti • Donald Davidson • Daniel Dennett • René Descartes • Fred Dretske • Fodor • Goldman • Martin Heidegger • David Hume • Edmund Husserl • William James • Frank Cameron Jackson • Immanuel Kant • David Lewis (philosopher) • John Locke • Gottfried Wilhelm Leibniz • Maurice Merleau-Ponty • Marvin Minsky • Thomas Nagel • Alva Noë • Derek Parfit • Plato • Hilary Putnam • Richard Rorty • Gilbert Ryle • John Searle • Wilfrid Sellars • Baruch Spinoza • Alan Turing • Michael Tye • Vasubandhu • Ludwig Wittgenstein • Stephen Yablo • Zhuangzi • *more...*

Theories	• Behaviorism
	• Biological naturalism
	• Dualism
	• Eliminative materialism
	• Emergent materialism
	• Epiphenomenalism
	• Functionalism
	• Interactionism
	• Naïve realism
	  ◦ Neurophenomenology
	• Neutral monism
	• New mysterianism
	• Nondualism
	• Occasionalism
	• Parallelism
	• Phenomenalism
	• Phenomenology
	• Physicalism
	  ◦ Type physicalism
	• Property dualism
	• Representational
	• Solipsism
	• Substance dualism

	• Abstract object
	• Chinese room
	• Creativity
	• Cognition
	• Cognitive closure
	• Concept
	• Consciousness
	• Hard problem of consciousness
	• Hypostatic abstraction
	• Idea
	• Identity
	• Intelligence
	  ◦ Artificial
	  ◦ Human

Concepts	• Intentionality • Introspection • Intuition • Language of thought • Mental event • Mental image • Mental process • Mental property • Mental representation • Mind • Mind–body problem • Pain • Problem of other minds • Propositional attitude • Qualia • Tabula rasa • Understanding • Zombie
Related	• Metaphysics • Philosophy of artificial intelligence / information / perception / self

- Category
- Philosophers category
- Project
- Task Force

- v
- t
- e

Philosophy of science

	• Analysis • Analytic–synthetic distinction • *A priori* and *a posteriori* • Causality ◦ Mill's Methods

Concepts	• Commensurability
	• Consilience
	• Construct
	• Correlation
	  ◦ function
	• Creative synthesis
	• Demarcation problem
	• Empirical evidence
	• Experiment
	  ◦ design
	• Explanatory power
	• Fact
	• Falsifiability
	• Feminist method
	• Functional contextualism
	• Hypothesis
	  ◦ alternative
	  ◦ null
	• *Ignoramus et ignorabimus*
	• Inductive reasoning
	• Intertheoretic reduction
	• Inquiry
	• Nature
	• Objectivity
	• Observation
	• Paradigm
	• Problem of induction
	• Scientific evidence
	  ◦ Evidence-based practice
	• Scientific law
	• Scientific method
	• Scientific pluralism
	• Scientific Revolution
	• Testability
	• Theory
	  ◦ choice
	  ◦ ladenness
	  ◦ scientific
	• Underdetermination

	• Unity of science
	• Variable
	  ◦ control
	  ◦ dependent and independent
	• more...

Theories	• Coherentism • Confirmation holism • Constructive empiricism • Constructive realism • Constructivist epistemology • Contextualism • Conventionalism • Deductive-nomological model • Epistemological anarchism • Evolutionism • Fallibilism • Foundationalism • Hypothetico-deductive model • Inductionism • Instrumentalism • Model-dependent realism • Naturalism • Physicalism • Positivism / Reductionism / Determinism • Pragmatism • Rationalism / Empiricism • Received view / Semantic view of theories • Scientific essentialism • Scientific formalism • Scientific realism / Anti-realism • Scientific skepticism • Scientism • Structuralism • Uniformitarianism • Verificationism

	• Vitalism

Philosophy of...	• Biology
	• Chemistry
	• Physics
	  ◦ Space and time
	• Social science
	  ◦ Archaeology
	  ◦ Economics
	  ◦ Geography
	  ◦ History
	  ◦ Linguistics
	  ◦ Psychology

Related topics	• Criticism of science • Descriptive science • Epistemology • Exact sciences • Faith and rationality • Hard and soft science • History and philosophy of science • Non-science ◦ Pseudoscience • Normative science • Protoscience • Questionable cause • Relationship between religion and science • Rhetoric of science • Science studies • Sociology of scientific ignorance • Sociology of scientific knowledge
	• Roger Bacon • Francis Bacon

Philosophers of science	**Precursors**	• Galileo Galilei • Isaac Newton • David Hume
	• Auguste Comte • Henri Poincaré • Pierre Duhem • Rudolf Steiner • Karl Pearson • Charles Sanders Peirce • Wilhelm Windelband • Alfred North Whitehead • Bertrand Russell • Otto Neurath • C. D. Broad • Michael Polanyi • Hans Reichenbach • Rudolf Carnap • Karl Popper • Carl Gustav Hempel • W. V. O. Quine • Thomas Kuhn • Imre Lakatos • Paul Feyerabend • Ian Hacking • Bas van Fraassen • Larry Laudan	

- Category
- Philosophy portal
- Science portal

- v
- t
- e

Psychology

- History

- Philosophy
- Portal
- Psychologist

Basic psychology	• Abnormal • Affective neuroscience • Affective science • Behavioral genetics • Behavioral neuroscience • Behaviorism • Cognitive/Cognitivism • Cognitive neuroscience ◦ Social • Comparative • Cross-cultural • Cultural • Developmental • Differential • Ecological • Evolutionary • Experimental • Gestalt • Intelligence • Mathematical • Moral • Neuropsychology • Perception • Personality • Psycholinguistics • Psychophysiology • Quantitative • Social • Theoretical	
	• Anomalistic • Applied behavior analysis • Assessment • Clinical	

Applied psychology	• Coaching • Community • Consumer • Counseling • Critical • Educational • Ergonomics • Feminist • Forensic • Health • Humanistic • Industrial and organizational • Legal • Media • Medical • Military • Music • Occupational health • Pastoral • Political • Positive • Psychometrics • Psychotherapy • Religion • School • Sport and exercise • Suicidology • Systems • Traffic
Methodologies	• Animal testing • Archival research • Behavior epigenetics • Case study • Content analysis • Experiments • Human subject research • Interviews • Neuroimaging

	• Observation • Psychophysics • Qualitative research • Quantitative research • Self-report inventory • Statistical surveys
Concepts	• Behavior • Behavioral engineering • Behavioral genetics • Behavioral neuroscience • Cognition • Competence • Consciousness • Consumer behavior • Emotions • Feelings • Human factors and ergonomics • Intelligence • Mind • Psychology of religion • Psychometrics • Terror management theory
	• Wilhelm Wundt • William James • Ivan Pavlov • Sigmund Freud • Edward Thorndike • Carl Jung • John B. Watson • Clark L. Hull • Kurt Lewin • Jean Piaget • Gordon Allport • J. P. Guilford • Carl Rogers • Erik Erikson

Psychologists	• B. F. Skinner • Donald O. Hebb • Ernest Hilgard • Harry Harlow • Raymond Cattell • Abraham Maslow • Neal E. Miller • Jerome Bruner • Donald T. Campbell • Hans Eysenck • Herbert A. Simon • David McClelland • Leon Festinger • George A. Miller • Richard Lazarus • Stanley Schachter • Robert Zajonc • Albert Bandura • Roger Brown • Endel Tulving • Lawrence Kohlberg • Noam Chomsky • Ulric Neisser • Jerome Kagan • Walter Mischel • Elliot Aronson • Daniel Kahneman • Paul Ekman • Michael Posner • Amos Tversky • Bruce McEwen • Larry Squire • Richard E. Nisbett • Martin Seligman • Ed Diener • Shelley E. Taylor • John Anderson • Ronald C. Kessler • Joseph E. LeDoux		

	• Richard Davidson • Susan Fiske • Roy Baumeister
Lists	• Counseling topics • Disciplines • Organizations • Outline • Psychologists • Psychotherapies • Research methods • Schools of thought • Timeline • Topics

- Wiktionary definition
- Wiktionary category
- Wikisource
- Wikimedia Commons
- Wikiquote
- Wikinews
- Wikibooks

- **v**
- **t**
- **e**

Philosophy

Branches	
Branches	• Applied philosophy • Logic • Metaphilosophy • Philosophy of information • Philosophy of language • Philosophy of mathematics • Philosophy of religion • Philosophy of science • Political philosophy • Practical philosophy • Social philosophy

	• Theoretical philosophy
Aesthetics	• Aesthetic response • Formalism • Institutionalism
Epistemology	• Empiricism • Fideism • Naturalism • Particularism • Rationalism • Skepticism • Solipsism
Ethics	• Consequentialism • Deontology • Virtue

Free will	• Compatibilism
	• Determinism
	  ◦ Hard
	• Incompatibilism
	  ◦ Hard
	• Libertarianism

Metaphysics	• Atomism • Dualism • Idealism • Monism • Naturalism • Realism
Mind	• Behaviorism • Eliminativism • Emergentism • Epiphenomenalism • Functionalism • Objectivism • Subjectivism
Normativity	• Absolutism • Particularism • Relativism • Nihilism • Skepticism • Universalism

Ontology	• Action • Event • Process
Reality	• Anti-realism • Conceptualism • Idealism • Materialism • Naturalism • Nominalism • Physicalism • Realism

By era

By era		• Ancient
		• Western
		  ◦ Medieval
		  ◦ Renaissance
		  ◦ Early modern
		  ◦ Modern
		  ◦ Contemporary

	Chinese	• Agriculturalism • Confucianism • Legalism • Logicians • Mohism • Chinese naturalism • Taoism • Yangism

	Greco-Roman	• Presocratic
		  ◦ Ionians
		  ◦ Pythagoreans
		  ◦ Eleatics
		  ◦ Atomists
		  ◦ Sophists
		• Cyrenaics
		• Cynicism
		• Eretrian school
		• Megarian school
		• Academy
		• Peripatetic school
		• Hellenistic philosophy
		  ◦ Pyrrhonism
		  ◦ Stoicism
		  ◦ Epicureanism

Ancient		◦ Academic Skepticism • Middle Platonism • School of the Sextii • Neopythagoreanism • Second Sophistic • Neoplatonism • Church Fathers
	Indian	• Hindu ◦ Samkhya ◦ Nyaya ◦ Vaisheshika ◦ Yoga ◦ Mīmāṃsā ◦ Ājīvika ◦ Ajñana ◦ Cārvāka • Jain ◦ Anekantavada ◦ Syādvāda • Buddhist ◦ Abhidharma ◦ Sarvāstivadā ◦ Pudgalavada ◦ Sautrāntika ◦ Madhyamaka ◦ Svatantrika and Prasangika ◦ Śūnyatā ◦ Yogacara ◦ Tibetan
	Persian	• Mazdakism • Mithraism • Zoroastrianism • Zurvanism
	East Asian	• Neotaoism • Tiantai • Huayan • Chan • Zen • Neo-Confucianism • Korean Confucianism
	European	• Christian • Augustinianism • Scholasticism • Thomism • Scotism • Occamism • Renaissance humanism

Medieval	**Indian**	• Vedanta
		  ◦ Acintya bheda abheda
		  ◦ Advaita
		  ◦ Bhedabheda
		  ◦ Dvaita
		  ◦ Nimbarka Sampradaya
		  ◦ Shuddhadvaita
		  ◦ Vishishtadvaita
		• Navya-Nyāya

	Islamic	• Aristotelianism • Averroism • Avicennism • Illuminationism • *ʿIlm al-Kalām* • Sufi
	Jewish	• Judeo-Islamic

Modern	• Anarchism
	• Classical Realism
	• Collectivism
	• Conservatism
	• Determinism
	• Dualism
	• Edo neo-Confucianism
	• Empiricism
	• Existentialism
	• Foundationalism
	• Historicism
	• Holism
	• Humanism
	  ◦ Anti-
	• Idealism
	  ◦ Absolute
	  ◦ British
	  ◦ German
	  ◦ Objective
	  ◦ Subjective
	  ◦ Transcendental
	• Individualism
	• Kokugaku
	• Liberalism
	• Materialism
	• Modernism
	• Monism
	• Naturalism
	• Natural law
	• Nihilism
	• New Confucianism
	• Neo-scholasticism
	• Pragmatism	

		• Phenomenology • Positivism • Reductionism • Rationalism • Social contract • Socialism • Transcendentalism • Utilitarianism
	People	• Cartesianism • Kantianism ◦ Neo • Kierkegaardianism • Krausism • Hegelianism • Marxism • Newtonianism • Nietzscheanism • Spinozism
Contemporary	**Analytic**	• Applied ethics • Analytic feminism • Analytical Marxism • Communitarianism • Consequentialism • Critical rationalism • Experimental philosophy • Falsificationism • Foundationalism / Coherentism • Internalism and externalism • Logical positivism • Legal positivism • Meta-ethics • Moral realism • Quinean naturalism • Normative ethics • Ordinary language philosophy • Postanalytic philosophy • Quietism • Rawlsian • Reformed epistemology • Systemics • Scientism • Scientific realism • Scientific skepticism • Transactionalism • Contemporary utilitarianism • Vienna Circle • Wittgensteinian
		• Critical theory

	Continental	• Deconstruction • Existentialism • Feminist • Frankfurt School • Hermeneutics • Neo-Marxism • New Historicism • Phenomenology • Posthumanism • Postmodernism • Post-structuralism • Social constructionism • Structuralism • Western Marxism
	Miscellaneous	• Kyoto School • Objectivism • Postcritique • Russian cosmism • *more...*

- **By region**

By region	**African**	• Bantu • Egyptian • Ethiopian • Africana
	Eastern	• Buddhist • Chinese • Indian • Indonesian • Japanese • Korean • Taiwanese • Vietnamese
	Middle Eastern	• Iranian • Islamic • Jewish • Pakistani • Turkish
		• American • Australian • British ◦ Scottish • Canada • Czech • Danish

	Western	• Dutch • Finland • French • German • Greek • Italian • Maltese • Polish • Slovene • Spanish
	Miscellaneous	• Amerindian • Aztec • Romanian • Russian • Yugoslav

- **Philosophy portal**
- **Category**