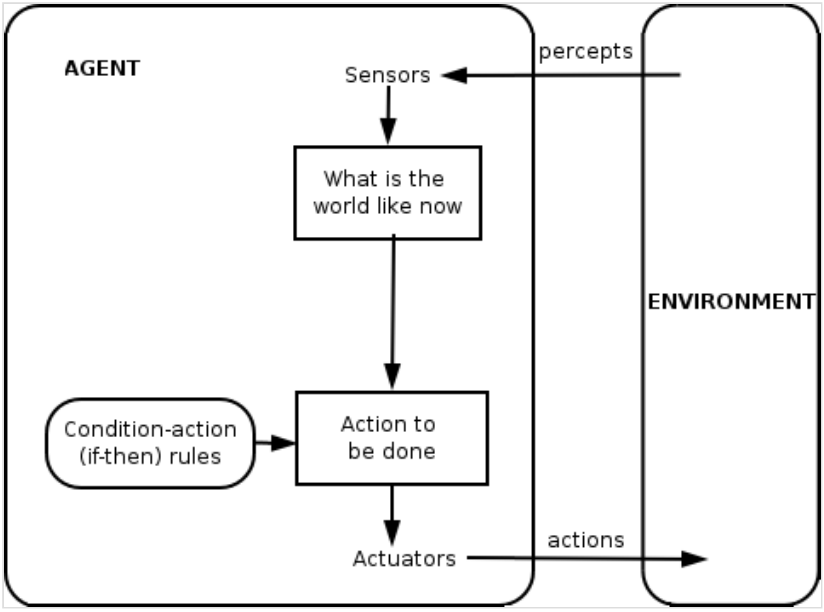
WIKIPEDIA
The Free Encyclopedia

Intelligent agent

In intelligence and artificial intelligence, an **intelligent agent** (**IA**) is an agent acting in an intelligent manner; It perceives its environment, takes actions autonomously in order to achieve goals, and may improve its performance with learning or acquiring knowledge. An intelligent agent may be simple or complex: A thermostat or other control system is considered an example of an intelligent agent, as is a human being, as is any system that meets the definition, such as a firm, a state, or a biome.[1]

Leading AI textbooks define "artificial intelligence" as the "study and design of intelligent agents", a definition that considers goal-directed behavior to be the essence of intelligence. Goal-directed agents are also described using a term borrowed from economics, "rational agent".[1]



Simple reflex agent diagram

An agent has an "objective function" that encapsulates all the IA's goals. Such an agent is designed to create and execute whatever plan will, upon completion, maximize the expected value of the objective function.[2] For example, a reinforcement learning agent has a "reward function" that allows the programmers to shape the IA's desired behavior,[3] and an evolutionary algorithm's behavior is shaped by a "fitness function".[4]

Intelligent agents in artificial intelligence are closely related to agents in economics, and versions of the intelligent agent paradigm are studied in cognitive science, ethics, the philosophy of practical reason, as well as in many interdisciplinary socio-cognitive modeling and computer social simulations.

Intelligent agents are often described schematically as an abstract functional system similar to a computer program. Abstract descriptions of intelligent agents are called **abstract intelligent agents** (**AIA**) to distinguish them from their real-world implementations. An **autonomous intelligent agent** is designed to function in the absence of human intervention. Intelligent agents are also closely related to software agents (an autonomous computer program that carries out tasks on behalf of users).

As a definition of artificial intelligence

Artificial Intelligence: A Modern Approach[5][6][2] defines an "agent" as

> "Anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators"

It defines a "rational agent" as:

> "An agent that acts so as to maximize the expected value of a performance measure based on past experience and knowledge."

It also defines the field of "artificial intelligence research" as:

> "The study and design of rational agents"

Padgham & Winikoff (2005) agree that an intelligent agent is situated in an environment and responds in a timely (though not necessarily real-time) manner to changes in the environment. However, intelligent agents must also proactively pursue goals in a flexible and robust way.[a] Optional desiderata include that the agent be rational, and that the agent be capable of belief-desire-intention analysis.[7]

Kaplan and Haenlein define artificial intelligence as "A system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation."[8] This definition is closely related to that of an intelligent agent.

Advantages

Philosophically, this definition of artificial intelligence avoids several lines of criticism. Unlike the Turing test, it does not refer to human intelligence in any way. Thus, there is no need to discuss if it is "real" vs "simulated" intelligence (i.e., "synthetic" vs "artificial" intelligence) and does not indicate that such a machine has a mind, consciousness or true understanding (i.e., it does not imply John Searle's "strong AI hypothesis"). It also doesn't attempt to draw a sharp dividing line between behaviors that are "intelligent" and behaviors that are "unintelligent"—programs need only be measured in terms of their objective function.

More importantly, it has a number of practical advantages that have helped move AI research forward. It provides a reliable and scientific way to test programs; researchers can directly compare or even combine different approaches to isolated problems, by asking which agent is best at maximizing a given "goal function". It also gives them a common language to communicate with other fields—such as mathematical optimization (which is defined in terms of "goals") or economics (which uses the same definition of a "rational agent").[9]

Objective function

An agent that is assigned an explicit "goal function" is considered more intelligent if it consistently takes actions that successfully maximize its programmed goal function. The goal can be simple ("1 if the IA wins a game of Go, 0 otherwise") or complex ("Perform actions mathematically similar to ones that succeeded in the past"). The "goal function" encapsulates all of the goals the agent is driven to act on; in the case of rational agents, the function also encapsulates the acceptable trade-offs between accomplishing conflicting goals. (Terminology varies; for example, some agents seek to maximize or minimize a "utility function", "objective function", or "loss function".)[6][2]

Goals can be explicitly defined or induced. If the AI is programmed for "reinforcement learning", it has a "reward function" that encourages some types of behavior and punishes others. Alternatively, an evolutionary system can induce goals by using a "fitness function" to mutate and preferentially replicate high-scoring AI systems, similar to how animals evolved to innately desire certain goals such as finding food.[10] Some AI systems, such as nearest-neighbor, instead of reason by analogy, these systems are not generally given goals, except to the degree that goals are implicit in their training data.[11] Such systems can still be benchmarked if the non-goal system is framed as a system whose "goal" is to accomplish its narrow classification task.[12]

Systems that are not traditionally considered agents, such as knowledge-representation systems, are sometimes subsumed into the paradigm by framing them as agents that have a goal of (for example) answering questions as accurately as possible; the concept of an "action" is here extended to encompass the "act" of giving an answer to a question. As an additional extension, mimicry-driven systems can be framed as agents who are optimizing a "goal function" based on how closely the IA succeeds in mimicking the desired behavior.[6][2] In the generative adversarial networks of the 2010s, an "encoder"/"generator" component attempts to mimic and improvise human text composition. The generator is attempting to maximize a function encapsulating how well it can fool an antagonistic "predictor"/"discriminator" component.[13]

While symbolic AI systems often accept an explicit goal function, the paradigm can also be applied to neural networks and to evolutionary computing. Reinforcement learning can generate intelligent agents that appear to act in ways intended to maximize a "reward function".[14] Sometimes, rather than setting the reward function to be directly equal to the desired benchmark evaluation function, machine learning programmers will use reward shaping to initially give the machine rewards for incremental progress in learning.[15] Yann LeCun stated in 2018 that "Most of the learning algorithms that people have come up with essentially consist of minimizing some objective function."[16] AlphaZero chess had a simple objective function; each win counted as +1 point, and each loss counted as -1 point. An objective function for a self-driving car would have to be more complicated.[17] Evolutionary computing can evolve intelligent agents that appear to act in ways intended to maximize a "fitness function" that influences how many descendants each agent is allowed to leave.[4]

The theoretical and uncomputable AIXI design is a maximally intelligent agent in this paradigm;[18] however, in the real world, the IA is constrained by finite time and hardware resources, and scientists compete to produce algorithms that can achieve progressively higher scores on benchmark tests with real-world hardware.[19]

Classes of intelligent agents

Russell and Norvig's classification

Russell & Norvig (2003) group agents into five classes based on their degree of perceived intelligence and capability:[20]

Simple reflex agents

Simple reflex agents act only on the basis of the current percept, ignoring the rest of the percept history. The agent function is based on the *condition-action rule*: "if condition, then action".